JMIR Mental Health

Journal Impact Factor (JIF) (2023): 4.8
Volume 11 (2024) ISSN 2368-7959 Editor in Chief: John Torous, MD, MBI

Contents

Reviews

Viewpoints

Corrigenda and Addendas

Research Letter

Editorial

Digital Tools to Facilitate the Detection and Treatment of Bipolar Disorder: Key Developments and Future Directions

Taiane de Azevedo Cardoso[1,2], PhD; Shruti Kochhar[2], MS; John Torous[3], MD, MBI; Emma Morton[4], PhD

[1]The Institute for Mental and Physical Health and Clinical Translation, School of Medicine, Deakin University, Geelong, Australia

[2]JMIR Publications, Toronto, ON, Canada

[3]Digital Psychiatry, Department of Psychiatry, Beth Israel Deaconess Medical Center, Boston, MA, United States

[4]School of Psychological Sciences, Monash University, Clayton, Australia

Corresponding Author:
Taiane de Azevedo Cardoso, PhD
JMIR Publications
130 Queens Quay E Suite 1100
Toronto, ON, M5A 0P6
Canada
Phone: 1 2899411482
Email: taiane.cardoso@jmir.org

Abstract

Bipolar disorder (BD) impacts over 40 million people around the world, often manifesting in early adulthood and substantially impacting the quality of life and functioning of individuals. Although early interventions are associated with a better prognosis, the early detection of BD is challenging given the high degree of similarity with other psychiatric conditions, including major depressive disorder, which corroborates the high rates of misdiagnosis. Further, BD has a chronic, relapsing course, and the majority of patients will go on to experience mood relapses despite pharmacological treatment. Digital technologies present promising results to augment early detection of symptoms and enhance BD treatment. In this editorial, we will discuss current findings on the use of digital technologies in the field of BD, while debating the challenges associated with their implementation in clinical practice and the future directions.

(JMIR Ment Health 2024;11:e58631) doi:10.2196/58631

KEYWORDS

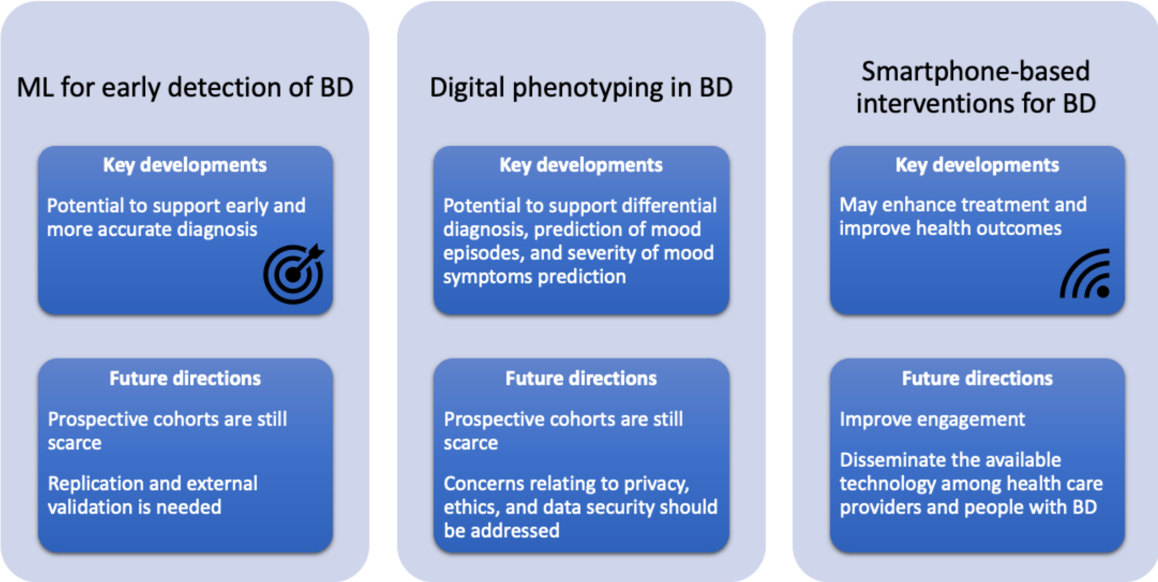
bipolar disorder; digital phenotyping; machine learning; mobile health interventions; mobile health; mHealth; apps

Introduction

Bipolar disorder (BD) is a chronic and recurrent mental illness that affects 2.4% of the worldwide population [1]. BD usually manifests in early adulthood, with the median age at onset found to be 33 years of age and a peak age at onset of 19.5 years of age [2]. BD presents a profound negative impact on individuals' lives with high rates of disability [3]. According to the Global Burden of Disease Study (2019), BD is the 12th leading cause of years lived with disability among young adults aged between 15 to 24 years [4].

Digital health technologies have been studied in the context of BD and are showing promising results in the early detection of the disorder [5,6] and depressive or manic episodes among individuals with the disorder [7], as well as the promotion of a better prognosis [8,9]. To understand this progress, we will review 3 promising and innovative areas of work (Figure 1). In this editorial, we will discuss the role of (1) machine learning techniques, (2) digital phenotyping, and (3) mobile health (mHealth) apps to enhance BD care. Additionally, the challenges and future directions for the implementation of digital health technologies in BD will be considered.

Figure 1. Digital tools for BD: key developments and future directions. BD: bipolar disorder; ML: machine learning.



Early Detection of BD and Reducing Misdiagnosis: Insights From Machine Learning Studies

A potential contributor to the disease burden in BD is the delay in obtaining an accurate diagnosis, which consequently delays the appropriate management and treatment of the disorder. A recent systematic review showed that the median delay in help seeking was 3.5 years, the median delay in diagnosis was 6.7 years, and the median duration of untreated BD was 5.9 years [10]. Another recent study found that the rate of misdiagnosis in BD was 76.8%, and most of those cases received a misdiagnosis of major depressive disorder (MDD) [11]. Despite the similarities in the clinical presentation of a depressive episode in MDD and BD, the treatment strategies recommended for each disorder are different, with antidepressants being the main pharmacological strategy in MDD [12] and mood stabilizers being recommended for BD [13]. Thus, having strategies for the early detection of BD is crucial to reduce the misdiagnosis rates and to provide the proper treatment early in the course of the disorder. In this section, we will be describing some machine learning studies aimed at (1) predicting mood disorder misdiagnosis, (2) predicting BD onset, and (3) differentiating BD from unipolar disorder. Finally, we will discuss the challenges of translating these findings into clinical practice.

A scoping review aimed at investigating the use of machine learning techniques for the detection of BD found that the majority of the studies used classification models (eg, random forest), included a sample size of fewer than 300 individuals, and included clinical data in the model [5]. The potential of new machine learning methods to better understand factors associated with misdiagnosis was exemplified in a recent study that reported a misdiagnosis rate of 50.97% [14]. In this study, any mismatch between the self-reported diagnosis and the clinical interview diagnosis was considered a misdiagnosis. The investigators used machine learning techniques to identify the predictors of misdiagnosis, and the mean accuracy of the predictive model was 70% [14]. This study showed that more severe depressive symptoms and unstable self-image were the strongest predictors of mood disorder misdiagnosis among the 1045 variables evaluated [14]. These results may be explained by the fact that patients usually seek treatment when they are severely depressed and that they may be underreporting hypomanic symptoms during a severe depressive episode. Consequently, they might be misdiagnosed with major depression instead of receiving the correct diagnosis of BD.

Another recent study highlights how a correct diagnosis may be made earlier. The clinical predictors of BD were described in a large birth cohort study, including 3748 subjects assessed at birth and 11, 15, 18, and 22 years of age [6]. The study used machine learning techniques and showed that the presence of suicide risk, generalized anxiety disorder, parental physical abuse, and financial problems at 18 years of age were the strongest predictors for a BD diagnosis at 22 years of age, with a balanced accuracy of 75% [6]. Additionally, the high-risk subgroup of BD showed a high frequency of drug use and depressive symptoms [6].

Several machine learning studies used digital phenotyping to classify BD and unipolar disorder [15,16]. In one study, daily smartphone-based self-assessments of mood and same-time passively collected smartphone data on smartphone usage were assessed for 6 months [15]. The main findings indicate that patients with BD, in an euthymic state, had a lower number of incoming phone calls per day compared to patients with unipolar depression also experiencing euthymia. In addition, during depressive states, patients with BD had a lower number of incoming and outgoing phone calls per day as compared with patients with unipolar depression [15]. BD was classified with an area under the curve (AUC) of 0.84 (overall; when mood state was not taken into consideration), 0.86 (during a depressive state), and 0.87 (during a euthymic state) in this study. However, when applying the leave-one-out cross-validation approach, the AUC for all models reduced (AUC=0.48 for the overall model, AUC=0.42 for the depressive state model, and AUC=0.46 for the euthymic state model), indicating that changes in combined smartphone-based data were highly individual [15]. Another digital phenotyping study using the mindLAMP app to collect geolocation, accelerometer, and screen/state reported an AUC of 0.62 for classifying patients with MDD or Bipolar I/II disorders [16]

The differing results noted above are common in machine learning research, especially where the underlying data and technology differ between studies. A task force discussing the scientific literature related to machine learning and big data–based studies showed that machine learning studies have included a variety of data to predict BD, including neuroimaging, genetics, electroencephalogram, neurophysiological data, blood biomarkers, text, facial expressions, speech, and ecological momentary assessments [17]. The task force emphasized that some limitations should be addressed to allow these findings to be translated to clinical practice, in particular the lack of external validation of the predictive models [17].

Digital Phenotyping to Detect Mood Symptoms and Mood Episodes in BD

The development of digital phenotyping is quickly evolving and expanding in the field of BD. Digital phenotyping involves collecting data (eg, location, activity, sleep, speech patterns), typically from smartphones, to monitor behavior, cognition, and mood [18]. Digital phenotyping may help facilitate the early detection of potentially problematic mood changes, therefore facilitating early intervention. Before digital phenotyping can be applied in usual care for BD, we must develop an understanding of which of the multitude of digital data collected by smartphones and wearable sensors can reliably and validly detect early warning signs of mood episodes. Importantly, while several studies have shown that digital phenotyping is a promising technique, it faces several challenges that need to be robustly addressed [19], which will be discussed in this section.

A systematic review describing the evidence about the use of portable digital tools for detecting BD, mood states, and mood symptoms found 62 studies assessing it in terms of four main areas: (1) smartphone apps designed to collect active (eg, mood self-assessments) or passive (eg, recording geolocation, step counts, call and text logs, sleep, etc) data; (2) wearable sensors for the monitoring of electrocardiography and actigraphy; (3) audio-visual recordings for the analysis of speech or facial expressions and upper body movements; and (4) multimodal tools, combining 2 or more of the above [7]. Two-thirds of the studies included applied machine learning approaches to classify BD versus healthy controls, to identify mood states, or to predict the severity of symptoms. They achieved mixed results, yielding fair to excellent classification performances, with accuracy globally ranging from 60% to 97% [7]. A recent review assessing the application of digital tools for major depressive episodes described the following digital phenotype for BD: (1) speech alterations during a depressive episode, including decreased speech pause and reduced fundamental frequency, while these speech features were increased during a hypomanic episode; (2) irrespective of the mood state, heart rate variability was reduced, but the change in heart rate variability in the interepisodic phases remained unclear, and (3) an electrodermal hypoactivity in a depressive episode was reported, which increased when patients were euthymic [20].

Regarding the challenges related to digital phenotyping, it is important to note that any data collected using digital devices is prone to bias and needs to be standardized to ensure accuracy not only across populations but also across different devices. Moreover, concerns relating to privacy, ethics, data security, and consent must be addressed. User comfort in sharing data differs depending on the data type (eg, users are more comfortable sharing health data than personal data such as location, communication logs, and social activity) and the recipient (eg, users have greater comfort sharing data directly with clinicians than having this entered into their electronic health record), and this may impact willingness to use digital phenotyping platforms [21]. As user engagement is essential for the success of any digital phenotyping tools [22], it is necessary to account for discrepancies in access, equity, and distribution of resources. Finally, more in-depth longitudinal studies are required to ascertain the relationship between biomarkers and long-term outcomes of health and well-being.

Smartphone-Based Interventions for BD

Psychosocial therapies and education in self-management strategies can improve outcomes in BD [23] and are recommended complements to pharmacological interventions in guidelines for BD treatment. However, access to these forms of care remains suboptimal, with less than 50% of individuals in treatment for BD receiving therapy with a psychologist, social worker, or self-help support group [24]. Smartphone apps have the potential to provide psychoeducation and facilitate several of the core components of psychosocial therapies (eg, self-monitoring, detecting and responding to mood episodes, stabilizing daily routines, improving emotion regulation, encouraging medication adherence, etc) [25]. Encouragingly, individuals with BD report high levels of access to smartphones and a willingness to receive psychosocial interventions via apps [26,27]. Several app-facilitated interventions have been developed and evaluated for BD, variously integrating self-monitoring, psychoeducation, cognitive-behavior therapy, and skills training, and targeting both symptoms and patient-valued outcomes such as functioning and quality of life [9]. However, evidence for their feasibility and efficacy is still preliminary, and interventions are yet to fully leverage the capabilities of apps for intervention personalization.

Two recent systematic reviews and meta-analyses investigated the role of smartphone-based interventions to improve clinical outcomes in BD and found conflicting results [8,9]. Liu et al [8] included 10 studies in their systematic review (7 randomized controlled trials and 3 single-arm trials) and concluded that smartphone-based interventions were effective in reducing manic and depressive symptoms in between-group (compared to controls) and within-group (comparing symptoms from baseline to postintervention in the intervention group). Anmella et al [9] included 13 studies in their qualitative synthesis of the findings and 5 studies in their meta-analysis. The meta-analyses comparing the pre-post change in depressive and (hypo)manic symptom severity, functioning, quality of life, and perceived stress between smartphone interventions and controls did not reach statistical significance for any outcome assessed [9]. The potential explanation for the conflicting findings is that the eligibility criteria were different between both studies. The most important difference is the fact that Liu et al [8] included not only smartphone-based apps but also phone calls from specialists to facilitate therapy and website interventions in the intervention group, while Anmella et al [9] excluded interventions not delivered through smartphones (eg, exclusive of phone calls, phone messaging, only SMS text messaging, or computer-delivered interventions) from the intervention group. Another difference between both studies is that Liu et al [8] did not restrict the inclusion criteria to individuals with BD and included a few studies that recruited a more heterogeneous population (eg, serious mental illness, mood disorders), while Anmella et al [9] only included studies where the participants were diagnosed with BD.

Given the heterogeneity of BD both between and within individuals, effective psychotherapy involves appropriately tailoring intervention content and delivery to the challenges and goals of a specific individual at a specific time. However, apps are yet to fully capitalize on the potential of smartphones to personalize intervention delivery in response to changes in clinical state. One app program, SIMPLe, personalizes content using ecological momentary assessment to identify potential prodromal mood changes and adapts the delivery of psychoeducation messages in response [28,29]. To advance our understanding of how to tailor just-in-time adaptive interventions for BD, microrandomized trials can be used to evaluate the immediate impact of diverse types of intervention prompts. For example, an evaluation of mobile acceptance and commitment therapy used this trial design to evaluate different categories of intervention and found that awareness-focused prompts paradoxically increased symptoms [30]. Beyond the clinical utility of personalization, this feature is also highly prioritized by individuals with BD themselves [31], who have expressed a desire for apps that make meaningful use of their data to customize intervention delivery and facilitate proactive support.

Improving the Dissemination and Uptake of Apps for BD

Although research-led studies have developed and evaluated mobile apps for BD, the dissemination and uptake of these apps in real-world contexts must be considered to reach the target population and maximize their impact. A recent web-based survey investigating the use of mobile apps to support mood and sleep self-management among individuals with BD found that 41.6% of participants reported using a self-management app related to mood and/or sleep [32]. The most nominated app for mood monitoring was Daylio, and the most reported app for sleep monitoring was Fitbit. Since these apps are designed to support the public with well-being concerns, this raises questions about why apps specifically designed for BD are not reaching this population. Two possibilities emerge: (1) apps designed for BD are not sufficiently acceptable or engaging in the eyes of the target audience and (2) individuals with BD may not be adequately supported to select the app that is best suited to their needs. To facilitate research-led apps reaching and impacting users with BD, we must consider their ability to create and sustain user engagement. Further, we must consider effective dissemination pathways, targeting both patients and the health care providers involved in the provision of care to people with BD (eg, clinicians, nurses, allied health professionals, and case managers).

Poor engagement is endemic to mental health apps in general, extending beyond just those aimed at BD [33], with most users of publicly available apps disengaging within 30 days. Specific to BD, a systematic review showed that adherence data in research trials was infrequently reported; of the 13 studies providing engagement data, the activity rate ranged from 58% to 91% [34]. The failure to consider the needs and goals of the targeted population likely contributes to startlingly poor levels of uptake and adherence. Involving users in the design of apps can help ensure their design, content, and feature selection are relevant, acceptable, and engaging. However, a recent review investigated the level of user involvement in the design of self-monitoring apps for BD [35] and found that 36% of the apps did not mention user involvement in the design, while 9% reported low, 36% reported medium, and 18% reported high user involvement. This review highlights the importance of including an appropriate sample size capable of adequately capturing users' needs so that technology can be better designed. Finally, it is recommended that users are involved early in the design process, and their involvement should not be limited solely to the design but also to all aspects of the research, ensuring end-to-end involvement. Case studies of apps using a co-design framework include the quality of life–focused LiveWell and PolarUs apps, both of which consulted people with BD throughout development [36,37]. Figure 2 depicts 2 screens from the PolarUs app: on the home screen (left image), users are prompted to engage in quality of life, sleep, and mood self-monitoring, and are provided with relevant resources [37]. Users can review their self-monitoring data over time (right image). Individuals with BD provided input into the app design (including icons, color scheme, and layout), navigation, features, and content.

Looking ahead, as more apps for BD are developed and made available to the public, patients with BD and health care providers will likely require support to navigate the digital health landscape, as research-led apps will compete for attention with commercial offerings that may have limitations in their privacy protections and efficacy [38]. Educational interventions to enhance digital health literacy may help individuals with BD to select the appropriate apps for their self-management goals. While in general, levels of digital health literacy are comparable for people with BD to the general population, a study found that individuals with BD who are younger, have completed less education, or are less familiar with mental health apps may require extra support to safely and productively navigate web-based health resources [39]. Recent steps have been taken to address the needs of these groups: a brief, informational video describing strategies to select safe, effective, and engaging mental health apps for BD was created [40], incorporating the perspectives of people with lived experience in the script and design. A still image from this video is presented in Figure 3 [40]. This resource was later expanded upon to create a web-based module [41], depicted in Figure 4, which contains additional information and resources to support people in evaluating app privacy policies, inclusion of evidence-based strategies for BD, and motivational techniques. Other resources like mindapps.org can help facilitate informed decision-making about mental health apps [38].

Health care providers are an important source of information and advice on smartphone apps, yet a web survey found that only 48.8% of health care providers reported discussing or recommending health apps to patients with BD [42]. Most of the apps recommended were related to core symptoms of BD, including mood and sleep. Among the health care providers who did not discuss health apps with patients with BD (51.2%), the predominant reason mentioned was the lack of familiarity with credible and suitable apps tailored for BD. The resources discussed above are also appropriate for use by clinicians wishing to learn more about appropriate and effective apps for BD [38,41]. These findings emphasize the importance of providing training aimed at increasing clinician self-efficacy in using mobile apps with patients, a strategy that should be considered by researchers developing new mHealth tools.

Figure 2. Interface of the PolarUs app (reproduced from Michalak et al [37], which is published under Creative Commons Attribution 4.0 International License [43]).

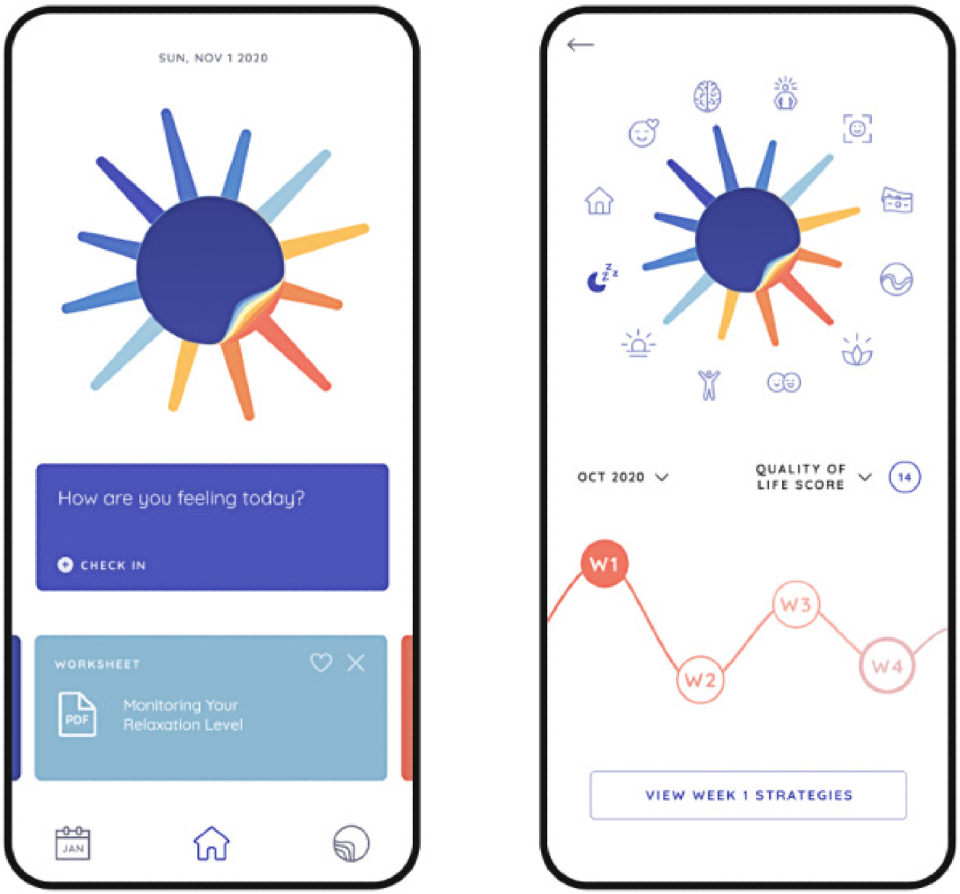
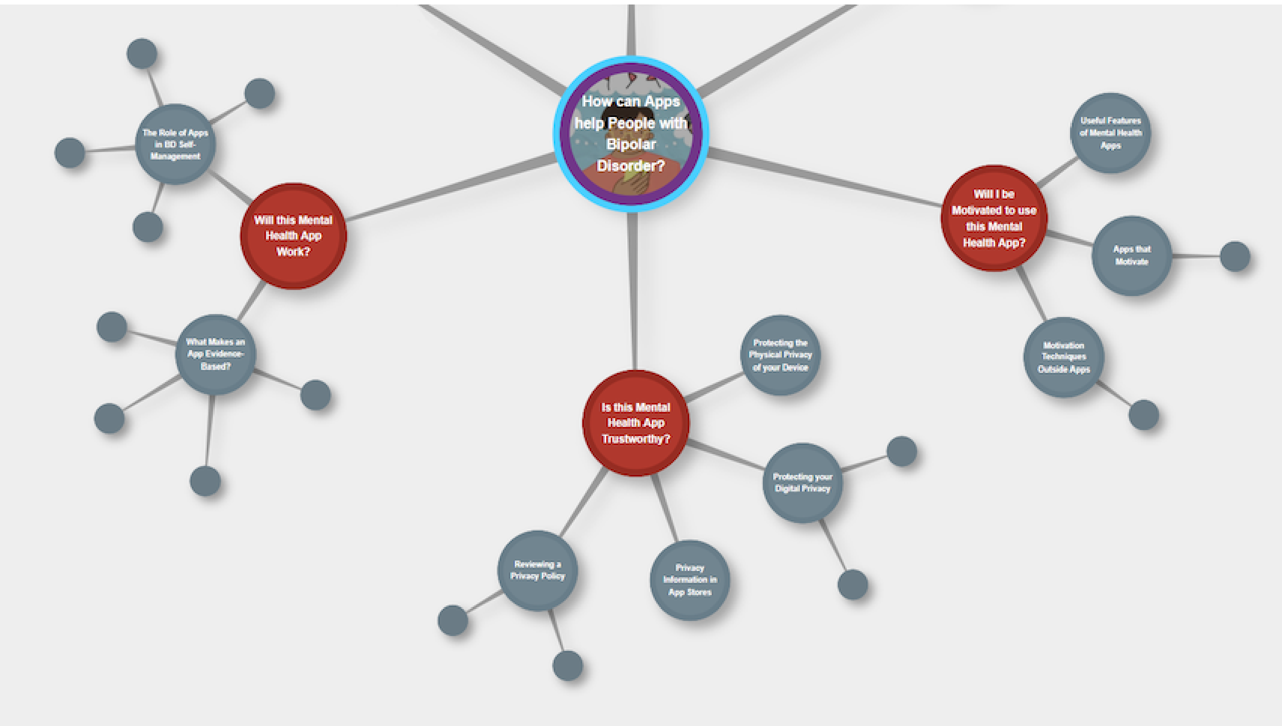


Figure 3. Choosing a bipolar disorder app that works for you (reproduced from [40], with permission from Erin Michalak).

Figure 4. Additional information and resources to support people in evaluating app privacy policies, inclusion of evidence-based strategies for BD, and motivational techniques (reproduced from [41], with permission from Erin Michalak). BD: bipolar disorder.



Conclusion

The evidence available to date indicates that digital technologies may help in the early detection of BD and mood episodes, as well as in enhancing treatment, improving health outcomes and consequently promoting a better prognosis for individuals with BD. However, there are important limitations that need to be addressed before these technologies can be translated to clinical practice, including the following: (1) external validation of the machine learning models developed to date, (2) need for well-designed prospective cohort studies to validate findings about digital phenotyping and early detection of BD and mood symptoms, (3) involvement of individuals with lived experience in the development of mobile apps, and (4) dissemination of the available technology among health care providers and directly to people with BD. Finally, adequately powered randomized controlled trials are still needed to evaluate the efficacy of mental health apps for BD. Additionally, there is a need to advance our understanding of how to tailor app-based interventions based on the valuable insights generated by digital phenotyping.

Authors' Contributions

TdAC contributed to writing the original draft. TdAC, EM, SK, and JT contributed to the conceptualization, writing, review, and editing of the manuscript. EM and JT contributed to supervision.

Conflicts of Interest

TdAC is a scientific editor at JMIR Publications, EM is an associate editor for *JMIR Mental Health*, SK is a managing editor at JMIR Publications, and JT is the editor-in-chief of *JMIR Mental Health*.

References

1. Merikangas KR, Jin R, He J, Kessler RC, Lee S, Sampson NA, et al. Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. Arch Gen Psychiatry 2011 Mar;68(3):241-251 [FREE Full text] [doi: 10.1001/archgenpsychiatry.2011.12] [Medline: 21383262]
2. Solmi M, Radua J, Olivola M, Croce E, Soardo L, Salazar de Pablo G, et al. Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. Mol Psychiatry 2022 Jan;27(1):281-295 [FREE Full text] [doi: 10.1038/s41380-021-01161-7] [Medline: 34079068]
3. Frey BN, Vigod S, de Azevedo Cardoso T, Librenza-Garcia D, Favotto L, Perez R, et al. The early burden of disability in individuals with mood and other common mental disorders in Ontario, Canada. JAMA Netw Open 2020 Oct 01;3(10):e2020213 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.20213] [Medline: 33104205]

4. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022 Feb;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
5. Jan Z, Ai-Ansari N, Mousa O, Abd-Alrazaq A, Ahmed A, Alam T, et al. The role of machine learning in diagnosing bipolar disorder: scoping review. J Med Internet Res 2021 Nov 19;23(11):e29749 [FREE Full text] [doi: 10.2196/29749] [Medline: 34806996]
6. Rabelo-da-Ponte FD, Feiten JG, Mwangi B, Barros FC, Wehrmeister FC, Menezes AM, et al. Early identification of bipolar disorder among young adults - a 22-year community birth cohort. Acta Psychiatr Scand 2020 Dec;142(6):476-485. [doi: 10.1111/acps.13233] [Medline: 32936930]
7. Saccaro LF, Amatori G, Cappelli A, Mazziotti R, Dell'Osso L, Rutigliano G. Portable technologies for digital phenotyping of bipolar disorder: a systematic review. J Affect Disord 2021 Dec 01;295:323-338. [doi: 10.1016/j.jad.2021.08.052] [Medline: 34488086]
8. Liu J, Xu K, Zhu G, Zhang Q, Li X. Effects of smartphone-based interventions and monitoring on bipolar disorder: a systematic review and meta-analysis. World J Psychiatry 2020 Nov 19;10(11):272-285 [FREE Full text] [doi: 10.5498/wjp.v10.i11.272] [Medline: 33269223]
9. Anmella G, Faurholt-Jepsen M, Hidalgo-Mazzei D, Radua J, Passos IC, Kapczinski F, et al. Smartphone-based interventions in bipolar disorder: systematic review and meta-analyses of efficacy. A position paper from the International Society for Bipolar Disorders (ISBD) Big Data Task Force. Bipolar Disord 2022 Sep;24(6):580-614 [FREE Full text] [doi: 10.1111/bdi.13243] [Medline: 35839276]
10. Scott J, Graham A, Yung A, Morgan C, Bellivier F, Etain B. A systematic review and meta-analysis of delayed help-seeking, delayed diagnosis and duration of untreated illness in bipolar disorders. Acta Psychiatr Scand 2022 Nov;146(5):389-405. [doi: 10.1111/acps.13490] [Medline: 36018259]
11. Shen H, Zhang L, Xu C, Zhu J, Chen M, Fang Y. Analysis of misdiagnosis of bipolar disorder in an outpatient setting. Shanghai Arch Psychiatry 2018 Apr 25;30(2):93-101 [FREE Full text] [doi: 10.11919/j.issn.1002-0829.217080] [Medline: 29736129]
12. Kennedy SH, Lam RW, McIntyre RS, Tourjman SV, Bhat V, Blier P, et al. Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 clinical guidelines for the management of adults with major depressive disorder: section 3. Pharmacological treatments. Can J Psychiatry 2016 Sep;61(9):540-560 [FREE Full text] [doi: 10.1177/0706743716659417] [Medline: 27486148]
13. Yatham LN, Kennedy SH, Parikh SV, Schaffer A, Bond DJ, Frey BN, et al. Canadian Network for Mood and Anxiety Treatments (CANMAT) and International Society for Bipolar Disorders (ISBD) 2018 guidelines for the management of patients with bipolar disorder. Bipolar Disord 2018 Mar;20(2):97-170 [FREE Full text] [doi: 10.1111/bdi.12609] [Medline: 29536616]
14. Benacek J, Lawal N, Ong T, Tomasik J, Martin-Key NA, Funnell EL, et al. Identification of predictors of mood disorder misdiagnosis and subsequent help-seeking behavior in individuals with depressive symptoms: gradient-boosted tree machine learning approach. JMIR Ment Health 2024 Jan 11;11:e50738 [FREE Full text] [doi: 10.2196/50738] [Medline: 38206660]
15. Faurholt-Jepsen M, Rohani DA, Busk J, Tønning ML, Frost M, Bardram JE, et al. Using digital phenotyping to classify bipolar disorder and unipolar disorder - exploratory findings using machine learning models. Eur Neuropsychopharmacol 2024 Feb 03;81:12-19. [doi: 10.1016/j.euroneuro.2024.01.003] [Medline: 38310716]
16. Langholm C, Breitinger S, Gray L, Goes F, Walker A, Xiong A, et al. Classifying and clustering mood disorder patients using smartphone data from a feasibility study. NPJ Digit Med 2023 Dec 21;6(1):238 [FREE Full text] [doi: 10.1038/s41746-023-00977-7] [Medline: 38129571]
17. Passos IC, Ballester PL, Barros RC, Librenza-Garcia D, Mwangi B, Birmaher B, et al. Machine learning and big data analytics in bipolar disorder: a position paper from the International Society for Bipolar Disorders Big Data Task Force. Bipolar Disord 2019 Nov;21(7):582-594. [doi: 10.1111/bdi.12828] [Medline: 31465619]
18. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA 2017 Oct 03;318(13):1215-1216. [doi: 10.1001/jama.2017.11295] [Medline: 28973224]
19. Ortiz A, Maslej MM, Husain MI, Daskalakis ZJ, Mulsant BH. Apps and gaps in bipolar disorder: a systematic review on electronic monitoring for episode prediction. J Affect Disord 2021 Dec 01;295:1190-1200. [doi: 10.1016/j.jad.2021.08.140] [Medline: 34706433]
20. Ettore E, Müller P, Hinze J, Benoit M, Giordana B, Postin D, et al. Digital phenotyping for differential diagnosis of major depressive episode: narrative review. JMIR Ment Health 2023 Jan 23;10:e37225 [FREE Full text] [doi: 10.2196/37225] [Medline: 36689265]
21. Nicholas J, Shilton K, Schueller SM, Gray EL, Kwasny MJ, Mohr DC. The role of data type and recipient in individuals' perspectives on sharing passively collected smartphone data for mental health: cross-sectional questionnaire study. JMIR Mhealth Uhealth 2019 Apr 05;7(4):e12578 [FREE Full text] [doi: 10.2196/12578] [Medline: 30950799]
22. Currey D, Torous J. Increasing the value of digital phenotyping through reducing missingness: a retrospective review and analysis of prior studies. BMJ Ment Health 2023 Feb;26(1) [FREE Full text] [doi: 10.1136/bmjment-2023-300718] [Medline: 37197799]

23. Miklowitz DJ, Efthimiou O, Furukawa TA, Scott J, McLaren R, Geddes JR, et al. Adjunctive psychotherapy for bipolar disorder: a systematic review and component network meta-analysis. JAMA Psychiatry 2021 Feb 01;78(2):141-150 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.2993] [Medline: 33052390]
24. Lembke A, Miklowitz DJ, Otto MW, Zhang H, Wisniewski SR, Sachs GS, et al. Psychosocial service utilization by patients with bipolar disorders: data from the first 500 participants in the Systematic Treatment Enhancement Program. J Psychiatr Pract 2004 Mar;10(2):81-87. [doi: 10.1097/00131746-200403000-00002] [Medline: 15330403]
25. Miklowitz DJ, Scott J. Psychosocial treatments for bipolar disorder: cost-effectiveness, mediating mechanisms, and future directions. Bipolar Disord 2009 Jun;11 Suppl 2:110-122. [doi: 10.1111/j.1399-5618.2009.00715.x] [Medline: 19538690]
26. Hidalgo-Mazzei D, Nikolova V, Kitchen S, Young A. Internet-connected devices ownership, use and interests in bipolar disorder: from desktop to mobile mental health. Digital Psychiatry 2019 May 28;2(1):1-7. [doi: 10.1080/2575517X.2019.1616476]
27. Daus H, Kislicyn N, Heuer S, Backenstrass M. Disease management apps and technical assistance systems for bipolar disorder: investigating the patients´ point of view. J Affect Disord 2018 Dec 15;229:351-357. [doi: 10.1016/j.jad.2017.12.059] [Medline: 29331693]
28. Hidalgo-Mazzei D, Mateu A, Reinares M, Murru A, Del MBC, Varo C, et al. Psychoeducation in bipolar disorder with a SIMPLe smartphone application: feasibility, acceptability and satisfaction. J Affect Disord 2016 Aug;200:58-66. [doi: 10.1016/j.jad.2016.04.042] [Medline: 27128358]
29. Hidalgo-Mazzei D, Reinares M, Mateu A, Nikolova VL, Bonnín CDM, Samalin L, et al. OpenSIMPLe: a real-world implementation feasibility study of a smartphone-based psychoeducation programme for bipolar disorder. J Affect Disord 2018 Dec 01;241:436-445. [doi: 10.1016/j.jad.2018.08.048] [Medline: 30145515]
30. Cochran A, Maronge JM, Victory A, Hoel S, McInnis MG, Thomas EB. Mobile acceptance and commitment therapy in bipolar disorder: microrandomized trial. JMIR Ment Health 2023 Apr 20;10:e43164 [FREE Full text] [doi: 10.2196/43164] [Medline: 37079363]
31. Morton E, Nicholas J, Lapadat L, O'Brien HL, Barnes SJ, Poh C, et al. Use of smartphone apps in bipolar disorder: an international web-based survey of feature preferences and privacy concerns. J Affect Disord 2021 Dec 01;295:1102-1109. [doi: 10.1016/j.jad.2021.08.132] [Medline: 34706421]
32. Morton E, Nicholas J, Yang L, Lapadat L, Barnes SJ, Provencher MD, et al. Evaluating the quality, safety, and functionality of commonly used smartphone apps for bipolar disorder mood and sleep self-management. Int J Bipolar Disord 2022 Apr 04;10(1):10 [FREE Full text] [doi: 10.1186/s40345-022-00256-6] [Medline: 35368207]
33. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
34. Patoz M, Hidalgo-Mazzei D, Pereira B, Blanc O, de Chazeron I, Murru A, et al. Patients' adherence to smartphone apps in the management of bipolar disorder: a systematic review. Int J Bipolar Disord 2021 Jun 03;9(1):19 [FREE Full text] [doi: 10.1186/s40345-021-00224-6] [Medline: 34081234]
35. Majid S, Reeves S, Figueredo G, Brown S, Lang A, Moore M, et al. The extent of user involvement in the design of self-tracking technology for bipolar disorder: literature review. JMIR Ment Health 2021 Dec 20;8(12):e27991 [FREE Full text] [doi: 10.2196/27991] [Medline: 34931992]
36. Jonathan GK, Dopke CA, Michaels T, Bank A, Martin CR, Adhikari K, et al. A smartphone-based self-management intervention for bipolar disorder (LiveWell): user-centered development approach. JMIR Ment Health 2021 Apr 12;8(4):e20424 [FREE Full text] [doi: 10.2196/20424] [Medline: 33843607]
37. Michalak EE, Barnes SJ, Morton E, O'Brien HL, Murray G, Hole R, et al. Supporting self-management and quality of life in bipolar disorder with the PolarUs App (Alpha): protocol for a mixed methods study. JMIR Res Protoc 2022 Aug 04;11(8):e36213 [FREE Full text] [doi: 10.2196/36213] [Medline: 35925666]
38. Lagan S, Ramakrishnan A, Lamont E, Ramakrishnan A, Frye M, Torous J. Digital health developments and drawbacks: a review and analysis of top-returned apps for bipolar disorder. Int J Bipolar Disord 2020 Dec 01;8(1):39 [FREE Full text] [doi: 10.1186/s40345-020-00202-4] [Medline: 33259047]
39. Morton E, Ho K, Barnes SJ, Michalak EE. Digital health literacy in bipolar disorder: international web-based survey. JMIR Ment Health 2021 Oct 19;8(10):e29764 [FREE Full text] [doi: 10.2196/29764] [Medline: 34665143]
40. Choosing a bipolar disorder app that works for you. YouTube. 2023 Oct 19. URL: https://youtu.be/q2SAeYr_pTc?si=d1SOzotm5FzrLIz3 [accessed 2024-03-20]
41. How can apps help people with bipolar disorder? CREST.BD. 2023. URL: https://crestbd.tapestry-tool.com/tapestry/reach-project/ [accessed 2024-03-20]
42. Morton E, Torous J, Murray G, Michalak EE. Using apps for bipolar disorder - an online survey of healthcare provider perspectives and practices. J Psychiatr Res 2021 May;137:22-28. [doi: 10.1016/j.jpsychires.2021.02.047] [Medline: 33647725]
43. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2024-03-26]

Abbreviations

AUC: area under the curve
BD: bipolar disorder
MDD: major depressive disorder
mHealth: mobile health

Edited by T Leung; submitted 20.03.24; this is a non–peer-reviewed article;accepted 25.03.24; published 01.04.24.

Please cite as:
de Azevedo Cardoso T, Kochhar S, Torous J, Morton E
Digital Tools to Facilitate the Detection and Treatment of Bipolar Disorder: Key Developments and Future Directions
JMIR Ment Health 2024;11:e58631
URL: https://mental.jmir.org/2024/1/e58631
doi:10.2196/58631
PMID:38557724

Comparison of Use Rates of Telehealth Services for Substance Use Disorder During and Following COVID-19 Safety Distancing Recommendations: Two Cross-Sectional Surveys

Adrijana Pusnik[1,*], BSN, RN, MPH; Bryan Hartzler[2,*], PhD; Olivia Vjorn[1,*], MS; Beth A Rutkowski[3,*], MPH; Michael Chaple[4,*], PhD; Sara Becker[5,*], PhD; Thomas Freese[3,*], PhD; Maureen Nichols[6,*], BA; Todd Molfenter[1,*], PhD

1

2

3

4

5

6

*all authors contributed equally

Corresponding Author:
Adrijana Pusnik, BSN, RN, MPH

Abstract

Background: The COVID-19 social distancing guidelines resulted in a dramatic transition to telephone and video technologies to deliver substance use disorder (SUD) treatment. Before COVID-19, the question was "Will telehealth ever take hold for SUD services?" Now that social distancing guidelines have been lifted, the question is "Will telehealth remain a commonly used care modality?"

Objective: The principal purpose of this investigation was to examine the extent to which telehealth use in SUD service settings persisted following the lifting of COVID-19 safety distancing recommendations. Additionally, the study aimed to explore practitioners' perceptions of telehealth convenience and value after its regular implementation during the pandemic. Specifically, the goal of this study was to compare telehealth activity between time intervals: May-August 2020 (during peak COVID-19 safety distancing recommendations) and October-December 2022 (following discontinuation of distancing recommendations). Specifically, we compared (1) telehealth technologies and services, (2) perceived usefulness of telehealth, (3) ease of use of telephone- and video-based telehealth services, and (4) organizational readiness to use telehealth.

Methods: An online cross-sectional survey consisting of 108 items was conducted to measure the use of telehealth technologies for delivering a specific set of SUD services in the United States and to explore the perceived readiness for use and satisfaction with telephonic and video services. The survey took approximately 25 - 35 minutes to complete and used the same 3 sets of questions and 2 theory-driven scales as in a previous cross-sectional survey conducted in 2020. Six of 10 Regional Addiction Technology Transfer Centers funded by the Substance Abuse and Mental Health Services Administration distributed the survey in their respective regions, collectively spanning 37 states. Responses of administrators and clinicians (hereafter referred to as staff) from this 2022 survey were compared to those obtained in the 2020 survey. Responses in 2020 and 2022 were anonymous and comprised two separate samples; therefore, an accurate longitudinal model could not be analyzed.

Results: A total of 375 staff responded to the 2022 survey (vs 457 in 2020). Baseline organizational characteristics of the 2022 sample were similar to those of the 2020 sample. Phone and video telehealth utilization rates remained greater than 50% in 2022 for screening and assessment, case management, peer recovery support services, and regular outpatient services. The perceived usefulness of phone-based telehealth was higher in 2022 than in 2020 (mean difference [MD] −0.23; P=.002), but not for video-based telehealth (MD −0.12; P=.13). Ease of use of video-based telehealth was perceived as higher in 2022 than in 2020 (MD−0.35; P<.001), but no difference was found for phone-based telehealth (MD −0.12; P=.11). From the staff's perspective, patients had greater readiness for using telehealth via phone than video, but the staff perceived their personal and organizational readiness for using telehealth as greater for video-based than for phone-based telehealth.

Conclusions: Despite lower telephone and video use in 2022 for telehealth services than in 2020, both modalities continue to be perceived positively. Future research may further determine the relative cost and clinical effectiveness of video-based services and thereby help to address some sources of the noted challenges to implementation by SUD organizations.

(JMIR Ment Health 2024;11:e52363) doi:10.2196/52363

KEYWORDS
telehealth; COVID-19; substance use disorders; telephone counseling; video counseling

Introduction

Use of telehealth, defined as the remote provision of health services via telecommunications, has proliferated over the past decade [1]; however, its uptake has been complex and inconsistent [2]. Despite compelling evidence that telehealth services result in equal or better clinical effectiveness and patient satisfaction relative to in-person services across multiple meta-analytic reviews [3-7], telehealth use had been extremely limited in the delivery of services for substance use disorders (SUDs). One national estimate in 2018 cited the use of telehealth in only 16% of SUD treatment programs [8]. Following the onset of COVID-19 in 2020, the implementation of corresponding safety distancing recommendations created what some have termed a disruptive innovation scenario [9] wherein the use of telephone- and video-based telehealth services was urgently supported at a federal level and began to occur among treatment providers at unprecedented levels [10].

The unforeseen circumstances of the global pandemic brought an opportunity for greater understanding of telehealth use among treatment organizations offering SUD services. Among the encouraging findings reported in the wake of new COVID-19 safety policies was the increased use of telehealth in engaging patients to receive medications for opioid use disorders, which was associated with greater treatment retention and a reduced chance of medically treated overdose [11]. Similarly, telehealth services in the midst of COVID-19 safety distancing recommendations left a favorable impression among clinical and administrative staff in terms of the clinical benefit and reach of vulnerable populations [12-14]. An additional related benefit was the perception of telehealth's favorable influence in both reducing SUD stigma and increasing support for patients in active recovery [15,16].

Concerns with telehealth also emerged during COVID-19. Multiple reports noted a greater acceptability of and preference for in-person services among some staff and patient groups [5,17], while others detailed concerns regarding the lack of human contact, confidentiality, and data security when using telehealth [18]. Perhaps unsurprisingly, the eventual loosening of COVID-19 safety distancing recommendations was reportedly associated with a significant reduction in telehealth use [10]. Taken together, a muddied picture of equivocal and rapidly shifting findings casts some doubt on what rates of telehealth use may be expected as treatment organizations proceed through the post-COVID-19 era.

Based on a national sampling of administrators and clinicians from 457 SUD treatment organizations in the United States in the months immediately following the institution of COVID-19 safety distancing recommendations (May to August 2020), a group affiliated with the Substance Abuse and Mental Health Services Administration (SAMHSA) Addiction Technology Transfer Center (ATTC) Network reported several findings concerning telehealth utilization by the addiction treatment community [19]. Foremost among these were: (1) a pattern of extensive telehealth (via telephone and video) utilization by 73% of the SUD treatment organizations, with its most prevalent application in screening and assessment intake (79%) and general outpatient services (82%); (2) strong organizational readiness to use, and satisfaction with, both telephone- and video-based technologies, albeit with the former technology deemed more accessible among patients with SUDs; and (3) validation of the Technology Acceptance Model (TAM) [20] in SUD service settings such that, for both telephone- and video-based technologies, perceived usefulness and ease of use predicted the organizational intent for continued utilization following the discontinuation of COVID-19 safety protocols. The current work offers a follow-up investigation of this 2020 survey using a similar cross-sectional national sampling of personnel from SUD treatment organizations from October to December 2022. The primary aim of this study was to examine the extent to which telehealth use in SUD service settings persisted following discontinuation of COVID-19 safety distancing recommendations. Specifically, we compared use rates for specific telehealth technologies and services, the perceived usefulness of telehealth, ease of use of telephone- and video-based telehealth services, and organizational readiness to use telehealth between the two time intervals: May-August 2020 (peak COVID-19 distancing recommendations) and October-December 2022 [1] (following discontinuation of distancing recommendations).

Methods

Study Design

The voluntary cross-sectional online survey (described in further detail below in the Survey Instrument subsection) to measure the use of telehealth technologies for delivering a specific set of SUD services and to explore the perceived readiness to use and satisfaction with telephonic and video services was developed using similar questions as our prior telehealth survey distributed during the peak of social distancing recommendations [1]. Six of 10 regional ATTCs distributed the survey to administrators, clinicians, and recovery personnel in their respective regions. The distribution collectively spanned 37 states. SAMHSA-funded ATTCs support the workforce for addiction treatment and recovery via regional ATTCs that correspond to the 10 regional offices of the US Department of Health and Human Services. Four regional ATTCs chose not to participate, citing concerns about potential survey fatigue among their stakeholders.

Data Collection

The 2022 survey was distributed on October 3, 2022, and data collection continued until January 6, 2023; the survey took approximately 8 - 15 minutes to complete [21]. Regional ATTCs distributed survey links to administrators and clinicians at substance use treatment organizations (hereafter referred to as "staff" for simplicity) via their regional mailing lists.

Survey Instrument

The survey instrument had 104 questions that included 3 sets of study-specific questions, designed to mimic the questions used in the 2020 survey, followed by 2 theory-driven scales. The method used to test survey usability and validity is described in Molfenter et al [19]. The first set of questions asked about the organization where the respondent worked. Specifically, respondents were asked to select their organization type (ie, health system, opioid treatment program, recovery community organization, and specialty addiction treatment providers such as nonopioid treatment programs), organization location (ie, tribal reservation, rural, small city, suburban, and urban), and organizational role (ie, administrators and personnel providing treatment and/or recovery services).

The second set of questions assessed the organization's use of the following technologies via binary yes/no variables: computerized screening and assessments, mobile app(s) during recovery, mobile app(s) during treatment, web portal for scheduling appointments, secure chats for recovery support sessions, text appointment reminders, text motivational messages, and video-based therapy to provide buprenorphine.

The third set of questions asked respondents which methods (telephone, video, or in person) were used for the following services: screening and assessment, buprenorphine therapy, case management, intensive outpatient treatment, peer recovery support, regular outpatient treatment, and residential counseling sessions.

The 2 theory-driven scales followed the above questions based on the Technology Acceptance Scale [19], which includes two subscales from the TAM: ease of use and perceived usefulness [20]. The ease-of-use scale assesses the ease of learning, customizing, and using a technology. Perceived usefulness assesses the extent to which the technology is perceived to enhance effectiveness, improve performance, increase productivity, and be useful. Items in these subscales were scored on 5-point Likert scales with endpoints of 1="strongly disagree" and 5="strongly agree."

The Organizational Readiness for Technology Use predictive tool was used to assess dimensions of organizational readiness for the use of telephone and video technologies [22,23]. Each item was evaluated using a 5-point Likert scale with endpoints of 1="strongly disagree" and 5="strongly agree." The inventory assessed the perceived feasibility of reimbursement for the technology during and after COVID-19; access to information technology experts, clinical champions, and billing experts to support the use of these technologies; ease of technology integration into the workflow; staff, facilities, and equipment to promote the technology; leadership, staff, and patient support; technology accessibility and affordability; and staff training.

Data Analysis

Responses from the 2020 cross-sectional survey were compared to responses in the 2022 cross-sectional survey. Since the responses in 2020 and 2022 were anonymous and comprised two separate samples, a true longitudinal model could not be analyzed. Each staff member answered the same set of telephone and video telehealth questions. We calculated frequencies and descriptive statistics for the questions about participating site characteristics and use of different telehealth services. To compare the different services used between 2020 and 2022, we analyzed each service separately and removed organizations that did not offer a particular service from that specific analysis. The following model comparison between the years 2020 and 2022 was performed using the glmer() function with a binomial distribution since "yes/no" binomial outcomes were present in the data:

$$\text{outcome} = \beta_0 + \beta_1\text{year} + \nu_0 + e,$$

with ν_0 representing the organization's random effects.

For models where each staff member had multiple responses, the telehealth type (telephone vs video) varied among and within staff members and organizations. Thus, responses were nested in staff members and staff members were nested within organizations. Using the methods described by Brauer and Curtin [24], the following random effects structure was used:

$$\text{outcome} = \beta_0 + \beta_1\text{type} + \beta_2\text{year} + \beta_3\text{type} \times \text{year} + u_0 + u_1\text{type} + \nu_0 + \nu_1\text{type} + e,$$

where u is the random intercept for organizations and e is the within-organization random error.

Ethical Considerations

The University of Wisconsin's Health Science Minimal Risk Institutional Review Board conducted a review of the study (2020 - 0551) and determined that it met the criteria for exempt human subjects research in accordance with the definition under 45 Code of Federal Regulations (CFR) 46 [25]. No incentive was provided to the respondents to complete the questionnaire and responses were collected using Research Electronic Data Capture (REDCap), a secure web application. The University of Wisconsin's institutional review board approved the presurvey information sheet, survey distribution, and recruitment of study participants.

Results

Basic Characteristics of the Samples

A total of 511 participants started the survey and 136 were removed in total (76 for incomplete responses, 2 that were part of our staff testing, and 8 for duplicate responses by the same individual). An incomplete response was removed if less than 4 of the 12 survey scales or sections (with demographics counting as one scale) were completed. For those who completed 4 or more of the sections, only completed sections were used in the relevant analyses. Duplicate responses were determined by matching email identifiers: only the first complete response was retained in cases of duplicate emails. IP addresses were not used due to shared staff computers.

The final sample in the 2022 survey included 375 responses from 325 unique organizations located across 37 states. The survey was distributed to 2102 organizations that provide SUD services, with an estimated return rate of 15% (325 organizations). The staff members predominantly reported their setting as an urban environment (177/375, 47.2%; see Table 1).

The distribution of responses between organizational settings was not significantly different from that of the 2020 sample (χ^2_3=4.48, *P*=.21). There was a fairly even distribution of staff members across the organizational types in 2022, with the most respondents (58%) being from a specialty treatment setting (23% for stand-alone sites and 2 - 5 sites, and 12% for 5+ organizational sites; see Table 1). The characteristics related to organizational type did differ significantly from those in 2020 (χ^2_4=13.34, *P*=.01) due to a higher percentage of participants from the recovery community in 2022. The distribution of staff members' roles within the organization also significantly differed between 2020 and 2022 (χ^2_1=8.38, *P*=.004), with a higher percentage of staff members in 2022 reporting that they provided treatment and/or recovery services than in 2020. To account for the differences in the samples, organization type and staff members' roles within the organization were added to the models comparing responses from 2020 and 2022.

Table . Characteristics of the participating organizations and staff.

Characteristic		Participating staff members in 2020 (n=581), n (%)	Participating staff members in 2022 (n=370), n (%)	χ^2	*df*	*P* value
Organization setting				4.48	3	.21
	Urban	244 (42)	177 (48)			
	Suburban	101 (17)	56 (15)			
	Small city	119 (21)	62 (17)			
	Rural	117 (20)	67 (18)			
	Tribal reservation	—[a]	8 (2)			
Organization type				13.34	4	.01
	Health system (Hospital, HMO[b], or primary care network)	101 (19)	53 (14)	2.42	1	.12
	Specialty behavioral health: stand-alone	121 (22)	84 (23)	0.02,	1	.90
	Specialty behavior health provider: at 2-5 sites[c]	211 (39)	83 (23)	1.54	1	.21
	Specialty behavior health provider: 5+ sites[c]	—	44 (12)	1.54	1	.21
	Opioid treatment program	59 (11)	41 (11)	<0.01	1	.97
	Recovery community	52 (10)	63 (17)	10.71	1	.001
Role within the organization				8.38	1	.004
	Administrator	205 (37)	103 (28)			
	Personnel providing treatment and/or recovery services	344 (63)	266 (72)			

[a]Not a response option in the 2020 survey.

[b]HMO: Health Maintenance Organization.

[c]These two categories were combined to compare to 2020.

Technology Use

The probability of using various technologies did not change from 2020 to 2022, except for the use of secure chats for recovery support sessions (*P*=.02; see Table 2). In 2020, SUD treatment staff were more likely to use secure chats for recovery support sessions than in 2022 (Table 2).

Table . Use of different technologies.

Technology	Probability (95% CI)		P value
	2020	2022	
Computerized screening and assessments	0.76 (0.68 - 0.82)	0.80 (0.71 - 0.87)	.24
Mobile app(s) during recovery	0.00 (0.00 - 0.00)	0.00 (0.00 - 0.00)	.89
Mobile app(s) during treatment	0.24 (0.17 - 0.31)	0.18 (0.12 - 0.27)	.14
Organizational web portal patients can use to schedule appointments	0.12 (0.07 - 0.18)	0.16 (0.10 - 0.24)	.10
Secure chats for recovery support sessions	0.44 (0.38 - 0.51)	0.33 (0.26 - 0.42)	.02
Text appointment reminders	0.60 (0.51 - 0.68)	0.68 (0.58 - 0.77)	.07
Text motivational messages	0.00 (0.00 - 0.00)	0.00 (0.00 - 0.00)	.93

Use of Different Telehealth Services

Overall, the respondent organizations' use of telephone- and video-based telehealth services significantly declined between 2020 and 2022, as shown in Figure 1 and Table 3. The only exceptions were video-based peer recovery services (P=.34) and video- and telephone-based therapy services in residential treatment programs (P=.79), all of which also declined but not significantly (Table 3). These data present a mixed picture (see Figure 1). Telehealth use remained encouragingly high (greater than 50%) for screening and assessment, case management, peer recovery support, and regular outpatient services, yet the utilization of these services significantly declined since the early months of the COVID-19 distancing recommendations (Figure 1 and Table 3).

Figure 1. Telehealth services availability level (2020 and 2022).

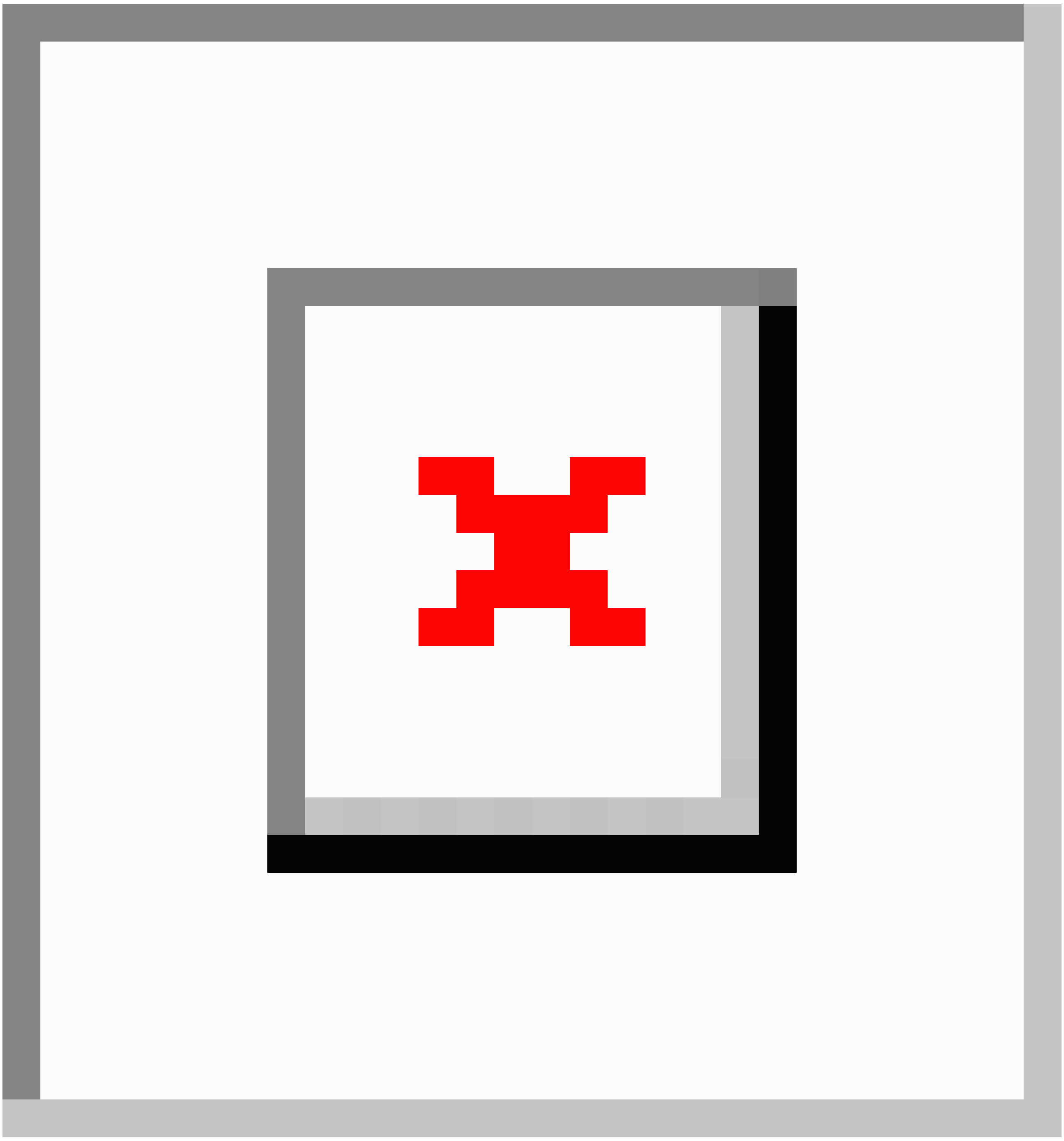


Table . Use of different telehealth services for 2020 versus 2022.

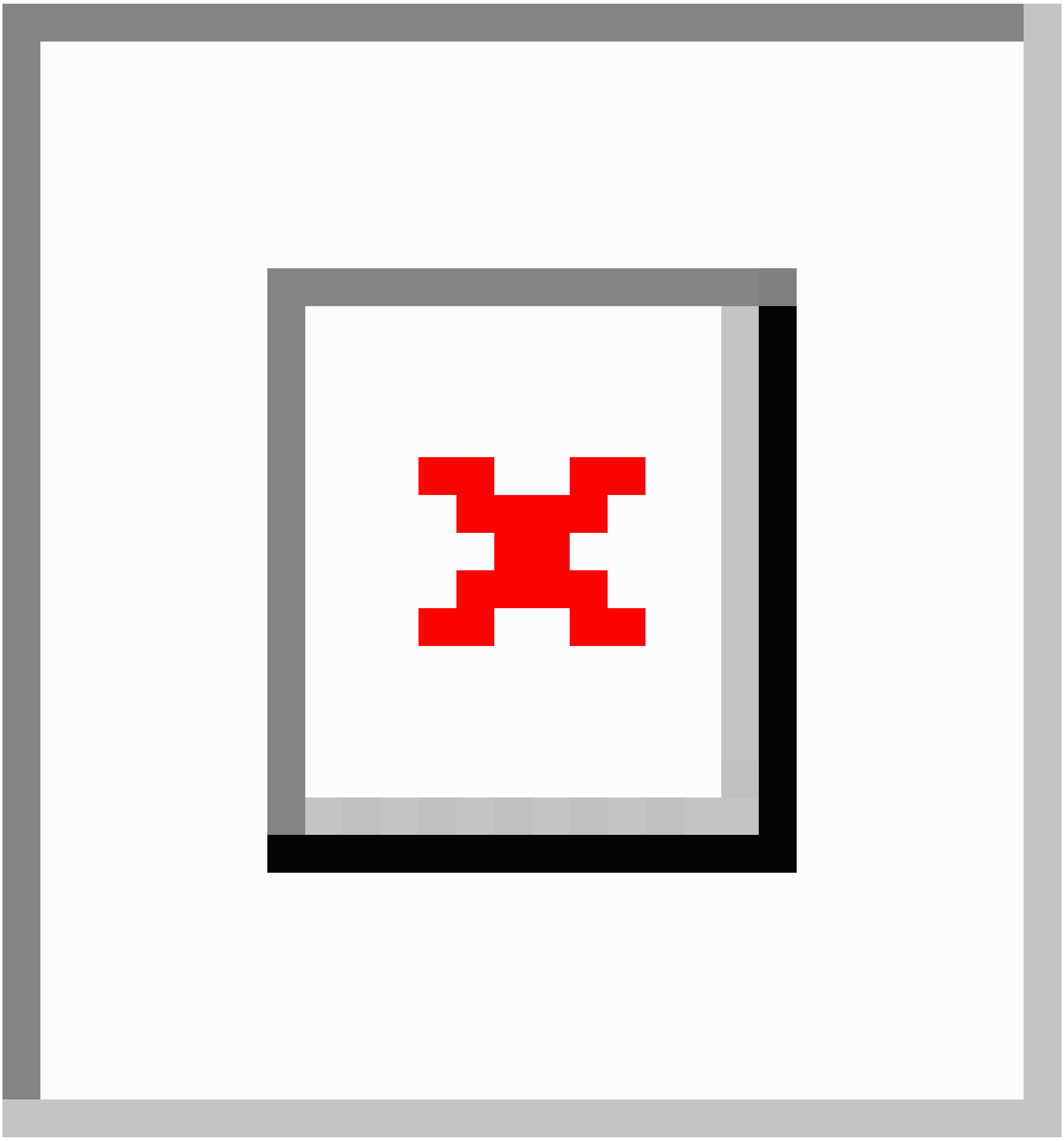
Telehealth services	Telephone-based service		Video-based service	
	χ^2 (*df*=1)	*P* value	χ^2 (*df*=1)	*P* value
Buprenorphine, suboxone, or subutex	15.39	<.001	15.36	<.001
Case management	22.91	<.001	5.44	.02
Intensive outpatient treatment	14.53	<.001	5.09	.02
Peer recovery support	6.85	.009	0.92	.34
Regular outpatient	50.6	<.001	7.51	.006
Screening and assessment/intake	45.11	<.001	7.22	.007
Therapy session residential settings	2.12	.14	0.07	.79

Perceived Usefulness and Ease of Use

Overall, video was perceived as more useful for telehealth than the phone (mean difference [MD] 0.16; *P*<.001) across the 2020 and 2022 samples (Figure 2). Telehealth (phone and video together) was perceived as more useful in 2022 than in 2020 (MD 0.23; *P*=.002). This overall effect was driven by a change in perceptions of telephone-based telehealth, for which perceived usefulness was higher in 2022 than in 2020 (MD −0.23; *P*=.002); however, there was no significant difference between years in the perceived usefulness of video-based telehealth (MD −0.12; *P*=.13).

By contrast, telephone was perceived as easier to use for telehealth than video (MD=−0.32, *P*<.001) across both the 2020 and 2022 samples. There was no statistically significant difference between 2020 and 2022 in perceived ease of use for telehealth (telephone and video together) (MD 0.12; *P*=.11). For video-based telehealth, ease of use was perceived higher in 2022 than in 2020 (MD −0.35; *P*<.001), whereas for telephone-based telehealth, there was no significant difference in ease of use between 2022 and 2020 (MD −0.12; *P*=.11).

Figure 2. Staff responses regarding the patient's perspective.

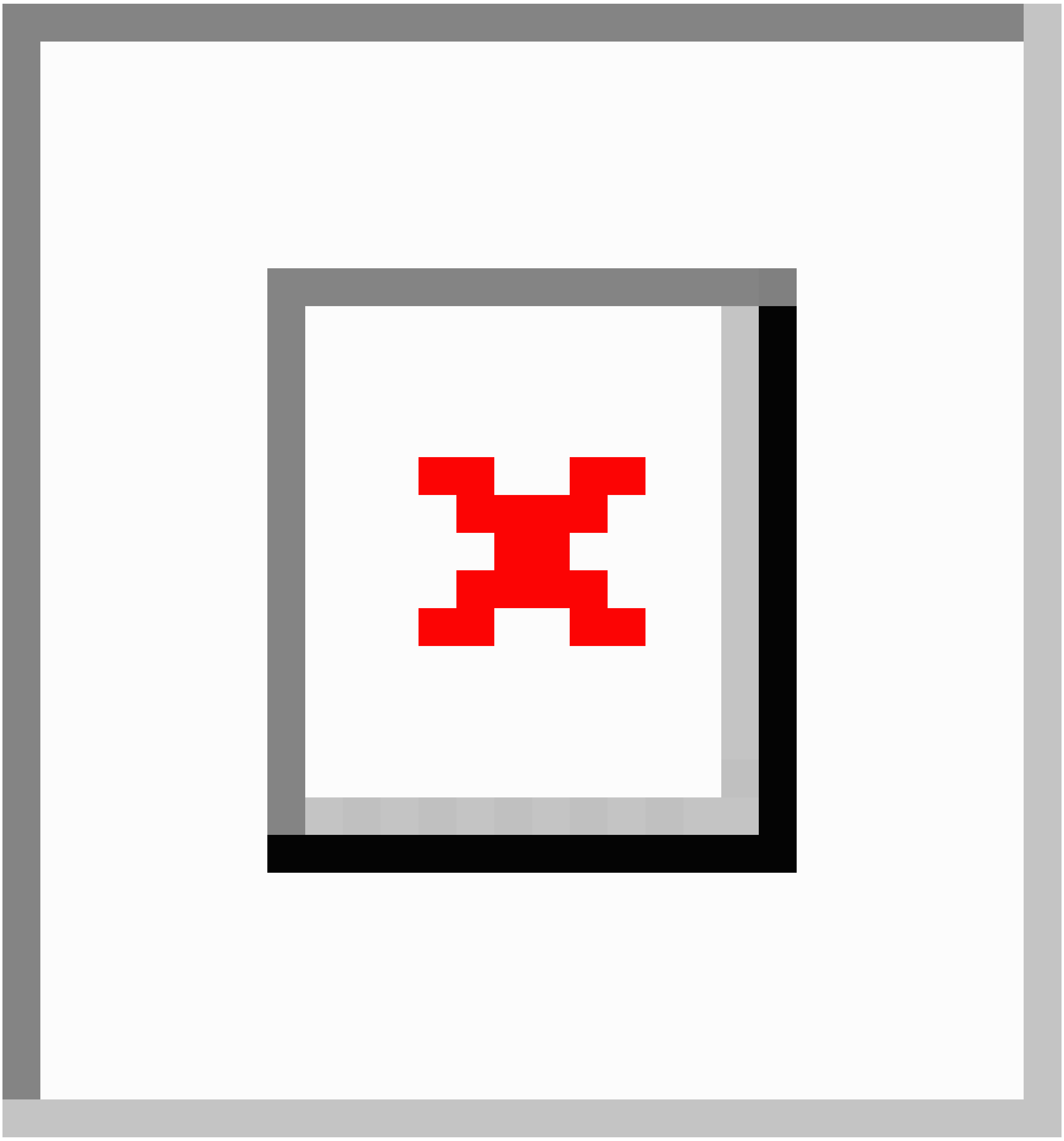


Organizational Readiness for Using Telehealth

Staff responses regarding organizational readiness showed that from their perspective, patients had greater readiness for using telehealth via telephone than via video (Figure 2). By contrast, staff perceived their personal and organizational readiness for using telehealth as greater for video than for the telephone, as shown in Figure 3. Staff ratings of both patient and organizational readiness for telehealth were higher in 2022 than in 2020.

This overall effect was driven by a change in perceptions of telehealth via video. Staff reported that patients found video telehealth easier to use in 2022 than in 2020 (MD −0.35; *P*<.001). In contrast, there was no difference in perceptions of ease of use of telephone-based telehealth between years (MD −0.10; *P*=.19). Additionally, from the staff's perspective, more patients had greater access to video telehealth in 2022 than in 2020 (MD −0.47; *P*<.001).

Figure 3. Staff responses regarding their personal and organizational readiness.



Discussion

The goal of this study was to compare the availability, perceived usefulness, and ease of use of telehealth services across two time intervals: during peak COVID-19 distancing requirements in 2020 and following the removal of distancing requirements in 2022 [19]. Relative to 2020, there was a notable decrease in the use of telehealth services across various clinical tasks, targets, and programs (ie, screening and assessment, case management, regular and intensive outpatient programs, medications for opioid use disorder), with telephonic services experiencing a more pronounced decline in utilization compared to video-based services. Despite the decline in use of telehealth services from 2020 to 2022, the majority (57% - 71%) of the 325 SUD organizations surveyed in 2022 reported the continued utilization of telehealth services, a rate that was much higher than national estimates, suggesting that 20% - 30% of SUD organizations offered telehealth prior to the COVID-19 pandemic [6]. The persistent use of telehealth services in 2022 highlights the popularity of the regulatory flexibilities that were implemented during the COVID-19 pandemic: such flexibilities were initially intended to be temporary but have endured due to substantial advocacy efforts led by providers, patients, investors, and policy makers [26-33].

The current findings replicated the patterns revealed in the 2020 survey reported by Molfenter et al [1], highlighting the perceived usefulness of video-based services and the perceived ease of use of telephonic services. Notably, while both effects were consistent with the prior report, they were of lesser magnitude following the discontinuation of COVID-19 safety protocols. Prior research indicates strong patient satisfaction with video-based services [34], either greater than or equivalent to satisfaction with telephone services. However, confidence in these results is limited by the lack of well-validated measures assessing satisfaction with technology: a 2022 systematic review identified 10 scales across 12 studies, finding 9 of the studies to be of "inadequate" or "doubtful" quality. Furthermore, the scalability of video-based services presents significant challenges. Cost is a major issue, encompassing the initial purchase and ongoing operation of videoconferencing platforms and securing reimbursement from state Medicaid systems and insurers. Another challenge is protecting patient anonymity and complying with the Health Insurance Portability and Accountability Act and 42 CFR Part 2, which provides privacy protections for SUD-related records [35,36]. The absence of an accreditation system for documenting compliance in telehealth service delivery leaves individual organizations in a precarious position concerning the 42 CFR regulatory requirements [37].

Another reason that video might consistently be rated as more difficult to use is organizational resistance encountered when unfamiliar technologies are incorporated into familiar staff roles, functions, and workflow. Fortunately, this challenge may be addressed via workforce education efforts. To that end, the ATTC Network has, in recent years, developed learning resources to aid this cause. One example is demonstration videos, which, as asynchronous learning resources, enable workforce members to individually access and observe models to visualize and approximate their future clinical practice behavior. Among available ATTC-sponsored demonstration videos are those modeling the video-based delivery of (1) empirically supported therapies (eg, motivational interviewing, cognitive behavioral therapy), (2) care interactions in the context of medications for opioid use disorder, and (3) effective clinical supervision practices. Several user-friendly resources to promote the use of telehealth for delivering evidence-based SUD practices can be found on the ATTC Network website [38]. Beyond workforce education efforts, a salient priority within the addiction research community should be further validation of the cost- and clinical effectiveness of video-based SUD services.

The current work should be interpreted in the context of methodological limitations. First and foremost, the convenience sampling approach used by regional ATTCs was prone to both selection and response biases. Despite the distribution of the electronic survey to large regional mailing lists and a wide range of organizational settings, the final response rate was only 15%. Second, it is possible that organizations that were more comfortable with telehealth might also have been more comfortable completing an electronic survey of 100+ questions, which could have introduced systematic response bias. Third, our reliance on survey methodology limited our ability to gather in-depth feedback, given the commonly stated "survey fatigue" during the COVID-19 pandemic [39]. A more detailed and granular account of workforce perceptions could have been derived via a qualitative or mixed methods inquiry. Finally, it should be restated that trends in perceptions of telehealth services at SUD treatment and recovery organizations were examined from two separate cross-sectional survey samples. Thus, longitudinal changes in the opinions of specific individuals polled as representatives of the addiction workforce should not be inferred.

These limitations notwithstanding, the results of this study suggest the initially promising outlook for telehealth services among the addiction workforce, as reported by Molfenter et al [1] in 2021 shortly after the initiation of COVID-19 social distancing recommendations. This outlook is now followed by reasons for both optimism and concern following the discontinuance of those social distancing recommendations. Of potential concern, a smaller percentage of SUD organizations indicated current telehealth service availability following removal of COVID-19 social distancing recommendations. However, this might be expected as some organizations or professionals could only deliver remote services during the pandemic, whereas following discontinuation of safety distancing recommendations, organizations could provide services either in person or via telehealth. Encouragingly, the availability of telehealth services remains common and endorsed by representatives from most SUD organizations in this sample, as is generally consistent with the findings of a recent review [40]. Both video-based and telephonic modalities for telehealth services continue to be perceived positively, with health professionals finding video services more useful but telephone services easier. Taken together with other recent research concerning these specific telehealth modalities [41-44], it seems that salient setting-, patient-, and population-level matching considerations may be needed to promote useful—and equitable—access to telephone- and video-based telehealth services. Additionally, structural support is likely required to overcome other challenges in availing telehealth services related to cost, reimbursement, and patient privacy. These are significant issues that deserve greater attention in future research on telehealth services at SUD service organizations, as does determination of the relative cost- and clinical effectiveness of specific telehealth services.

Acknowledgments

This study was funded by the following Substance Abuse and Mental Health Services Administration (SAMHSA) Addiction Technology Transfer Center (ATTC) cooperative agreements: UR1TI080211, UR1TI080201, UR1TI080203, UR1TI082504 UR1TI080207, and UR1TI080209. We are grateful to Maureen Fitzgerald for her leadership and assistance with proofreading and communications.

Conflicts of Interest

TM has <1% stock ownership in Center for Health Enhancement Systems Studies (CHESS) Mobile Health. CHESS Mobile Health does not provide or facilitate the delivery of telephonic or video care services. TM has worked extensively with his institution to manage any conflicts of interest. Other authors have no conflicts to declare.

References

1. Molfenter T, Roget N, Chaple M, et al. Use of telehealth in substance use disorder services during and after COVID-19: online survey study. JMIR Ment Health 2021 Feb 8;8(2):e25835. [doi: 10.2196/25835] [Medline: 33481760]
2. Roy J, Levy DR, Senathirajah Y. Defining telehealth for research, implementation, and equity. J Med Internet Res 2022 Apr 13;24(4):e35037. [doi: 10.2196/35037] [Medline: 35416778]
3. King VL, Stoller KB, Kidorf M, et al. Assessing the effectiveness of an internet-based videoconferencing platform for delivering intensified substance abuse counseling. J Subst Abuse Treat 2009 Apr;36(3):331-338. [doi: 10.1016/j.jsat.2008.06.011] [Medline: 18775625]
4. Tarp K, Bojesen AB, Mejldal A, Nielsen AS. Effectiveness of optional videoconferencing-based treatment of alcohol use disorders: randomized controlled trial. JMIR Ment Health 2017 Sep 29;4(3):e38. [doi: 10.2196/mental.6713] [Medline: 28963093]
5. Mark TL, Treiman K, Padwa H, Henretty K, Tzeng J, Gilbert M. Addiction treatment and telehealth: review of efficacy and provider insights during the COVID-19 pandemic. Psychiatr Serv 2022 May;73(5):484-491. [doi: 10.1176/appi.ps.202100088] [Medline: 34644125]
6. Frueh BC, Henderson S, Myrick H. Telehealth service delivery for persons with alcoholism. J Telemed Telecare 2005;11(7):372-375. [doi: 10.1258/135763305774472060] [Medline: 16238840]
7. Kodjebacheva GD, Culinski T, Kawser B, Coffer K. Satisfaction with telehealth services compared with nontelehealth services among pediatric patients and their caregivers: systematic review of the literature. JMIR Pediatr Parent 2023 Apr 27;6:e41554. [doi: 10.2196/41554] [Medline: 37000504]
8. Molfenter T, Brown R, O'Neill A, Kopetsky E, Toy A. Use of telemedicine in addiction treatment: current practices and organizational implementation characteristics. Int J Telemed Appl 2018 Mar;2018:3932643. [doi: 10.1155/2018/3932643] [Medline: 29713341]
9. Christensen CM, McDonald R, Altman EJ, Palmer JE. Disruptive innovation: an intellectual history and directions for future research. J Manag Stud 2018 Nov;55(7):1043-1078. [doi: 10.1111/joms.12349]
10. Pifer R. Telehealth use falls nationally for third month in a row: fair health. Healthcare Dive. 2021. URL: https://www.healthcaredive.com/news/telehealth-use-falls-nationally-for-third-month-in-a-row-fair-health/603009/ [accessed 2024-07-03]
11. Jones CM, Shoff C, Hodges K, et al. Receipt of telehealth services, receipt and retention of medications for opioid use disorder, and medically treated overdose among medicare beneficiaries before and during the COVID-19 pandemic. JAMA Psychiatry 2022 Oct 1;79(10):981-992. [doi: 10.1001/jamapsychiatry.2022.2284] [Medline: 36044198]
12. AlRasheed R, Woodard GS, Nguyen J, et al. Transitioning to telehealth for COVID-19 and beyond: perspectives of community mental health clinicians. J Behav Health Serv Res 2022 Oct;49(4):524-530. [doi: 10.1007/s11414-022-09799-z] [Medline: 35538275]
13. Békés V, Aafjes-van Doorn K, Zilcha-Mano S, Prout T, Hoffman L. Psychotherapists' acceptance of telepsychotherapy during the COVID-19 pandemic: a machine learning approach. Clin Psychol Psychother 2021 Nov;28(6):1403-1415. [doi: 10.1002/cpp.2682] [Medline: 34723404]
14. Fischer SH, Predmore Z, Roth E, Uscher-Pines L, Baird M, Breslau J. Use of and willingness to use video telehealth through the COVID-19 pandemic. Health Aff 2022;41(11):1645-1651. [doi: 10.1377/hlthaff.2022.00118] [Medline: 36343311]
15. Tse J, LaStella D, Chow E, et al. Telehealth acceptability and feasibility among people served in a community behavioral health system during the COVID-19 pandemic. Psychiatr Serv 2021 Jun;72(6):654-660. [doi: 10.1176/appi.ps.202000623] [Medline: 33957769]
16. Kennedy AJ, George JS, Rossetti G, et al. Providing low-barrier addiction treatment via a telemedicine consultation service during the COVID-19 pandemic in Los Angeles, county: an assessment 1 year later. J Addict Med 2023 Jan;17(1):e64-e66. [doi: 10.1097/ADM.0000000000001034] [Medline: 35839323]
17. Predmore ZS, Roth E, Breslau J, Fischer SH, Uscher-Pines L. Assessment of patient preferences for telehealth in post-COVID-19 pandemic health care. JAMA Netw Open 2021 Dec 1;4(12):e2136405. [doi: 10.1001/jamanetworkopen.2021.36405] [Medline: 34851400]
18. Bouabida K, Lebouché B, Pomey MP. Telehealth and COVID-19 pandemic: an overview of the telehealth use, advantages, challenges, and opportunities during COVID-19 pandemic. Healthcare 2022 Nov 16;10(11):2293. [doi: 10.3390/healthcare10112293] [Medline: 36421617]
19. Cross-Technology Transfer Center (TTC) Workgroup on Virtual Learning. Virtual reality for behavioral health workforce development in the era of COVID-19. J Subst Abuse Treat 2021 Feb;121:108157. [doi: 10.1016/j.jsat.2020.108157] [Medline: 33223379]

20. Davis FD. A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results. Phd thesis. : Massachusetts Institute of Technology; 1986 URL: https://dspace.mit.edu/handle/1721.1/15192 [accessed 2024-07-03]
21. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
22. Gustafson DH, Brennan PF, Hawkins RP. Investing in E-Health: Springer; 2007.
23. Wen KY, Gustafson DH, Hawkins RP, et al. Developing and validating a model to predict the success of an IHCS implementation: the readiness for implementation model. J Am Med Inform Assoc 2010 Nov;17(6):707-713. [doi: 10.1136/jamia.2010.005546] [Medline: 20962135]
24. Brauer M, Curtin JJ. Linear mixed-effects models and the analysis of nonindependent data: a unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. Psychol Methods 2018 Sep;23(3):389-411. [doi: 10.1037/met0000159] [Medline: 29172609]
25. 45 CFR 46. U.S. Department of Health and Human Services Office for Human Research Protections. URL: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html [accessed 2024-07-21]
26. Thomas EE, Haydon HM, Mehrotra A, et al. Building on the momentum: sustaining telehealth beyond COVID-19. J Telemed Telecare 2022 May;28(4):301-308. [doi: 10.1177/1357633X20960638] [Medline: 32985380]
27. Bestsennyy O, Gilbert G, Harris A, Rost J. Telehealth: a quarter-trillion-dollar post-COVID-19 reality. McKinsey & Company. 2021. URL: https://www.mckinsey.com/industries/healthcare/our-insights/telehealth-a-quarter-trillion-dollar-post-covid-19-reality [accessed 2024-07-03]
28. Slomski A. Telehealth success spurs a call for greater post-COVID-19 license portability. JAMA 2020 Sep 15;324(11):1021-1022. [doi: 10.1001/jama.2020.9142] [Medline: 32930757]
29. TIP 63: medications for opioid use disorder. : U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Substance Abuse Treatment; 2021 URL: https://store.samhsa.gov/product/tip-63-medications-opioid-use-disorder/pep21-02-01-002 [accessed 2024-07-03]
30. Levander XA, Pytell JD, Stoller KB, Korthuis PT, Chander G. COVID-19-related policy changes for methadone take-home dosing: a multistate survey of opioid treatment program leadership. Subst Abus 2022;43(1):633-639. [doi: 10.1080/08897077.2021.1986768] [Medline: 34666636]
31. Green TC, Bratberg J, Finnell DS. Opioid use disorder and the COVID 19 pandemic: a call to sustain regulatory easements and further expand access to treatment. Subst Abus 2020;41(2):147-149. [doi: 10.1080/08897077.2020.1752351] [Medline: 32314951]
32. Keyes KM, Cerdá M, Brady JE, Havens JR, Galea S. Understanding the rural-urban differences in nonmedical prescription opioid use and abuse in the United States. Am J Public Health 2014 Feb;104(2):e52-e59. [doi: 10.2105/AJPH.2013.301709] [Medline: 24328642]
33. Palombi LC, St Hill CA, Lipsky MS, Swanoski MT, Lutfiyya MN. A scoping review of opioid misuse in the rural United States. Ann Epidemiol 2018 Sep;28(9):641-652. [doi: 10.1016/j.annepidem.2018.05.008] [Medline: 29921551]
34. Richter KP, Shireman TI, Ellerbeck EF, et al. Comparative and cost effectiveness of telemedicine versus telephone counseling for smoking cessation. J Med Internet Res 2015 May 8;17(5):e113. [doi: 10.2196/jmir.3975] [Medline: 25956257]
35. Fact sheet 42 CFR Part 2 Final Rule. U.S. Department of Health and Human Services. 2024. URL: https://www.hhs.gov/hipaa/for-professionals/regulatory-initiatives/fact-sheet-42-cfr-part-2-final-rule/index.html [accessed 2024-07-03]
36. Substance abuse disorder treatment: 42 CFR Part 2 and HIPAA. Compliancy Group. URL: https://compliancy-group.com/substance-abuse-disorder-treatment-42-cfr-part-2-and-hipaa/ [accessed 2024-07-03]
37. Knopf A. Confidentiality of substance use disorder treatment records: update. Brown Univ Child Adolesc Psychopharm Update 2020 Jun;22(6):1-3. [doi: 10.1002/cpu.30496]
38. Telehealth video demonstrations. ATTC. URL: https://attcnetwork.org/northwest-telehealth-video-demonstrations/ [accessed 2024-07-03]
39. de Koning R, Egiz A, Kotecha J, et al. Survey fatigue during the COVID-19 pandemic: an analysis of neurosurgery survey response rates. Front Surg 2021 Aug;8:690680. [doi: 10.3389/fsurg.2021.690680] [Medline: 34458314]
40. Pham H, Lin C, Zhu Y, et al. Telemedicine-delivered treatment for substance use disorder: a scoping review. J Telemed Telecare 2023 Aug 3:1357633X231190945. [doi: 10.1177/1357633X231190945] [Medline: 37537907]
41. Lin C, Pham H, Zhu Y, et al. Telemedicine along the cascade of care for substance use disorders during the COVID-19 pandemic in the United States. Drug Alcohol Depend 2023 Jan 1;242:109711. [doi: 10.1016/j.drugalcdep.2022.109711] [Medline: 36462230]
42. Chang JE, Lindenfeld Z, Thomas T, Waldman J, Griffin J. Patient characteristics associated with phone versus video telemedicine visits for substance use treatment during COVID-19. J Addict Med 2022 Nov;16(6):659-665. [doi: 10.1097/ADM.0000000000000985] [Medline: 35195540]
43. Mattocks KM, Moore DT, Wischik DL, Lazar CM, Rosen MI. Understanding opportunities and challenges with telemedicine-delivered buprenorphine during the COVID-19 pandemic. J Subst Abuse Treat 2022 Aug;139:108777. [doi: 10.1016/j.jsat.2022.108777] [Medline: 35346533]

44. Gujral K, Van Campen J, Jacobs J, Kimerling R, Zulman DM, Blonigen D. Impact of VA's video telehealth tablets on substance use disorder care during the COVID-19 pandemic. J Subst Use Addict Treat 2023 Jul;150:209067. [doi: 10.1016/j.josat.2023.209067] [Medline: 37164153]

Abbreviations

ATTC: Addiction Technology Transfer Center
CFR: Code of Federal Regulations
MD: mean difference
REDCap: Research Electronic Data Capture
SAMHSA: Substance Abuse and Mental Health Services Administration
SUD: substance use disorder
TAM: Technology Acceptance Model

Edited by J Torous; submitted 15.09.23; peer-reviewed by D Hilty, H Alshawaf; revised version received 29.05.24; accepted 07.06.24; published 12.08.24.

Please cite as:
Pusnik A, Hartzler B, Vjorn O, Rutkowski BA, Chaple M, Becker S, Freese T, Nichols M, Molfenter T
Comparison of Use Rates of Telehealth Services for Substance Use Disorder During and Following COVID-19 Safety Distancing Recommendations: Two Cross-Sectional Surveys
JMIR Ment Health 2024;11:e52363
URL: https://mental.jmir.org/2024/1/e52363
doi:10.2196/52363

Original Paper

Developing a Framework to Infer Opioid Use Disorder Severity From Clinical Notes to Inform Natural Language Processing Methods: Characterization Study

Melissa N Poulsen[1*], MPH, PhD; Philip J Freda[2*], PhD; Vanessa Troiani[3], PhD; Danielle L Mowery[4], PhD

[1]Department of Population Health Sciences, Geisinger, Danville, PA, United States

[2]Department of Computational Biomedicine, Cedars-Sinai Medical Center, West Hollywood, CA, United States

[3]Department of Autism and Developmental Medicine, Geisinger, Danville, PA, United States

[4]Department of Biostatistics, Epidemiology and Informatics, Institute of Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, United States

[*]these authors contributed equally

Corresponding Author:
Melissa N Poulsen, MPH, PhD
Department of Population Health Sciences
Geisinger
100 North Academy Avenue
Danville, PA, 17822
United States
Phone: 1 5702149322
Email: mpoulsen@geisinger.edu

Abstract

Background: Information regarding opioid use disorder (OUD) status and severity is important for patient care. Clinical notes provide valuable information for detecting and characterizing problematic opioid use, necessitating development of natural language processing (NLP) tools, which in turn requires reliably labeled OUD-relevant text and understanding of documentation patterns.

Objective: To inform automated NLP methods, we aimed to develop and evaluate an annotation schema for characterizing OUD and its severity, and to document patterns of OUD-relevant information within clinical notes of heterogeneous patient cohorts.

Methods: We developed an annotation schema to characterize OUD severity based on criteria from the *Diagnostic and Statistical Manual of Mental Disorders, 5th edition*. In total, 2 annotators reviewed clinical notes from key encounters of 100 adult patients with varied evidence of OUD, including patients with and those without chronic pain, with and without medication treatment for OUD, and a control group. We completed annotations at the sentence level. We calculated severity scores based on annotation of note text with 18 classes aligned with criteria for OUD severity and determined positive predictive values for OUD severity.

Results: The annotation schema contained 27 classes. We annotated 1436 sentences from 82 patients; notes of 18 patients (11 of whom were controls) contained no relevant information. Interannotator agreement was above 70% for 11 of 15 batches of reviewed notes. Severity scores for control group patients were all 0. Among noncontrol patients, the mean severity score was 5.1 (SD 3.2), indicating moderate OUD, and the positive predictive value for detecting moderate or severe OUD was 0.71. Progress notes and notes from emergency department and outpatient settings contained the most and greatest diversity of information. Substance misuse and psychiatric classes were most prevalent and highly correlated across note types with high co-occurrence across patients.

Conclusions: Implementation of the annotation schema demonstrated strong potential for inferring OUD severity based on key information in a small set of clinical notes and highlighting where such information is documented. These advancements will facilitate NLP tool development to improve OUD prevention, diagnosis, and treatment.

(JMIR Ment Health 2024;11:e53366) doi:10.2196/53366

KEYWORDS

annotation; clinical notes; natural language processing; opioid related disorders; opioid use disorder; substance use disorders; adult; adults; opioid; annotation schema; severity score; substance misuse; mental health

Introduction

Background

Opioid use disorder (OUD), the problematic pattern of opioid use leading to clinically significant distress or impairment, has remained a significant public health burden for over 2 decades in the United States [1]. In 2021, over 9000 opioid-related overdose deaths involved heroin and nearly 17,000 involved a prescription opioid [2]. In 2021, overdose deaths involving any opioid exceeded 80,000, with 88% involving synthetic opioids like fentanyl. In addition to this significant loss of life, approximately 2.7 million people were diagnosed with OUD in 2021 [3], with the economic cost of the US opioid epidemic estimated to be over US $1 trillion in 2017 [4].

OUD is a *Diagnostic and Statistical Manual of Mental Disorders, 5th edition* (DSM-5) clinical diagnosis with varying levels of severity (mild to severe) based on 11 diagnostic criteria endorsed for a given patient. Criteria include items that assess physiological and behavioral symptoms, as well as harmful health and social consequences of opioid use. Access to patients' OUD status and severity is valuable to patient care. For example, a clinician might use such information to inform their approach for helping manage a patient's pain (eg, whether to use opioid analgesics or the dose required). Information about OUD severity aids in clinical decision-making regarding appropriate treatment [5], such as determining whether a patient should be referred for psychotherapy versus an inpatient or outpatient treatment facility. OUD severity has also been proposed for use in measurement-based care to track indicators of disease and improved outcomes [6]. For these reasons, accurately identifying OUD status and severity holds crucial importance to the development of patient care strategies and clinical outcome prediction.

Detecting and Characterizing OUD in Electronic Health Records

Electronic health records (EHRs) include discrete fields containing information such as diagnostic labels or medication orders, as well as fields that contain free text from clinical notes, encounters, and laboratory testing. Numerous algorithms using diagnostic codes have been designed to address problematic opioid use, including identification of patients at risk for prescription opioid misuse [7], OUD prediction [8], nonmedical opioid use detection [9,10], identification of OUD [11], characterization of problematic opioid use [12], and overdose risk prediction [13]. However, the success of such algorithms is limited when diagnostic codes are minimally applied to patient charts. Information about substance use disorders may be missing from EHRs due to a variety of factors, including disjointed care across hospitals, lack of specialty diagnostic expertise, or stigma of a given diagnosis. Thus, OUD can be challenging to isolate within a patient's medical record [14,15]. Furthermore, most algorithms have been developed in the context of research focusing on patients with prescription opioid misuse or chronic pain. Patients with chronic pain with opioid prescriptions are at increased risk for opioid misuse [16,17], but only focusing on such populations may exclude patient populations with illicit opioid use or otherwise outside of chronic pain treatment. Previous research indicates clinical notes provide a source of rich information that could improve efforts to identify and characterize OUD [12,18,19].

Importance of Annotation for Developing and Evaluating Natural Language Processing Tools

Natural language processing (NLP) tools have the potential to improve OUD detection and severity characterization, but many NLP frameworks require high quality data with reliably labeled OUD-related information. Common workflows for generating such a data set entail developing a schema (eg, containing classes [entities and events] and attributes [qualifiers]) representing OUD-related information, creating a codebook of instruction for the annotation process, conducting an agreement study to assess schema reliability, and facilitating consensus review of disagreements to generate a reference standard for benchmarking the NLP system [20]. Prior to these steps, it is imperative to understand how and where relevant information is documented in EHRs to inform data extraction and subsequent automation. Generally, this step is not well described in the scientific literature, nor are the documentation patterns of such information well characterized in studies. This step can be critical to informing intelligent search of EHRs and limiting the note types necessary for operationalizing the algorithm, thereby reducing computational effort and potentially improving accuracy. Although prior studies have used NLP to identify problematic opioid use from EHRs [21-26], few have described an annotation process and none have reported documentation patterns for OUD-relevant information within clinical notes.

Study Objectives

Our long-term goal is to develop an automated NLP method to identify OUD arising from prescription or illicit opioid use and characterize the severity of such use that can be used for future EHR-based studies to drive informatics solutions to improve the prevention, diagnosis, and treatment of OUD through clinical care. Toward this goal, we developed an annotation schema to characterize severity and documented patterns of OUD-relevant information. Using clinical notes of varying type and across encounter settings from a large integrated health system, we annotated OUD symptoms and other relevant information, comparing several heterogenous cohorts—including patients with chronic pain, OUD diagnoses, and receiving medication treatment for OUD—to explore the following questions:

- How accurately can OUD severity be inferred from text from a small number of targeted clinical notes per patient?
- Where and how do clinical teams document OUD-related information, in terms of clinical note types and encounter settings?

- How is OUD-related information documented over time relative to an opioid- or OUD-specific health care encounter?
- What is the frequency of OUD-related concepts and their co-occurrence in clinical notes?

This information, along with the annotation schema developed and described in this study, may be useful in future development of NLP methods to identify and characterize the severity of OUD.

Methods

Study Design

We developed an annotation schema to identify patients with OUD stemming from either prescription or illicit opioid use and characterize OUD severity based on DSM-5 criteria. We applied the schema to deidentified clinical notes from patients with varying evidence of OUD and a comparison group with minimal exposure to opioid analgesics.

Ethical Considerations

The Geisinger Institutional Review Board and University of Pennsylvania reviewed and approved the protocol for this study (2021-0113).

Study Population

We obtained clinical notes from EHRs of 100 adult patients from Geisinger, a large integrated health system that serves a largely rural area of central and northeast Pennsylvania. We used stratified random sampling to select 20 individuals from each of 5 mutually exclusive groups, stratifying by sex and age categories (18-29, 30-39, 40-49, 50-59, and 60 years and older) to ensure diversity and equal representation in the data set. Study groups were selected from preexisting data sets used in prior studies [11,27,28] to represent various methods of identifying patients with likely or diagnosed OUD from EHRs and a control group (Textbox 1). Further, 2 groups represented chronic pain patients with (at least mild) OUD confirmed through chart review [27] and patient report as to whether their opioid use began with an opioid analgesic prescription (group CP-RX) or not (group CP-nonRX). In total, 2 groups had at least one OUD diagnostic code, but differed as to whether they had an order for medication treatment of OUD such as buprenorphine (group OUD-TX) or not (group OUD-DX). The control group had a single opioid analgesic order in their EHR. Diagnoses and orders used to define study groups occurred within this study's period of January 2012 to March 2020.

Textbox 1. Inclusion and exclusion criteria for each study group.

CP-RX

- Inclusion criteria:
  - Chronic pain: chronic pain was defined as having at least two opioid analgesic prescriptions for nonprogressive musculoskeletal pain.
  - Opioid use disorder (OUD; mild, moderate, or severe) confirmed through chart review.
  - Opioid use began with opioid analgesic prescription. Based on self-report in a survey question.
- Exclusion criteria:
  - Non-European ancestry: individuals with non-European ancestry were excluded because this study's sample was originally assembled for a genetic study.

CP-nonRX

- Inclusion criteria:
  - Chronic pain.
  - OUD (mild, moderate, or severe) confirmed through chart review.
  - Opioid use did not begin with opioid analgesic prescription. Based on self-report in a survey question.
- Exclusion criteria:
  - Non-European ancestry: individuals with non-European ancestry were excluded because this study's sample was originally assembled for a genetic study.

OUD-DX

- Inclusion criteria:
  - At least one diagnosis code for OUD. International Classification of Disease codes used to define OUD were based on Jennings et al [29].
- Exclusion criteria:
  - Chronic pain.
  - Order for medications for OUD including buprenorphine, buprenorphine-naloxone, and naltrexone. Geisinger providers did not prescribe methadone for OUD treatment.

OUD-TX

- Inclusion criteria:
  - At least one diagnosis code for OUD.
  - Order for medications for OUD.
- Exclusion criteria:
  - Chronic pain.

Control

- Inclusion criteria:
  - In total, 1 opioid analgesic order.
- Exclusion criteria:
  - Chronic pain.
  - Diagnosis code for OUD.
  - Order for medications for OUD.

Data Collection

We obtained notes from inpatient, outpatient, and emergency department (ED) encounters. In total, 11 note types were obtained, selected based on clinician input: outpatient clinic notes, progress notes, ancillary progress notes, history and progress (H&P) notes, discharge summaries, ED notes, ED provider notes, ED triage notes, ED support staff notes, communication notes, and lactation notes. We obtained notes for 3 encounter dates per patient: an index date representing either the first observed OUD diagnosis (for OUD-DX and OUD-TX groups and some patients in the CP-RX and CP-nonRX groups) or the most recent opioid analgesic order (for some patients in the CP-RX and CP-nonRX groups and all patients in the control group) and the encounters immediately prior to and following the index date. Multiple notes per patient were obtained for some encounter dates. Notes were deidentified using Philter [30].

Annotation Schema Development and Procedures

Schema development was based upon pilot work [18]. We revised the pilot schema to map classes onto DSM-5 criteria for characterizing OUD severity and to clarify or eliminate ambiguous concepts. We leveraged the extensible Human Oracle Suite of Tools, an open-source text tool [31] to annotate notes. In total, 2 authors (MNP and PJF) separately reviewed and annotated notes across 15 batches (batched by note type). Annotation was completed at the sentence level, assigning full sentences to one or more relevant classes. After each batch, the 2 reviewers adjudicated discordances through discussion, with other study team members providing input when discordances remained unresolved. We varied the note type annotated in consecutive batches to ensure portability of the schema across note types.

Severity Score

We calculated a severity score for each patient based on annotations in their notes. Scores used 18 of the annotation schema classes, which mapped onto the DSM-5 criteria for characterizing OUD severity (Multimedia Appendix 1). The "crosswalk" between classes and DSM-5 criteria was based on a systematic chart review process developed by Palumbo et al [27] and adapted by Poulsen et al [11]. Scores ranged from 0 to 11. Severity was categorized based on DSM-5 guidelines (0-1=no OUD; 2-3=mild OUD; 4-5=moderate OUD; >6=severe OUD).

We calculated positive predictive values (PPVs) for detecting moderate or severe OUD among patients with annotations for the 4 study groups with likely or diagnosed OUD, and again for patients categorized into 2 groups based on their index encounter reason (OUD diagnosis or opioid analgesic order). PPVs were calculated as the number of patients with a severity score >4 divided by the total number of patients with annotations. We did not calculate PPVs separately for severity category (mild, moderate, and severe), as there was insufficient information in EHRs on which to base such a comparison. We also calculated PPVs among the full sample of patients (regardless of whether they had annotations), a more conservative quantification of the severity score's validity that accounts for the lack of OUD-relevant information observed in the reviewed notes.

Results

Annotation of Clinical Notes

The annotation schema contained 27 classes, 12 of which included attributes (Figure 1; Multimedia Appendix 2). Interannotator agreement (IAA) for the 2 reviewers ranged from 20% to 100% across the 15 batches of reviewed notes but was above 70% for all except 4 batches (Figure 2). Lower IAA occurred with less frequently annotated classes and for classes with an attribute denoting a historic concept (eg, *psychiatric condition current* vs *historic*; IAA results not shown by class).

Figure 1. Annotation schema for characterizing OUD severity from clinical note text. OUD: opioid use disorder.

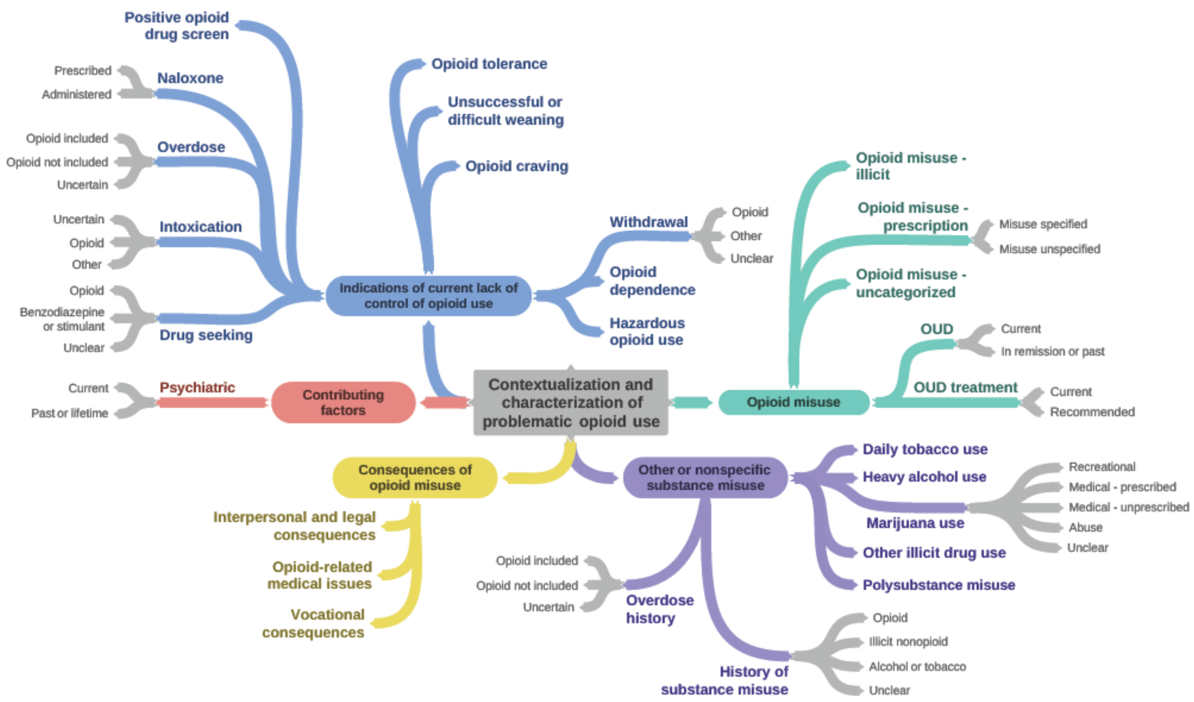
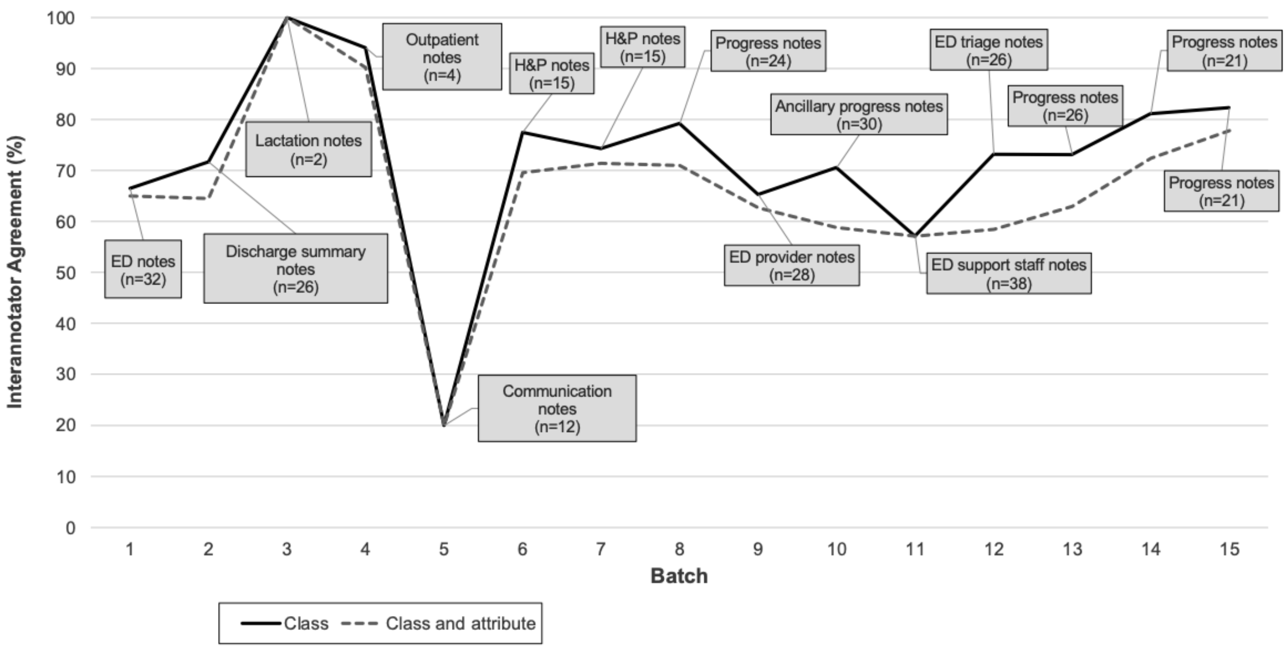


Figure 2. Interannotator agreement across batches of notes with note type and number of notes reviewed per batch. Each batch was a different note type. ED: emergency department; H&P: history and progress.



From the 100 sampled patients and 320 associated notes, we annotated 1436 sentences within 186 notes from 82 patients over 15 batches. The remaining notes did not yield any annotations (ie, they did not contain text relevant to the classification schema). Most patients without annotations were controls (11/18, 61%) and most had an index encounter based on an opioid analgesic order (16/18, 89%). They also had fewer notes available (mean 1.8, SD 1.1) and no notes were from the inpatient setting.

Among the 4 noncontrol groups, 73 patients had annotations (Table 1). OUD-TX was the only group in which all 20 sampled patients had at least one note with text relevant for annotation and the group accounted for the largest proportion of annotations.

Table 1. Counts of patients, notes, and annotated sentences for each study group.

Study group	Count of patients with annotations, n (%)	Index encounter[a]		Count of notes with annotations, n (%)	Count of annotated sentences, n (%)
		OUD[b] diagnosis, n (%)	Opioid order, n (%)		
CP-RX	16 (20)	8 (50)	8 (50)	29 (16)	170 (12)
CP-nonRX	18 (22)	14 (78)	4 (22)	32 (18)	277 (19)
OUD-DX	19 (23)	19 (100)	0 (0)	50 (27)	363 (25)
OUD-TX	20 (24)	20 (100)	0 (0)	61 (34)	602 (42)
Control	9 (11)	0 (0)	9 (100)	14 (8)	24 (2)

[a]Reason for selecting index encounter, either a diagnosis of OUD or an order for an opioid analgesic.

[b]OUD: opioid use disorder.

Can Severity of Problematic Opioid Use Be Inferred From a Limited Number of Clinical Notes?

We used annotated classes to calculate severity scores for the 82 patients with annotations. All control group patients had a score of 0. The mean severity score among the 73 patients in noncontrol groups was 5.1 (SD 3.2). The majority (48/73, 66%) had a score >6 (indicating severe OUD), 4 of 73 (5%) had a score of 4-5 (indicating moderate OUD), 2 of 73 (3%) had a score of 2 (indicating mild OUD), and 19 of 73 (26%) had a score of 0-1 (indicating no OUD). Severity scores were highest for the OUD-TX group and lowest for the CP-RX group (Table 2). The mean severity score among these 73 patients was 6.0 (SD 2.7) for those whose index encounter was selected based on an OUD diagnosis and 0.6 (1.7) for those with an index encounter based on an opioid analgesic order.

Table 2. Average OUD[a] scores among 82 patients with annotated sentences.

Study group or index encounter reason[b]	Mean (SD) OUD score	Count of patients by OUD severity				PPVs[c] for moderate or severe OUD	
		None, n (%)	Mild, n (%)	Moderate, n (%)	Severe, n (%)	Patients with annotations	All sampled patients
CP-RX	3.4 (3.4)	8 (50)	0 (0)	1 (6)	7 (44)	0.50	0.40
CP-nonRX	4.5 (3.9)	7 (39)	0 (0)	1 (6)	10 (56)	0.61	0.55
OUD-DX	5.2 (2.5)	3 (16)	1 (5)	1 (5)	14 (74)	0.79	0.75
OUD-TX	6.8 (2.2)	1 (5)	1 (5)	1 (5)	17 (85)	0.90	0.90
Control	0.0 (0.0)	9 (100)	0 (0)	0 (0)	0 (0)	N/A[d]	N/A
OUD diagnosis	6.0 (2.7)	8 (13)	2 (3)	4 (7)	47 (77)	0.84	0.81
Opioid analgesic order	0.6 (1.7)	11 (92)	0 (0)	0 (0)	1 (1)	0.08	0.03

[a]OUD: opioid use disorder.

[b]Reason for selecting index encounter, either an OUD diagnosis or an order for an opioid analgesic.

[c]PPV: positive predictive value.

[d]N/A: not applicable.

The PPV for detecting moderate or severe OUD among the 73 noncontrol group patients with annotations was 0.71. PPVs were highest for OUD-TX group at 0.90 (Table 2). At 0.84, the PPV for detecting moderate or severe OUD using the notes of patients whose index encounter was selected based on an OUD diagnosis was higher than patients whose index encounter was based on an opioid analgesic order (PPV=0.08).

The mean number of notes per patient that were reviewed and annotated differed slightly by degree of severity (Table 3), but these differences were not statistically significant (1-way ANOVA for mean number of notes reviewed: $F_{3,78}$=0.60; P=.62; for mean number of notes annotated: $F_{3,56}$=1.46; P=.24). We observed no consistent patterns in the frequency of note types or encounter types by severity.

History of substance misuse was among the most prevalent classes across severity groups and was present in 56-126 (17%-38%) of the 332 notes (Table 3). For those with “no OUD,” the classes *psychiatric condition* and *daily tobacco use* were also common. The class *OUD treatment* automatically led to a classification of “severe OUD” and was present in 36 (16%) of the 226 notes among patients scored as “severe OUD.”

Table 3. Note characteristics by OUD[a] severity among 100 sampled study patients with 332 notes.

Note characteristics		Not annotated	OUD severity			
			None	Mild	Moderate	Severe
Patients, n						
		18	28	2	4	48
Notes reviewed, n						
		32	98	12	17	226
Number of notes reviewed per patient, mean (SD)						
		1.8 (1.2)	3.3 (2.3)	5.0 (0.0)	4.0 (3.0)	3.8 (2.3)
Total number of notes annotated, n						
		N/A[b]	90	12	15	214
Number of notes with annotations per patient, mean (SD)						
		N/A	2.3 (2.0)	5.0 (0.0)	3.8 (3.0)	3.3 (2.2)
Note type, n (%)						
	Progress notes	17 (53)	24 (24)	1 (8)	7 (41)	45 (20)
	H&P[c] notes	0 (0)	21 (21)	6 (5)	1 (6)	46 (20)
	ED[d] notes	4 (13)	9 (9)	1 (8)	2 (12)	33 (15)
	ED provider notes	3 (9)	22 (22)	2 (17)	5 (29)	28 (12)
	Discharge summaries	0 (0)	16 (16)	2 (17)	1 (6)	38 (17)
	ED triage notes	2 (6)	1 (1)	0 (0)	0 (0)	15 (7)
	Ancillary progress notes	0 (0)	1 (1)	0 (0)	0 (0)	11 (5)
	Communication notes	0 (0)	1 (1)	0 (0)	0 (0)	2 (1)
	OPT[e] clinic notes	1 (3)	2 (1)	0 (0)	0 (0)	1 (0)
	ED support staff notes	5 (16)	1 (1)	0 (0)	1 (6)	6 (3)
	Lactation notes	0 (0)	0 (0)	0 (0)	0 (0)	1 (0)
Encounter type, n (%)						
	OPT	18 (56)	25 (26)	1 (8)	3 (18)	31 (14)
	ED	14 (44)	33 (34)	5 (42)	10 (59)	97 (43)
	ED to IPT[f]	0 (0)	21 (21)	6 (50)	4 (24)	67 (30)
	IPT	0 (0)	18 (18)	0 (0)	0 (0)	31 (14)
Most frequent classes (percentage of notes with class)[a]						
		N/A	History of substance misuse (38%)	History of substance misuse (22%)	Psychiatric condition (22%)	History of substance misuse (18%)
		N/A	Psychiatric condition (31%)	Withdrawal (22%)	History of substance misuse (17%)	OUD treatment (16%)

[a]OUD: opioid use disorder.

[b]N/A: not applicable.

[c]H&P: history and progress.

[d]ED: emergency department.

[e]OPT: outpatient.

[f]IPT: inpatient.

Where Is OUD-Relevant Information Found in Clinical Notes and How Is it Documented Over Time?

Progress notes, the most common note type in the sample, had the largest total number of annotations, was among the highest yielding note types with an average of 9.3 annotations per note (SD 10.4), had the greatest diversity of classes represented, and was the only note type in which we observed the class *OUD* (signifying a definitive OUD diagnosis; Figure 3; Table 4). However, progress notes had a lower proportion of notes with annotations compared to other note types. H&P notes had the largest proportion of notes with annotations, followed by discharge summaries, ED provider notes, and ED notes (Table 4). These 4 note types also represented a large diversity of classes.

Figure 3. Heatmap of class frequencies by note type. ED: emergency department; H&P: history and progress; OUD: opioid use disorder.

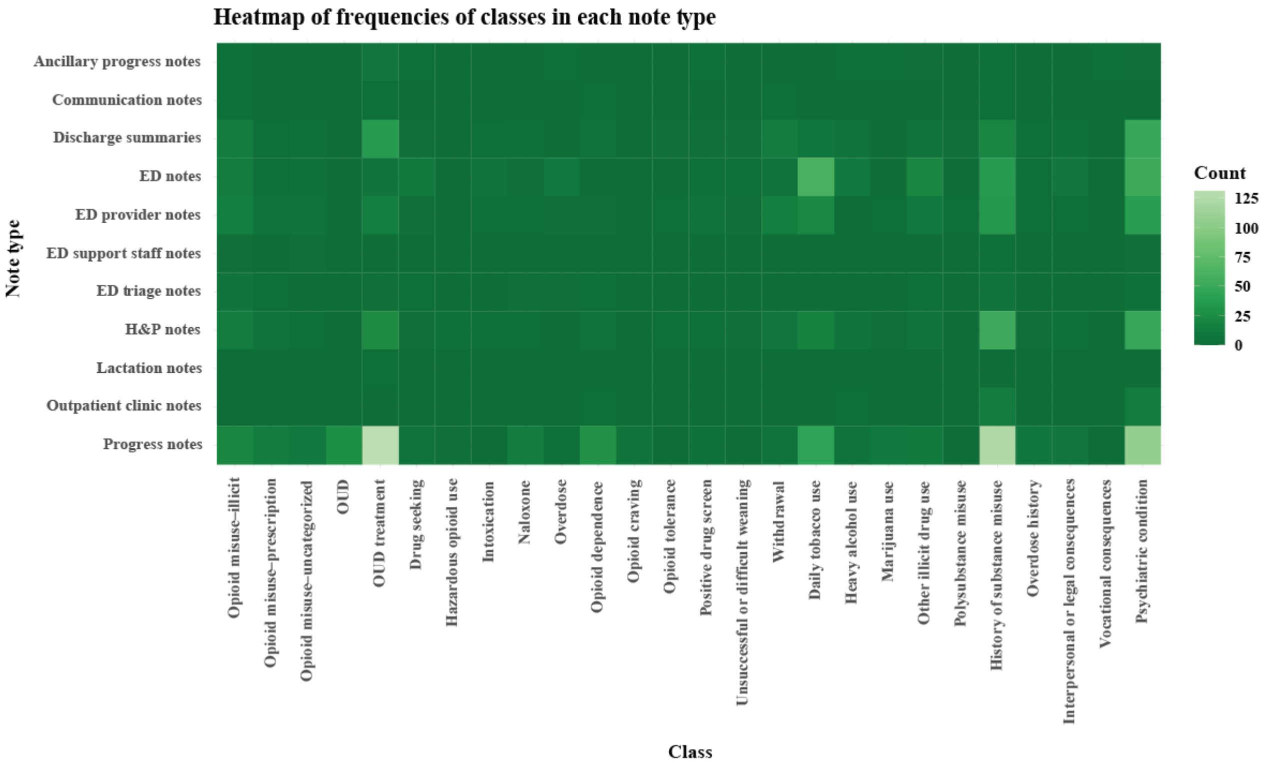


Table 4. Count of notes and annotated sentences by note type among the 1436 annotated sentences.

Note type	Number of notes	Number of notes with annotated sentences, n (%)	Counts of notes with annotated sentences by encounter setting (percentage of total)				Number of annotated sentences, n (%)	Mean (SD) number of annotated sentences per note
			ED[a], n (%)	ED to IPT[b], n (%)	OPT[c], n (%)	IPT, n (%)		
Progress notes	92	62 (67)	1 (2)	7 (11)	3 (5)	51 (82)	574 (40)	9.3 (10.4)
H&P[d] notes	30	27 (90)	8 (30)	13 (48)	6 (22)	0 (0)	196 (14)	7.3 (5.7)
ED notes	32	25 (78)	21 (84)	4 (16)	0 (0)	0 (0)	237 (17)	9.5 (9.9)
ED provider notes	28	23 (82)	18 (78)	5 (22)	0 (0)	0 (0)	177 (12)	7.7 (7.0)
Discharge summaries	26	23 (88)	5 (22)	10 (43)	8 (35)	0 (0)	172 (12)	7.5 (6.4)
ED triage notes	26	9 (35)	9 (100)	0 (0)	0 (0)	0 (0)	22 (2)	2.4 (1.7)
Ancillary progress notes	30	9 (30)	2 (22)	4 (44)	3 (33)	0 (0)	19 (1)	2.1 (1.2)
Communication notes	12	3 (25)	0 (0)	2 (67)	1 (33)	0 (0)	5 (0)	1.7 (0.6)
OPT clinic notes	4	2 (50)	0 (0)	0 (0)	0 (0)	2 (100)	26 (2)	13.0 (15.6)
ED support staff notes	38	2 (5)	1 (50)	1 (50)	0 (0)	0 (0)	7 (0)	3.5 (2.1)
Lactation notes	2	1 (50)	0 (0)	0 (0)	1 (100)	0 (0)	1 (0)	1.0 (N/A[e])

[a]ED: emergency department.

[b]IPT: inpatient.

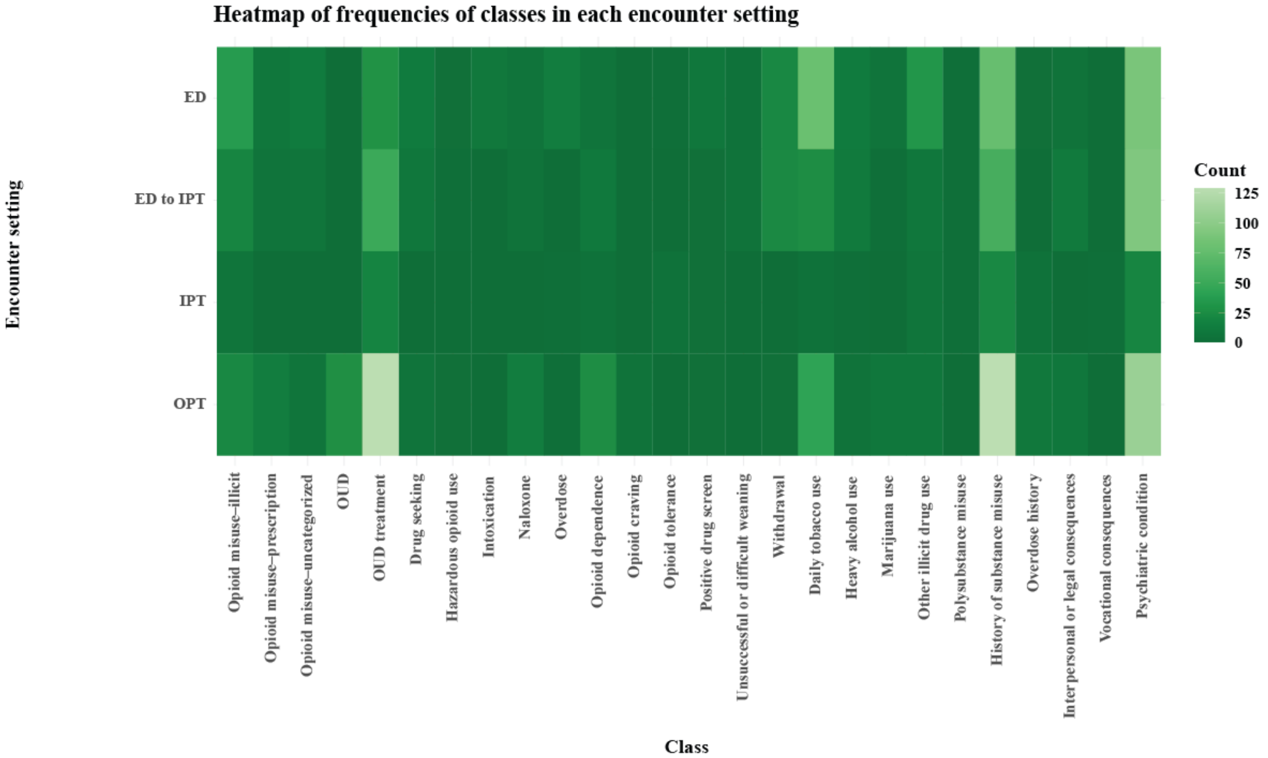
[c]OPT: outpatient.

[d]H&P: history and progress.

[e]N/A: not applicable.

Regarding encounter type, compared to the inpatient setting, notes from ED and outpatient settings had the highest proportions of notes with annotations and a high diversity of classes represented (Figure 4).

Figure 4. Heatmap of class frequencies by encounter setting. ED: emergency department; IPT: inpatient; OPT: outpatient; OUD: opioid use disorder.

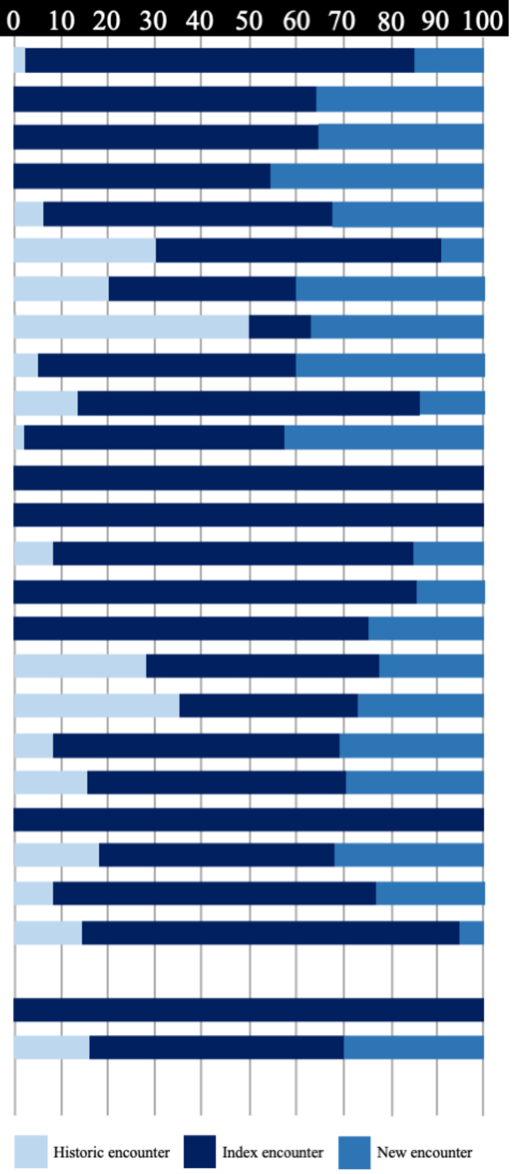


Some classes appeared across varied note types, such as *OUD treatment*, *daily tobacco use*, *history of substance misuse*, and *psychiatric condition*, whereas others tended to only appear in a particular note type (eg, *OUD, naloxone, opioid craving, and overdose history* in progress notes; *drug seeking* in ED notes).

The largest proportion of annotations was observed at the index encounter for all classes except *intoxication* (Figure 5). The classes *opioid craving*, *opioid tolerance*, *polysubstance misuse*, and *vocational consequences* were only observed at the index encounter. Annotations were more common in "new" versus "historic" encounters, and several classes were only observed in "new" encounters and not "historic" encounters.

Figure 5. Count of total annotated sentences by class and proportion by encounter order among the 1436 annotated sentences. Note selection was centered around an index encounter, during which a patient first met group inclusion criteria. “Historic” and “new” refer to notes from encounters immediately preceding and proceeding the index encounter, respectively. OUD: opioid use disorder.

Class type (Aggregate sum)	Class	Total (Annotated sentences)
Opioid misuse (380)	Opioid misuse - illicit	82
	Opioid misuse - prescription	25
	Opioid misuse - uncategorized	23
	OUD	26
	OUD treatment	224
Indications of current lack of control of opioid use or dependence (188)	Drug seeking	23
	Hazardous opioid use	5
	Intoxication	8
	Naloxone	22
	Overdose	15
	Opioid dependence	42
	Opioid craving	4
	Opioid tolerance	4
	Positive drug screen	13
	Unsuccessful or difficult weaning	7
	Withdrawal	45
Other or nonspecific present or past substance misuse (538)	Daily tobacco use	149
	Heavy alcohol use	26
	Marijuana use	13
	Other illicit drug use	52
	Polysubstance misuse	3
	History of substance misuse	282
	Overdose history	13
Consequences of opioid misuse (22)	Interpersonal or legal consequences	21
	Opioid-related medical issues	0
	Vocational consequences	1
Contributing factors (308)	Psychiatric condition	308
1436		



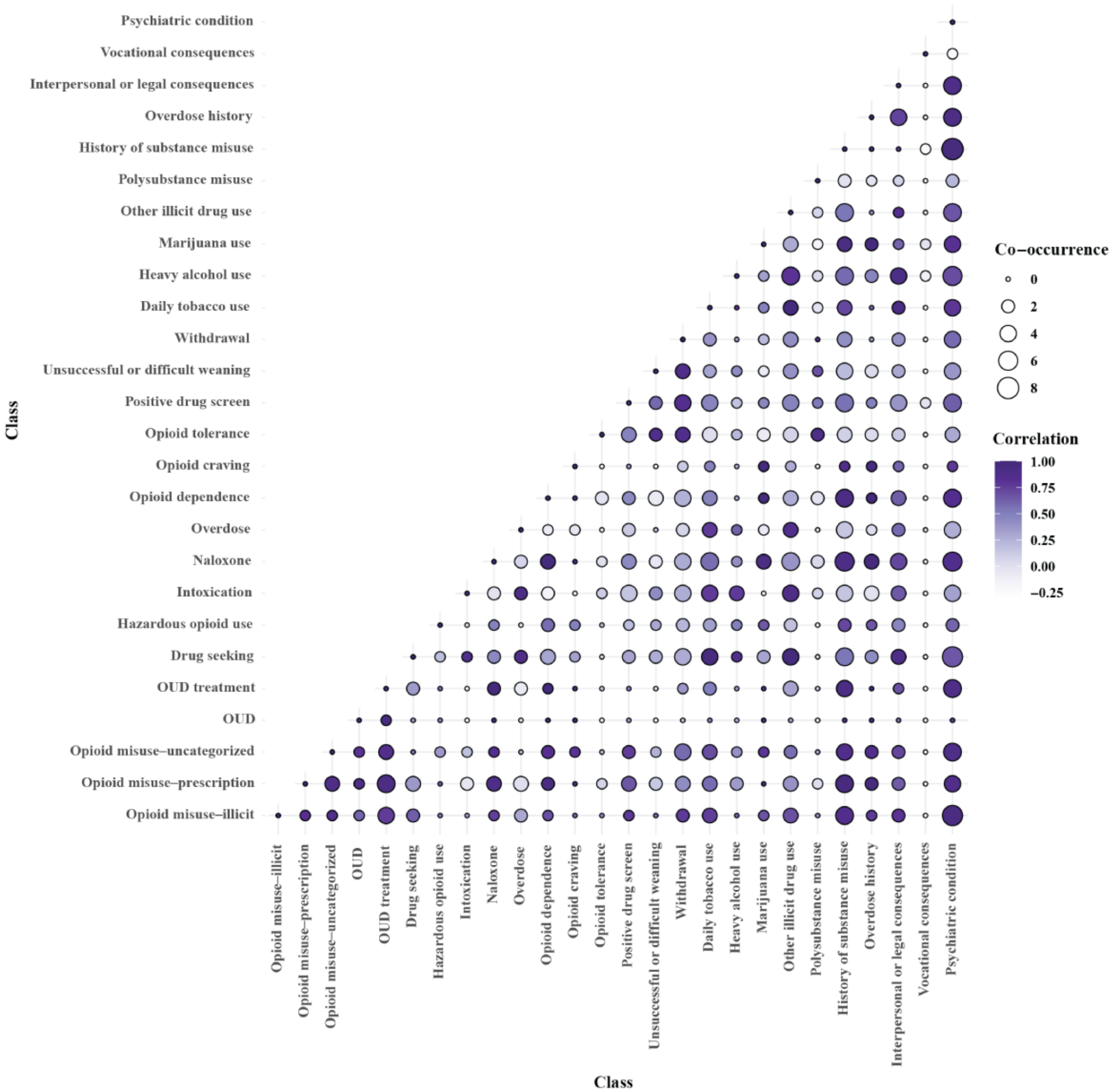
Which OUD-Related Concepts Are Most Common in Clinical Notes?

The largest number of annotations involved classes representing substance use not specific to opioids, with the most common classes being *history of substance misuse* and *daily tobacco use* (Figure 5). Except for the class *overdose history*, none of these classes contributed to the OUD severity score. Classes representing opioid misuse had the second largest number of annotations, with *OUD treatment* being the most common. *Psychiatric condition*, representing a contributing factor to OUD, was also commonly observed. Few annotations with classes representing consequences of opioid misuse occurred; this category included some of the least common classes including *vocational consequences* and *opioid-related medical issues*. Several of the least commonly assigned classes represented current lack of control of opioid use.

Which OUD-Related Concepts Are Found Together in Clinical Notes?

Several class pairs had highly correlated frequency distributions across note types, indicating similar documentation frequency (Figure 6). *Psychiatric condition*, *history of substance misuse*, *opioid misuse-illicit*, and *opioid misuse-prescription* were highly correlated with many other classes. Conversely, classes with the lowest correlations in their frequency distribution with other classes included *opioid tolerance*, *vocational consequences*, and *polysubstance misuse.*

Figure 6. Correlation between and co-occurrence of classes. Color depicts correlations between distributions of class pairs documented across note types, demonstrating how similarly 2 given classes are observed across the 11 note types. High correlation indicates classes are documented with similar frequency distributions across note types; conversely, low correlation indicates classes have different frequency distributions across notes. Bubble sizes depict class frequency co-occurrence at the patient level, showing how frequently 2 given classes are experienced across patients. Higher frequency ranges indicate 2 class pairs are more frequently experienced together by a given patient; conversely lower frequency ranges indicate 2 class pairs are less frequently experienced together. OUD: opioid use disorder.



When evaluated across patients, the classes *psychiatric condition* and *history of substance misuse* also had high frequency of co-occurrence with other classes, as did *other illicit drug use* (Figure 6). Classes with the lowest frequency of co-occurrence with other classes included *opioid tolerance*, *opioid craving*, *vocational consequences*, and *OUD*.

Discussion

Principal Results

Through development and evaluation of an annotation schema to characterize OUD severity and documentation of patterns of OUD-relevant information, this study illustrated severity can be inferred from a limited number of clinical notes. Severity, determined by capturing features associated with DSM-5 OUD severity criteria, followed the expected range of scores, with the highest severity observed for patients receiving OUD treatment and the lowest among those with prescriptions for chronic pain. While severity was determined using a range of note types, we found the most relevant information in outpatient notes typically used for acute care and within a subset of note types. The prevalence of schema classes varied widely, providing information regarding the concepts most useful for developing NLP tools.

Inferring Severity of Problematic Opioid Use

To our knowledge, this is the first study to develop an annotation schema to characterize OUD severity. Most patients received a severity score indicating severe OUD or no OUD. The paucity of scores indicating mild or moderate OUD may be explained by our approach to mapping classes to DSM-5 criteria, the selection of study cohorts, and the limited number of notes reviewed. The "crosswalk" of classes and DSM-5 criteria was based on prior work that included clinician input [11,27], but down weighting specific classes could be justified in future uses and would yield lower severity scores. Additionally, the resulting PPVs for moderate or severe OUD aligned with expectations for this study's cohorts. Control group individuals all had a severity score of 0, evidence of the specificity of our approach. The OUD-TX group's PPV of 0.90 is consistent with successful EHR-based algorithms for other conditions [32]. This was expected since these individuals received medication treatment for OUD, which, if documented in the notes we reviewed, would yield a score indicative of severe OUD. The OUD-DX group also had a high PPV, which is unsurprising given their OUD diagnosis. OUD is often underdiagnosed [14,15]; thus, the existence of an OUD diagnosis would be expected to indicate a true disorder, with signs and symptoms likely documented. That said, prior studies have found International Classification of Diseases codes are insufficient for accurately identifying OUD [11,12,14], which may explain why the PPV for OUD-DX was somewhat lower than the OUD-TX group. Both chronic pain groups had moderate PPVs and individuals in these groups were more likely to have a score indicating no OUD. This is likely explained by the index encounter reason—half of the CP-RX group and nearly a quarter of the CP-nonRX group were selected based on an opioid analgesic order. Most individuals whose index encounter was based on an opioid analgesic order had a score indicative of "no OUD." Although these 2 cohorts had OUD confirmed through a chart review process in prior studies [27,28], the confirmation was only for mild OUD. It is also probable that given the limited number of notes reviewed for this study, additional notes may have yielded information indicative of OUD, which may have occurred later than the 3 selected encounter dates. Based on these findings, we believe that using a limited number of notes is likely sufficient for characterizing OUD severity only in the presence of other confirming information, such as an OUD diagnosis or a medication order for OUD treatment.

Documentation Patterns of Information Relevant to OUD in Clinical Notes

To inform development of NLP methods, we explored documentation patterns of OUD-related information. We observed the most relevant information—as reflected by the proportion of notes with annotated sentences—were in ambulatory and outpatient settings, as opposed to the inpatient setting. This suggests clinical notes from ED and outpatient settings may yield more relevant information regarding OUD; however, we do not recommend excluding notes from the inpatient setting given that our pilot work demonstrated the ubiquity of OUD-relevant information in hospital discharge summaries [18]. Regarding note type, H&P notes, discharge summaries, ED provider notes, and ED notes yielded the most information pertaining to OUD, and the most diverse information. Finally, relevant information was most often found within notes from the index encounter and the following encounter, demonstrating the importance of selecting notes in reference to relevant structured EHR information and suggesting the sensitivity of OUD severity scores may improve with review of additional notes following the index encounter. The index encounter was defined based on an OUD diagnostic code or the patient's most recent opioid analgesic order; having such a meaningful anchor may be particularly important for characterizing severity with limited notes. To optimize efficiency, future studies of OUD should consider focusing on the notes likely to yield the most relevant information. However, our findings should be considered preliminary, as we did not surveil all notes, opting for those surrounding the index encounter. Documentation patterns were also likely influenced by the reason for the encounter (eg, whether related to pain management versus medication treatment for OUD). Thus, findings may differ with a review of a patient's full history of clinical notes.

Opioid misuse, nonopioid substance abuse, and psychiatric classes were the most common annotations in our cohort. The prevalence of opioid misuse classes is not unexpected given that 4 of 5 groups had confirmed OUD. Relatedly, substance use disorders and psychiatric disorders tend to be comorbid with OUD [33]. That said, *history of substance misuse* and *psychiatric condition* were also the most common classes among individuals characterized as having no OUD based on severity criteria. Although nonopioid substance use disorders and psychiatric disorders are risk factors for OUD, they are also commonly comorbid with one another [33]. Indeed, genetic predispositions for substance use disorders, including OUD, and psychiatric disorders can be shared via genetic pleiotropy [34]. The high prevalence and co-occurrence of substance use disorders and psychiatric disorders was also observed across note types and patients. Taken together, these results highlight the importance of substance use and psychiatric classes for developing OUD-related NLP tools.

Absence of certain concepts is also informative for NLP tool development. Consistent with previous work [18,27], some classes relevant to OUD severity were rarely observed, including those representing consequences of opioid misuse and lack of control of opioid use. Reviewing a larger set of notes in an expanded patient sample could yield more frequent documentation of some classes, but some classes may represent concepts not traditionally recorded by clinicians (eg, *vocational consequences*). Documentation by clinicians of additional concepts such as the vocational, social, legal, and medical consequences patients face due to opioid use could be useful for patient care, particularly given the potential utility of such information in characterizing OUD severity. Future work to develop NLP frameworks related to OUD should consider that information captured in clinical notes may change over time with secular changes that occur in the drug landscape, clinical practices and documentation, and patients' care-seeking behaviors.

Limitations

This study used clinical notes from an integrated health system, allowing for review of multiple note types across encounter settings. Findings may be different in health systems in which information is not integrated across settings; for example, opioid analgesic orders may be missing and OUD treatment siloed. Although we evaluated cohorts with varying evidence of opioid misuse, inclusion of other patient cohorts may yield different findings. In particular, because the cohorts were originally assembled for a genomic study, the CP-RX and CP-nonRX study groups included individuals only of European ancestry, limiting the generalizability of findings to other racial groups. Replication studies are necessary to understand whether study findings generalize to other settings and patient populations.

Conclusions

Understanding how and where OUD-relevant information is captured in EHRs is essential to informing development of NLP tools to identify and characterize the severity of OUD. We developed an annotation schema to determine OUD severity and highlighted document patterns of OUD-relevant information in clinical notes, which may be informative for future NLP frameworks related to OUD. Findings suggest OUD-relevant information is more prevalent in a subset of note types in ambulatory and outpatient settings—particularly H&P notes, discharge summaries, ED provider notes, and ED notes—and that certain information relevant to OUD may only be captured in certain note types or may be infrequently documented. Furthermore, when reviewing a limited number of notes, having a meaningful anchor such as an OUD diagnostic code or recent opioid analgesic order is important for characterizing OUD severity. Findings also demonstrate the potential for inferring severity of OUD from key information contained in a limited number of clinical notes, paving the way for development of informatics solutions to improve the prevention, diagnosis, and treatment of OUD through clinical care.

Acknowledgments

MNP received funding to complete this work from the National Institute on Drug Abuse of the National Institutes of Health (K01DA049903). VT received funding to complete this work from the National Institute on Drug Abuse (R01DA044015 and P50DA054071) and from the Pennsylvania Department of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Authors' Contributions

All authors participated in the conceptualization of this study and its methodology and contributed to the original draft of the paper. MNP was responsible for funding acquisition and project administration. MNP and PJF carried out the annotation work and analysis and they created the visualizations used in this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Crosswalk of class or attribute combinations and Diagnostic and Statistical Manual of Mental Disorders, 5th edition (DSM-5) criteria for opioid use disorder (OUD).

[DOCX File , 18 KB - mental_v11i1e53366_app1.docx]

Multimedia Appendix 2

Classes with brief definitions, example annotated sentences, and the count of annotated sentences by class or attribute.

[DOCX File , 19 KB - mental_v11i1e53366_app2.docx]

References

1. Jalali MS, Botticelli M, Hwang RC, Koh HK, McHugh RK. The opioid crisis: a contextual, social-ecological framework. Health Res Policy Syst 2020 Aug 06;18(1):87 [FREE Full text] [doi: 10.1186/s12961-020-00596-8] [Medline: 32762700]
2. CDC Wonder online database 2021. Centers for Disease Control and Prevention. 2023. URL: https://wonder.cdc.gov/ [accessed 2023-12-22]
3. National survey on drug use and health 2021 (NSDUH-2016-DS001). Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality. 2021. URL: https://www.datafiles.samhsa.gov/dataset/national-survey-drug-use-and-health-2021-nsduh-2021-ds0001 [accessed 2023-12-23]
4. Florence C, Luo F, Rice K. The economic burden of opioid use disorder and fatal opioid overdose in the United States, 2017. Drug Alcohol Depend 2021;218:108350 [FREE Full text] [doi: 10.1016/j.drugalcdep.2020.108350] [Medline: 33121867]
5. Strain E, Peavy KM. Opioid use disorder: treatment overview. UpToDate. 2023. URL: https://www.uptodate.com/contents/opioid-use-disorder-treatment-overview [accessed 2023-12-22]

6. Marsden J, Tai B, Ali R, Hu L, Rush AJ, Volkow N. Measurement-based care using DSM-5 for opioid use disorder: can we make opioid medication treatment more effective? Addiction 2019;114(8):1346-1353 [FREE Full text] [doi: 10.1111/add.14546] [Medline: 30614096]
7. Rice JB, White AG, Birnbaum HG, Schiller M, Brown DA, Roland CL. A model to identify patients at risk for prescription opioid abuse, dependence, and misuse. Pain Med 2012;13(9):1162-1173 [FREE Full text] [doi: 10.1111/j.1526-4637.2012.01450.x] [Medline: 22845054]
8. Hasan MM, Young GJ, Patel MR, Modestino AS, Sanchez LD, Noor-E-Alam M. A machine learning framework to predict the risk of opioid use disorder. Mach Learn Appl 2021;6:100144 [FREE Full text] [doi: 10.1016/j.mlwa.2021.100144]
9. Carrell DS, Albertson-Junkans L, Ramaprasan A, Scull G, Mackwood M, Johnson E, et al. Measuring problem prescription opioid use among patients receiving long-term opioid analgesic treatment: development and evaluation of an algorithm for use in EHR and claims data. J Drug Assess 2020;9(1):97-105 [FREE Full text] [doi: 10.1080/21556660.2020.1750419] [Medline: 32489718]
10. Canan C, Polinski JM, Alexander GC, Kowal MK, Brennan TA, Shrank WH. Automatable algorithms to identify nonmedical opioid use using electronic data: a systematic review. J Am Med Inform Assoc 2017;24(6):1204-1210 [FREE Full text] [doi: 10.1093/jamia/ocx066] [Medline: 29016967]
11. Poulsen MN, Nordberg CM, Troiani V, Berrettini W, Asdell PB, Schwartz BS. Identification of opioid use disorder using electronic health records: beyond diagnostic codes. Drug Alcohol Depend 2023;251:110950. [doi: 10.1016/j.drugalcdep.2023.110950] [Medline: 37716289]
12. Schirle L, Jeffery A, Yaqoob A, Sanchez-Roige S, Samuels DC. Two data-driven approaches to identifying the spectrum of problematic opioid use: a pilot study within a chronic pain cohort. Int J Med Inform 2021;156:104621 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104621] [Medline: 34673309]
13. Lo-Ciganic WH, Donohue JM, Yang Q, Huang JL, Chang CY, Weiss JC, et al. Developing and validating a machine-learning algorithm to predict opioid overdose in Medicaid beneficiaries in two US states: a prognostic modelling study. Lancet Digit Health 2022;4(6):e455-e465 [FREE Full text] [doi: 10.1016/S2589-7500(22)00062-0] [Medline: 35623798]
14. Arifkhanova A, Prieto JT, Davidson AJ, Al-Tayyib A, Hawkins E, Kraus E, et al. Defining opioid-related problems using a health care safety net institution's inpatient electronic health records: limitations of diagnosis-based definitions. J Addict Med 2023;17(1):79-84. [doi: 10.1097/ADM.0000000000001041] [Medline: 35914026]
15. Rowe CL, Santos GM, Kornbluh W, Bhardwaj S, Faul M, Coffin PO. Using ICD-10-CM codes to detect illicit substance use: a comparison with retrospective self-report. Drug Alcohol Depend 2021;221:108537. [doi: 10.1016/j.drugalcdep.2021.108537] [Medline: 33621806]
16. Brat GA, Agniel D, Beam A, Yorkgitis B, Bicket M, Homer M, et al. Postsurgical prescriptions for opioid naive patients and association with overdose and misuse: retrospective cohort study. BMJ 2018;360:j5790 [FREE Full text] [doi: 10.1136/bmj.j5790] [Medline: 29343479]
17. Tetsunaga T, Tetsunaga T, Nishida K, Kanzaki H, Misawa H, Takigawa T, et al. Drug dependence in patients with chronic pain: a retrospective study. Medicine (Baltimore) 2018;97(40):e12748 [FREE Full text] [doi: 10.1097/MD.0000000000012748] [Medline: 30290690]
18. Poulsen MN, Freda PJ, Troiani V, Davoudi A, Mowery DL. Classifying characteristics of opioid use disorder from hospital discharge summaries using natural language processing. Front Public Health 2022;10:850619 [FREE Full text] [doi: 10.3389/fpubh.2022.850619] [Medline: 35615042]
19. Singleton J, Li C, Akpunonu PD, Abner EL, Kucharska-Newton AM. Using natural language processing to identify opioid use disorder in electronic health record data. Int J Med Inform 2023;170:104963. [doi: 10.1016/j.ijmedinf.2022.104963] [Medline: 36521420]
20. Velupillai S, Mowery D, South BR, Kvist M, Dalianis H. Recent advances in clinical natural language processing in support of semantic analysis. Yearb Med Inform 2015;10(1):183-193 [FREE Full text] [doi: 10.15265/IY-2015-009] [Medline: 26293867]
21. Alzeer A, Jones JF, Bair MJ, Liu X, Alfantouck LA, Patel J, et al. A comparison of text mining versus diagnostic codes to identify opioid use problem: a retrospective study. Research Square. Preprint posted online on March 05 2020 2020 [FREE Full text] [doi: 10.21203/rs.3.rs-16124/v1]
22. Carrell DS, Cronkite D, Palmer RE, Saunders K, Gross DE, Masters ET, et al. Using natural language processing to identify problem usage of prescription opioids. Int J Med Inform 2015;84(12):1057-1064. [doi: 10.1016/j.ijmedinf.2015.09.002] [Medline: 26456569]
23. Haller IV, Renier CM, Juusola M, Hitz P, Steffen W, Asmus MJ, et al. Enhancing risk assessment in patients receiving chronic opioid analgesic therapy using natural language processing. Pain Med 2017;18(10):1952-1960 [FREE Full text] [doi: 10.1093/pm/pnw283] [Medline: 28034982]
24. Hylan TR, Von Korff M, Saunders K, Masters E, Palmer RE, Carrell D, et al. Automated prediction of risk for problem opioid use in a primary care setting. J Pain 2015;16(4):380-387 [FREE Full text] [doi: 10.1016/j.jpain.2015.01.011] [Medline: 25640294]
25. Lingeman JM, Wang P, Becker W, Yu H. Detecting opioid-related aberrant behavior using natural language processing. AMIA Annu Symp Proc 2018;2017:1179-1185 [FREE Full text] [Medline: 29854186]

26. Sharma B, Dligach D, Swope K, Salisbury-Afshar E, Karnik NS, Joyce C, et al. Publicly available machine learning models for identifying opioid misuse from the clinical notes of hospitalized patients. BMC Med Inform Decis Mak 2020;20(1):79 [FREE Full text] [doi: 10.1186/s12911-020-1099-y] [Medline: 32349766]
27. Palumbo SA, Adamson KM, Krishnamurthy S, Manoharan S, Beiler D, Seiwell A, et al. Assessment of probable opioid use disorder using electronic health record documentation. JAMA Netw Open 2020;3(9):e2015909 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.15909] [Medline: 32886123]
28. Tormohlen KN, Mojtabai R, Seiwell A, McGinty EE, Stuart EA, Tobin KE, et al. Co-occurring opioid use and depressive disorders: patient characteristics and co-occurring health conditions. J Dual Diagn 2021;17(4):296-303 [FREE Full text] [doi: 10.1080/15504263.2021.1979349] [Medline: 34581663]
29. Jennings MV, Lee H, Rocha DB, Bianchi SB, Coombes BJ, Crist RC, et al. Identifying high-risk comorbidities associated with opioid use patterns using electronic health record prescription data. Complex Psychiatry 2022;8(1-2):47-55 [FREE Full text] [doi: 10.1159/000525313] [Medline: 36545045]
30. Norgeot B, Muenzen K, Peterson TA, Fan X, Glicksberg BS, Schenk G, et al. Protected health information filter (Philter): accurately and securely de-identifying free-text clinical notes. NPJ Digit Med 2020;3:57 [FREE Full text] [doi: 10.1038/s41746-020-0258-y] [Medline: 32337372]
31. South B, Shen S, Leng J, Forbush T, DuVall S, Chapman W. A prototype tool set to support machine-assisted annotation. Stroudsburg, PA: Association for Computational Linguistics; 2012 Presented at: BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing; June 8, 2012; Montréal, Canada p. 130-139.
32. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc 2013;20(e1):e147-e154 [FREE Full text] [doi: 10.1136/amiajnl-2012-000896] [Medline: 23531748]
33. Freda PJ, Moore JH, Kranzler HR. The phenomics and genetics of addictive and affective comorbidity in opioid use disorder. Drug Alcohol Depend 2021;221:108602 [FREE Full text] [doi: 10.1016/j.drugalcdep.2021.108602] [Medline: 33652377]
34. Hu R, Dai Y, Jia P, Zhao Z. ANCO-GeneDB: annotations and comprehensive analysis of candidate genes for alcohol, nicotine, cocaine and opioid dependence. Database (Oxford) 2018;2018:bay121 [FREE Full text] [doi: 10.1093/database/bay121] [Medline: 30403795]

Abbreviations

DSM-5: Diagnostic and Statistical Manual of Mental Disorders, 5th edition
ED: emergency department
EHR: electronic health record
H&P: history and progress
IAA: interannotator agreement
NLP: natural language processing
OUD: opioid use disorder
PPV: positive predictive value

Edited by J Torous; submitted 06.10.23; peer-reviewed by A Jeffery, P Imhof; comments to author 22.11.23; revised version received 30.11.23; accepted 02.12.23; published 15.01.24.

Please cite as:
Poulsen MN, Freda PJ, Troiani V, Mowery DL
Developing a Framework to Infer Opioid Use Disorder Severity From Clinical Notes to Inform Natural Language Processing Methods: Characterization Study
JMIR Ment Health 2024;11:e53366
URL: https://mental.jmir.org/2024/1/e53366
doi:10.2196/53366
PMID:38224481

Original Paper

Feasibility, Acceptability, and Potential Efficacy of a Self-Guided Internet-Delivered Dialectical Behavior Therapy Intervention for Substance Use Disorders: Randomized Controlled Trial

Alexander R Daros[1], PhD; Timothy H Guimond[2,3], MD, PhD; Christina Yager[2], MSW; Emma H Palermo[4], BA; Chelsey R Wilks[5], PhD; Lena C Quilty[1,3], PhD

[1]Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, ON, Canada

[2]Addictions Division, Centre for Addiction and Mental Health, Toronto, ON, Canada

[3]Department of Psychiatry, University of Toronto, Toronto, ON, Canada

[4]Penn Center for Mental Health, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

[5]Department of Psychological Science, University of Missouri-St Louis, St. Louis, MO, United States

Corresponding Author:
Alexander R Daros, PhD
Campbell Family Mental Health Research Institute
Centre for Addiction and Mental Health
1025 Queen Street West
Toronto, ON, M6J1H1
Canada
Phone: 1 5192533000 ext 2236
Email: daros.alexander@gmail.com

Abstract

Background: People with alcohol and substance use disorders (SUDs) often have underlying difficulties in regulating emotions. Although dialectical behavioral therapy is effective for SUDs, it is often difficult to access. Self-guided, internet-delivered dialectical behavioral therapy (iDBT) allows for expanded availability, but few studies have rigorously evaluated it in individuals with SUDs.

Objective: This study examines the feasibility, acceptability, and potential efficacy of an iDBT intervention in treatment-seeking adults with SUDs. We hypothesized that iDBT would be feasible, credible, acceptable, and engaging to people with SUDs. We also hypothesized that the immediate versus delayed iDBT group would show comparatively greater improvements and that both groups would show significant improvements over time.

Methods: A 12-week, single-blinded, parallel-arm, randomized controlled trial was implemented, with assessments at baseline and at 4 (acute), 8, and 12 weeks (follow-up). A total of 72 community adults aged 18 to 64 years were randomized. The immediate group (n=38) received access to iDBT at baseline, and the delayed group (n=34) received access after 4 weeks. The intervention (*Pocket Skills* 2.0) was a self-guided iDBT via a website, with immediate access to all content, additional text and email reminders, and additional support meetings as requested. Our primary outcome was substance dependence, with secondary outcomes pertaining to feasibility, clinical outcomes, functional disability, and emotion dysregulation, among other measures. All outcomes were assessed using self-report questionnaires.

Results: iDBT was perceived as a credible and acceptable treatment. In terms of feasibility, 94% (68/72) of the participants started iDBT, 13% (9/68) were early dropouts, 35% (24/68) used it for the recommended 8 days in the first month, and 50% (34/68) were still active 4 weeks later. On average, the participants used iDBT for 2 hours and 24 minutes across 10 separate days. In the acute period, no greater benefit was found for the immediate group on substance dependence, although we did find lower depression (b=−2.46; P=.02) and anxiety (b=−2.22; P=.02). At follow-up, there were greater benefits in terms of reduced alcohol (b=−2.00; P=.02) and nonalcoholic substance (b=−3.74; P=.01) consumption in the immediate access group. Both groups demonstrated improvements in substance dependence in the acute (b=−1.73; P<.001) and follow-up period (b=−2.09; P<.001). At follow-up, both groups reported reduced depression, anxiety, suicidal behaviors, emotional dysregulation, and functional disability.

Conclusions: iDBT is a feasible and acceptable intervention for patients with SUDs, although methods for improving engagement are warranted. Although results did not support efficacy for the primary outcome at 4 weeks, findings support reductions in substance dependence and other mental health concerns at 12 weeks. Notwithstanding the limitations of this study, the results suggest the potential value of iDBT in the treatment of SUDs and other mental health conditions.

Trial Registration: ClinicalTrials.gov NCT05094440; https://clinicaltrials.gov/show/NCT05094440

(JMIR Ment Health 2024;11:e50399) doi:10.2196/50399

KEYWORDS

depression; anxiety; emotion dysregulation; digital interventions; dialectical behavior therapy; substance use disorder; alcohol use disorder; randomized controlled trial; eHealth; mobile phone

Introduction

Background

Alcohol and substance use disorders (SUDs) are the leading causes of death and disability worldwide [1,2]. These conditions are often chronic, leading to elevated risks of co-occurring medical and mental health conditions, involvement with the criminal justice system, and loss of workplace productivity [1-4]. In 2019, the past-year use of alcohol, cannabis, tobacco, and illicit substances was 77%, 21%, 14%, and 3.6%, respectively, in Canadians [5]. Increased consumption during the COVID-19 pandemic in Canada and around the world has been linked to greater substance-related harms and concurrent mental health symptoms, such as depression, anxiety, and hopelessness [6]. Various evidence-based psychological treatments are available for SUDs; however, the availability and demand for these services come at a time when internet and mobile delivery formats are being promoted in care pathways [7]. These formats hold considerable public health promise in reducing the burden associated with SUDs. For example, a recent systematic review highlighted that existing mobile interventions were effective and rated as acceptable by people with SUDs [8].

Psychological Treatments for SUDs

Although pharmacological treatments exist for some substances (eg, alcohol and opioids), they have mixed evidence in treating other SUDs (eg, cannabis and stimulants [9]). Thus, psychological treatments remain a necessary therapeutic avenue for SUDs and may be particularly promising for those with multiple substance use concerns. Although psychological treatments vary greatly in their approach and theoretical framework, they tend to produce moderate effect size reductions in substance dependence [10,11]. To date, the greatest evidence supports cognitive behavioral and motivational enhancement approaches for treating SUDs.

SUDs rarely occur in isolation and often co-occur with depressive, anxiety, bipolar, and traumatic stressor disorders [12]. Psychological treatments are well suited to treat multiple conditions simultaneously when they incorporate a transdiagnostic focus or approach. There is growing consensus that people with SUDs, regardless of a specific substance, report higher difficulties in regulating their emotions compared with control samples and often use alcohol or other substances to cope with negative emotions [13]. More broadly, difficulties in emotion regulation appear to be a transdiagnostic risk factor underlying not only the development and course of SUDs but also depressive, anxiety, bipolar, and traumatic stressor disorders [14,15]. They also represent a promising treatment target, as emotion regulation skills tend to improve during psychological treatments for SUDs, along with more general improvements in self-efficacy and coping [16,17]. One psychological intervention that may be of substantial interest is dialectical behavior therapy (DBT), which was developed to treat individuals with high emotion dysregulation and includes comprehensive skills training in the domains of mindfulness, distress tolerance, emotion regulation, and interpersonal effectiveness.

DBT is a third-wave psychological intervention designed for patients with complex and severe behavioral, emotional, and interpersonal dysfunction [18,19]. DBT was first developed and found to be effective for severe clinical presentations related to suicidal behavior, nonsuicidal self-injury, and borderline personality disorder in adolescents and adults (refer to the study by Neacsiu et al [20] for review). Over time, DBT was reconceptualized as a transdiagnostic intervention appropriate for other mental health conditions and now includes specific content relevant to SUDs as well as other addictive behaviors [16,21,22]. Nevertheless, outpatient programs offering DBT are often safeguarded for those with acute suicide risk and behavioral problems. Importantly, although DBT was originally developed as a year-long multimodal intervention, evidence suggests that relatively brief formats focusing on DBT skills training (eg, 8-32 wk) are effective in treating SUDs, either as a primary condition or a co-occurring presentation in numerous clinical trials [16,22-24]. Despite these promising results, further research is needed to support the potential benefits of digital formats of DBT skills training, particularly within inclusive samples that reflect those seeking support for SUD.

Internet-Delivered DBT

Another way to increase the availability of DBT is through internet and mobile delivery formats. Thus far, research on internet-delivered DBT (iDBT) has been promising. In a review of 11 studies, iDBT was feasible and effective, although these results were based on small sample sizes, and few studies adopted a more rigorous methodology (eg, randomized controlled trials [RCTs] [25]). Various methods have been used, such as therapist-led sessions delivered via web-based videoconferencing [26], asynchronous material delivered via email [27], self-guided stand-alone websites [28,29], and therapist-guided programs [30]. Studies that evaluated potential efficacy suggested that iDBT was at least as effective as control conditions (waitlist or face-to-face) and was accepted by users. However, web-based delivery is not without harm or adverse events. One large-scale trial comparing integrated care management and skills training (ie, 4 self-guided DBT skills) for those with suicidal ideation found that the latter condition led to an increased risk of self-harm [31]. A discussion of the study suggested that it faced, among other issues, an implementation failure [32]. Thus, these and other considerations should be incorporated in future work.

In a seminal study, Wilks et al [30] evaluated therapist-guided iDBT in a sample of participants who are suicidal and alcohol dependent in a completely remote manner. This 8-week waitlist-controlled RCT delivered video trainings on mindfulness (2 wk), addiction (2 wk), emotion regulation (3 wk), and distress tolerance (1 wk) using an e-learning web-based platform along with handouts and worksheets delivered via email. The content was developed in collaboration with the developer of DBT. The intervention produced significant reductions in suicidal ideation, alcohol consumption, and emotion dysregulation. Although the treatment was deemed safe and acceptable to participants, there was substantial dropout, and technical issues were reported as a barrier to adherence [33]. Nevertheless, those who remained in the study reported that it was useful.

Following this work, a more advanced iDBT intervention called *Pocket Skills* (version 1.0) was created to overcome the accessibility and engagement issues encountered previously [34]. It is available through an internet browser on any device (ie, computer, tablet, or smartphone) and offers an interactive experience by using a chatbot along with embedded video lessons and practice. *Pocket Skills* 1.0 was evaluated in a single-arm trial as an adjunct intervention in individuals with a range of mental disorders completing in-person DBT for 4 weeks. The results of the study were promising, with both quantitative and qualitative evidence for its feasibility, acceptability, and potential use as an adjunct. We developed this study based on these 2 previous studies.

Current Study

This study aims to evaluate version 2.0 of *Pocket Skills* and advance the literature in several ways. First, the current investigation evaluates *Pocket Skills* 2.0, which includes some of the content from version 1.0, as well as revised and novel materials that have not yet been evaluated. Second, the delivery of iDBT in this study was predominantly self-guided, with limited therapist guidance compared with the previous trial that used iDBT intervention as a therapeutic adjunct [34]. Third, this investigation represented a more controlled study of *Pocket Skills* 2.0 as a stand-alone treatment in a sample of treatment-seeking adults with SUDs who were not receiving any other forms of psychological treatments. Finally, this investigation randomized participants to immediate versus delayed access to advance the previous single-arm study. A 12-week single-blinded parallel-arm waitlist-controlled RCT was initiated, with participants randomized to receive immediate access to the intervention or delayed access after 4 weeks. The 4-week intervention and follow-up periods are in line with previous implementations of self-guided digital mental health interventions [35-37]. These studies have found that attrition rates start to increase steadily after 4 weeks and especially after 7 to 8 weeks (eg, >50%).

Specifically, we hypothesized that greater than 50% of participants would start the intervention (H1a); not drop out early (H1b); engage with the intervention at a recommended dose of twice a week (or 8 d) in the first 4 weeks (H1c); and would still be using the intervention after 4 weeks (H1d). We also hypothesized that participants would rate the intervention as credible and acceptable on established measures (H1e). Second, we hypothesized that (H2a) participants in the immediate versus delayed iDBT group would show significantly greater improvements in our primary outcome of substance use dependence at the acute (week 4) and follow-up periods (week 12) in the form of an interaction effect (group×time). In addition, we hypothesized (H2b) significantly greater improvements for the immediate versus delayed iDBT group for our secondary outcomes (ie, depression, anxiety, emotion dysregulation, suicidality, functional disability, dispositional mindfulness, DBT skills, risky behaviors, and frequency of alcohol and substance use). Third, we hypothesized that iDBT would (H3) produce significant main effect improvements in both groups in the acute (week 4) and follow-up phases (week 12) of the intervention for all outcome measures.

Methods

Study Design

A 2-arm, single-blinded, parallel-group, preregistered RCT design was implemented, comparing individuals who received iDBT immediately with those who were first wait-listed for 4 weeks and then offered the intervention (delayed iDBT group). Assessments were completed at baseline and at 4 weeks, with additional follow-ups at 8 and 12 weeks. A CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist was completed with more detailed information on the study design (Multimedia Appendix 1).

Ethical Considerations

All study procedures were approved by the Centre for Addiction and Mental Health research ethics board (#016/2021), and this research complied with the Declaration of Helsinki of 1975, as revised in 2000. Data is stored in a de-identified format to safeguard participant information.

Participants and Recruitment

Enrollment ran from August 2022 to March 2023, and all follow-ups were completed by June 2023. Participants were recruited from psychiatric hospital clinician referrals, waitlists, and research registries and from the surrounding community through several methods of advertisement (eg, hospital and other websites, social media posts, private DBT clinics, and local community organizations). All advertisements sought individuals who wanted to reduce their alcohol or substance use and specifically stated that they would be offered an internet-delivered intervention.

All prospective participants were initially informed about the study and were prescreened for eligibility over the phone.

Inclusion criteria were as follows: (1) aged 18 to 65 years; (2) fluent in English; (3) understanding and willingness to comply with study requirements; (4) referred to addictions programming at our hospital or seeking treatment from the community, but not currently receiving any CBT or DBT intervention (support groups and psychiatric services were allowed); (5) alcohol or SUD in the past year; (6) use of alcohol or substance in the past month; (7) access to the internet (and assumed literacy); and (8) at least contemplation levels of wanting to reduce alcohol or substance use on the Contemplation Ladder measure [38]. Exclusion criteria included (1) any known practical factors that would preclude participation, (2) acute psychiatric (ie, suicidality, psychotic disorder) or medical condition (ie, acute intoxication or withdrawal) requiring medical attention, and (3) participation in another psychological intervention or treatment study. We did not exclude participants based on whether they were taking psychotropic medications or not.

Registration

The trial was registered with the ClinicalTrials.gov database (NCT05094440) on October 14, 2021. A revised registration was published on September 6, 2022, in line with changes to our protocol between our pilot study and this study. In this study, our analysis focused on the measures included in registration. One modification of the registered protocol was made, that is, the addition of a DBT skills measure at all time points to permit the evaluation of how this intervention was linked to changes in this key treatment target. Feasibility, acceptability, and engagement metrics were decided a priori for study implementation and were included in our study-specific protocol.

Randomization and Blinding

Participants were randomized to immediate or delayed iDBT using a blinded envelope system to ensure allocation concealment, with 3 randomization blocks (4, 6, and 8 participants). The randomization procedure was blinded to the participants and the experimenter who ran all baseline sessions (ARD). Thus, neither party knew which group the participant would be allocated to until after the informed consent and baseline procedures were completed. None of the participants withdrew immediately following randomization. The experimenters were not blinded to the procedures following the baseline session, including the follow-up assessments and contact. All follow-up assessments were conducted remotely and consisted solely of self-report measures.

Procedure

Eligible participants attended a 45-minute baseline session via a videoconference, where they provided informed consent (electronically), completed a demographic questionnaire and semistructured diagnostic interview, and were randomized into either immediate or delayed access groups. At the end of the baseline session, those randomized to the immediate group were provided the iDBT website URL and an invitation code and completed the sign-in procedure (15 min) with the experimenter during the videoconference call. Those randomized to the delayed access group were scheduled for an additional appointment in 4 weeks, where they met with the experimenter again and completed the sign-in procedure (15 min). Thus, although the time spent with the experimenter was approximately the same, the delayed group met with the experimenter via videoconference twice. Each participant was sent a guide to the intervention via email, with a suggested 8-week protocol.

Follow-up questionnaires, completed via REDCap (Research Electronic Data Capture; Vanderbilt University), were automatically distributed via email or text every 4 weeks. Text and email reminders for the follow-up questionnaires were sent daily for up to 4 days until completed, starting 2 days before each assessment was due. To support engagement, additional text messages were sent to consenting participants (56/72, 78%) twice a week for the first 4 weeks following the start of iDBT in both groups (following this point, reminders were discontinued). These text messages contained a link to a short REDCap survey that encouraged use, queried whether participants wanted a follow-up call, and reported any technical issues. Participants could request additional calls or meetings with the experimenter (via REDCap survey or email) to troubleshoot or clarify different components of the website; however, <10 of these calls or meetings took place throughout the study. Participants were compensated up to CAD $70 (US $45.5) for the completion of these procedures (CAD $10 [US $6.5] for baseline and CAD $20 [US $13] each for the 4-, 8-, and 12-week assessments). On average, participants were compensated CAD $59 (US $38.35), including those who did not collect their final payment.

Intervention

Pocket Skills 2.0 is an iDBT intervention developed by author CRW in collaboration with Microsoft Research and Dr Marsha Linehan; it is built upon the most recent DBT manual available [18]. It uses a web-based portal built on the Microsoft Azure platform that is compatible with any internet browser in addition to the Android and iOS mobile operating systems. This iDBT intervention incorporates lessons following the core modules of DBT as well as a specific module focused on addiction (Table 1 provides more details, and Figure 1 provides the screenshots). Within each module, participants selected a specific skill and were presented with a brief video featuring Dr Linehan introducing the skill and its uses. A practice session then ensues with the rule-based chatbot, which allows for feedback through both open-ended text input and a closed selection of responses. The chatbot guides users on how to select skills to use in different situations that may arise as well as the ability to gain points and unlock additional content, which increases user engagement. After logging in for the first time, participants were prompted to complete an introductory module in which they entered a nickname and set personal goals. Following the completion of this module, participants were able to enter any of the 5 DBT modules offered freely, without the need to unlock any content. The procedure in which iDBT was delivered in this study differs from that in previous studies (refer to Multimedia Appendix 2 [30,34] for a comparison).

Table 1. List of skills covered within the Pocket Skills 2.0 internet-delivered dialectical behavior therapy (DBT) intervention.

Module	DBT skills	Brief training description
Mindfulness	Introduction to mindfulness; wise mind; observing, describing, and participating; and nonjudgment, one-mindfully, and effectively	Introduces the foundational skills to develop nonjudgmental awareness of the present and practice mindfulness with skillful effectiveness.
Emotion regulation	Introduction to emotion regulation; understanding emotions; check the facts, opposite action, and problem-solving; accumulating positives and pleasant events; and building mastery and coping ahead	Teaches the functions of emotions, how to describe them, and skills to reduce the frequency and quantity of unwanted emotions. Also teaches skills to build resilience against future negative emotions.
Distress tolerance	Introduction to distress tolerance; TIP[a], distraction (ACCEPTS[b]), and self-soothe; pros and cons; and Help Me Cope!	Teaches skills to weather crises and intense negative emotions, manage experiential changes, and produce emotional and cognitive change. Help Me Cope! helps the user pick a coping strategy based on a few contextual questions.
Interpersonal effectiveness	Introduction to interpersonal effectiveness; DEARMAN[c], GIVE[d], and FAST[e]; and Dime Game	Teaches skills to navigate interpersonal situations and needs more effectively. Dime Game helps the user evaluate a situation for how firmly to make a request or say no.
Addiction	Introduction to addiction; pros and cons (addiction context); dialectical abstinence and clear mind; and community reinforcement and burning bridges	Helps learners find a middle path between sobriety and unrestrained substance use. Helps learners develop a clear mind and other strategies to stop or reduce problematic substance use.

[a]TIP: temperature, intense exercise, and paced breathing.

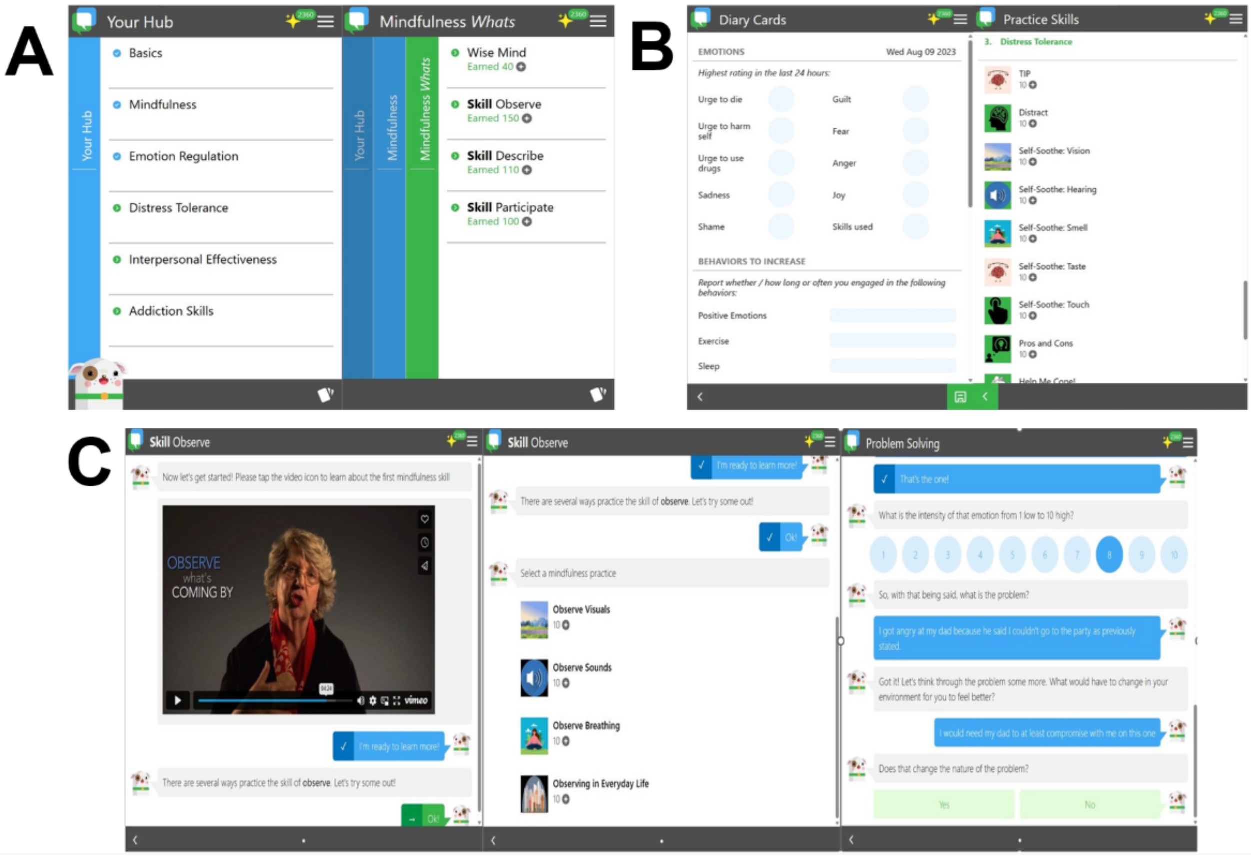
[b]ACCEPTS: activities, contributing, comparisons, emotions, pushing away, thoughts, and sensations.

[c]DEARMAN: describe, express, assert, reinforce, be mindful, appear confident, and negotiate.

[d]GIVE: be gentle, act interested, validate, and use an easy manner.

[e]FAST: be fair, no apologies, stick to values, and be truthful.

Figure 1. Screenshots depicting different features of the Pocket Skills 2.0 internet-delivered dialectical behavior therapy intervention: (A) displays the main Your Hub page, with the next screen showing the submenu selection within the Mindfulness module; (B) shows the optional Diary Card page to input various skills training targets, with the next screen showing the Practice skills page with quicker access to skills training without lessons; and (C) shows the initial portion of the Mindfulness Observe skill lesson, with an embedded video featuring Dr Linehan. The second screen shows the chatbot initializing an interactive skills training exercise, and the third screen shows the types of open- and closed-ended response options along with an example Likert-type rating scale.



Measures

Diagnostic Interviews

The Diagnostic Assessment and Research Tool version 4.0 [39] was used to assess depressive, anxiety, bipolar, obsessive-compulsive, trauma and stressor, alcohol, and SUDs according to the *Diagnostic and Statistical Manual of Mental Disorders, Fifth edition* [40]. We also screened the presence of psychotic disorders. All interviews were completed by the first author, who is a licensed clinical psychologist.

Feasibility and Credibility Measures

For feasibility, we calculated the proportion of randomized participants who started the intervention (by signing in on the first day of access). Of those who started the intervention, we calculated the proportion that (1) dropped out of the intervention after starting (indicated by not logging in after the first day), (2) recorded at least 1 activity after 4 weeks, and (3) completed a recommended dose of using the intervention twice per week for the first month (8 d total). Next, we administered the 6-item Credibility and Expectancy Scale [41] at baseline to assess whether participants had favorable opinions of the intervention and its potential effectiveness before starting treatment. In line with previous work [42], the first 3 items were used to evaluate credibility (using a 9-point Likert scale), whereas a single item (item 4) was used to evaluate expectancy of clinical improvement (using an 11-point Likert scale ranging from 0% to 100%).

Acceptability and Engagement Measures

We used the 6-item Treatment Acceptability Questionnaire [43], which was administered at weeks 4, 8, and 12 to assess ratings of acceptability, perceived effectiveness, and trustworthiness using a 7-point Likert scale. In this analysis, we used only the week 4 and 12 scores. From the intervention source, we examined several metrics tied to engagement or use: the total amount of time spent on the website, the number of interactions with the website (eg, clicks, page views, and text inputs), unique days of log-in, and days of use spread. We then recalculated these metrics for the first 4 weeks, consistent with the acute period of the intervention.

Primary Outcome

The Substance Dependence Scale [44] is a 5-item self-report scale used to assess the severity of alcohol or substance dependence at the baseline and follow-up assessments. Higher scores indicated a higher level of substance dependence. Participants were first asked to indicate which class of substance (including alcohol) they were experiencing the most difficulties abstaining from, even if they reported no use in the past month. The ω reliability coefficient in this study was 0.95. All primary and secondary outcome measures were administered at each assessment point.

Secondary Outcomes

The Patient Health Questionnaire, Depression subscale [45], is a 9-item self-report measure used to assess depressive symptoms over the past 2 weeks, with excellent internal reliability and clinical utility in predicting depression. The ω reliability coefficient was 0.93.

The Generalized Anxiety Disorder-7 Scale [46] is a 7-item self-report measure used to assess generalized anxiety symptoms over the past 2 weeks, with excellent internal reliability and clinical utility in predicting generalized anxiety disorder. The ω reliability coefficient was 0.95.

The Suicidal Behaviors Questionnaire-Revised [47] is a 4-item measure of suicidal thoughts and attempts as well as future intent over the past month, with evidence for its reliability and clinical utility. The total score ranges from 3 to 18, with scores ≥8 indicating significant suicidal risk within clinical samples. The ω reliability coefficient was 0.87.

The World Health Organization Disability Assessment Schedule 2.0 [48] is a 12-item self-report measure assessing functional disability over the past month in several domains (cognition, mobility, self-care, and getting along with others). Higher scores indicate greater functional disability. The ω reliability coefficient was 0.94.

The Difficulties in Emotion Regulation Scale, Short Form [49], is a 16-item self-report measure with excellent internal consistency, assessing emotion dysregulation based on a 6-facet model first described by Gratz and Roemer [50]. Higher scores suggest greater emotion dysregulation difficulties. The ω reliability coefficient was 0.83.

The Mindful Attention Awareness Scale [51] is a 15-item self-report measure of dispositional mindfulness in the form of open or receptive awareness and attention to what is taking place in the present over the past month. Higher scores, which were summed and then averaged, reflected higher levels of dispositional mindfulness. Owing to an administrative error, the anchors were reversed when presented to participants for the entire duration of the study. Therefore, we reversed all scores to ensure a standard interpretation as above. The ω reliability coefficient was 0.94.

The DBT Ways of Coping Checklist [52] is a 59-item self-report measure that assesses the frequency of maladaptive and adaptive skills used to manage difficult situations over the past month, with good internal consistency and test-retest reliability. In this study, we only used the 38-item adaptive skills subscale, which includes skillful behaviors often learned in DBT without using DBT-specific language. The ω reliability coefficient was 0.80.

The National Institutes of Drug Abuse–modified Alcohol, Smoking, and Substance Involvement Screening Test is an adaptation of the original measure [53] used to assess alcohol, smoking, and substance use involvement. This measure was used to assess tobacco, cannabis, cocaine, amphetamine-type stimulants, inhalants, sedatives or sleeping pills, hallucinogens, and opioids. Each class of substance was rated for frequency over the past month using an ordinal scale: 0=never; 1=once or twice; 2=3 or 4 times; 3=5, 6, or 7 times; 4=2 or 3 times a week; 5=4 or 5 times a week; and 6=daily or almost daily. The ω reliability coefficient was 0.23, likely because of the heterogeneity and range of substances used in our sample.

The Daily Drinking Questionnaire [54] was used to assess the frequency of alcohol use on each day of a typical week. Participants were asked how many standard drinks they had consumed on a typical Monday in the past month, with separate questions for each day of the week. Responses were recoded into an ordinal scale: 0=none, 1=1 to 2 standard drinks, 2=3 to 4 standard drinks, 3=5 to 7 standard drinks, 4=8 to 10 standard drinks, 5=11 to 14 standard drinks; and 6=≥15 standard drinks. The ω reliability coefficient was 0.96.

The Risky, Impulsive, and Self-Destructive Questionnaire [55] is an inventory of 38 risky, impulsive, and self-destructive behaviors that sometimes cause problems for people. For brevity and to avoid overlap with other measures, we only used the 4-item *risky sexual behavior* subscale and the 4-item *reckless behavior* subscale. We recoded the frequency of responses, which were evaluated over the past month, into an ordinal scale: 1=none, 2=once or twice, 3=3 to 4 times, 4=5 to 6 times, 5=7 to 9 times, and 6=≥10 times. The ω reliability coefficient was 0.86.

Statistical Analysis

Overview

No outcome measure data were missing from the baseline, and participants returned at least partially completed follow-up questionnaires at rates of 94% (68/72; week 4), 78% (56/72; week 8), and 81% (58/72; week 12). At follow-up, scores for outcome measures were only used if there were <10% of items missing, and we treated outcome measures with no data as missing. The frequency of nonalcoholic substance use, standard alcoholic drinks per day, and risky impulsive behaviors was first recoded using ordinal values to approximately equate each scale with respect to their frequency of occurrence. For each measure, we took the average of each ordinal item score and then rounded the average value to the nearest one to serve as the dependent variable. This rounding was required as an ordinal regression relies on categorizing each value of the ordinal dependent variable as a factor variable. There are several ways to analyze ordinal variables, and this procedure was supported by our biostatistical consultation team.

Descriptive statistics were used to evaluate treatment feasibility, acceptability, and engagement data. Chi-square test, Fisher exact test, and 2-tailed *t* test analyses were used to evaluate baseline differences. Engagement data consisted of time stamped logs of each interaction (ie, clicks, page views, and text inputs) with the website, organized hierarchically within persons, with a total of 39,884 observations. To capture the time spent on iDBT, we ordered the data in Excel (Microsoft Corporation) according to time within persons and calculated a difference score (delta time) between rows. This difference score assessed the time between one meaningful interaction and the next. We then applied a filter to remove any difference scores >30 minutes to account for participants taking breaks or not returning to the app until the next day, capturing 93% of the data. A 10-minute filter captured 92% of the data; however, we wanted to account for playing video content, which could run up to 10 minutes, and the potential of practicing skills live, while remaining on 1 of the web pages. Once the filter was applied, we also calculated the number (and spread) of dates the app was used as well as the number of observations per person, which we called the meaningful interactions calculation. Sensitivity analyses were then performed by examining the same metrics over the first 4 weeks and the time spent on each iDBT module.

All other statistics were run in the statistical program R (version 4.2.1; R Foundation for Statistical Computing). To evaluate the internal consistency of our measures over time, we calculated the between-person ω reliability coefficient [56] statistic using the *omegaSEM* function from the *MultilevelTools* package (version 0.1.1). To characterize changes over time for our continuous variables, we ran a series of linear mixed models with the *lme4* package (version 1.1-26 [57]), with each primary and secondary outcome serving as a dependent variable in separate models. To characterize changes over time for our ordinal variables, we ran additional linear mixed cumulative link models using the *ordinal* package (version 2022.11-16) with separate models for each outcome. As per recommendations, we adjusted each model by incorporating the baseline dependent variable value for each person irrespective of whether the difference was significant between groups [58].

All models included a random intercept for a person and relied on restricted maximum likelihood estimation. We omitted any random slope effects throughout the analyses because all our independent variables were level-2 grouping variables.

Each primary and secondary outcome variable was assessed with models containing an interaction effect (group×time, as factor variables) and main effects only (group+time, as factor variables) along with a continuous covariate controlling for the baseline assessment of each outcome per person. The final model chosen for interpretation was the better fitting model based on lower Akaike information criterion and Bayes information criterion values. Therefore, if the model fit was improved by the inclusion of the interaction term, we report that model; otherwise, we removed the interaction term and report the model with the main effects only. All model comparisons were evaluated using maximum likelihood estimation with the *lmerTest* (version 3.1-3 [59]) package, which uses the Satterthwaite *df* method. To further reduce the number of statistical tests reported, we also opted to interpret only the week 4 and week 12 contrasts against baseline as these were the most pertinent time points to address our hypotheses. Our α significance level was P=.05, and all statistical tests were 2 tailed. The outputs of each final statistical model are provided in Multimedia Appendix 2 for full transparency.

Power

To achieve at least a medium effect size reduction in substance dependence, as suggested by Wilks et al [30], we would require a minimum sample of 60 as per G*Power (version 3.1.9.7; Cohen f=0.15; 2 groups; 4 measurements over 12 wk; power=0.95; α=.05; and correlation between measures of at least 0.70) [60]. With an expected attrition rate of approximately 20%, we aimed to recruit approximately 72 to 75 individuals in total.

Results

Hypothesis 1: Feasibility and Acceptability

Participant Enrollment and Demographic Characteristics

Initially, 116 individuals were assessed for eligibility, and 72 participants aged 18 to 64 years completed all baseline procedures and were randomized. Demographic and clinical characteristics of the sample are presented in Table 2; these characteristics did not differ between groups, suggesting that the randomization procedure was successful. Of the 72 participants, 9 (13%) met the full threshold criteria for >1 SUD. The primary nonalcohol substance disorder across the sample was cannabis (22/72, 31%); nicotine (10/72, 14%); stimulants (7/72, 10%); and sedative, hypnotic, or anxiolytic (1/72, 1%). None of the participants had a current opioid use disorder. Participants met the criteria for a median of 3 psychiatric diagnoses overall (mean 3.30, SD 1.69; range 1-7).

At baseline, 46% (33/72) of the participants reported taking psychotropic medications in the past month, 22% (16/72) of the participants reported seeing a psychiatrist in the past month, and 8% (6/72) of the participants reported attending a community resource (eg, Alcoholics Anonymous and peer support group) in the past month. These rates did not increase when the participants reported the same services at each follow-up. As we did not restrict new options for care following baseline, 4 participants reported having access to outpatient programing at week 4, but only 2 reported this at both weeks 8 and 12. Figure 2 summarizes the study flow of participants in a CONSORT (Consolidated Standards of Reporting Trials) diagram.

Table 2. Demographic characteristics of the total intent-to-treat sample and by condition, with statistical comparisons (N=72).

Characteristics	Total ITT[a] (N=72)	Immediate iDBT[b] (n=38)	Delayed iDBT (n=34)	Group comparison statistical value	Group comparison *P* value
Age (y), mean (SD)	34.1 (11.9)	33.4 (10.5)	34.8 (13.3)	t_{70}=0.50	.62
Sex, n (%)				χ^2_2=0.2	.69
Female	47 (65)	24 (63)	23 (68)	—[c]	—
Male	25 (35)	14 (37)	11 (32)	—	—
Other	0 (0)	0 (0)	0 (0)	—	—
Gender,n (%)[d]				χ^2_2=0.5	.77
Woman	43 (60)	22 (58)	21 (62)	—	—
Man	24 (33)	13 (34)	11 (32)	—	—
Other (nonbinary, transgender, genderfluid, or other)	6 (8)	4 (10)	2 (6)	—	—
Sexual orientation, n (%)[d]				χ^2_3=1.4	.70
Heterosexual	44 (61)	22 (58)	22 (65)	—	—
Lesbian or gay	5 (7)	3 (8)	2 (6)	—	—
Bisexual	11 (15)	5 (13)	6 (18)	—	—
Other (pansexual, queer, asexual, questioning or not sure, or prefer not to answer)	12 (17)	8 (21)	4 (12)	—	—
Race or ethnicity, n (%)[d]				χ^2_5=4.8	.45
Black (African, North American, and Caribbean)	7 (10)	2 (5)	5 (15)	—	—
East or Southeast Asian	5 (7)	4 (10)	1 (3)	—	—
Latin American	5 (7)	3 (8)	2 (6)	—	—
South Asian	9 (12)	5 (13)	4 (12)	—	—
White	48 (67)	25 (66)	23 (68)	—	—
Other (First Nations, Middle Eastern, mixed, or not listed)	8 (11)	6 (16)	2 (6)	—	—
Marital status, n (%)				χ^2_4=4.9	.18
Single	39 (54)	25 (66)	14 (41)	—	—
Dating	18 (25)	8 (21)	10 (29)	—	—
Married	10 (14)	3 (8)	7 (21)	—	—
Other (divorced, widowed, or separated)	5 (7)	2 (5)	3 (9)	—	—
Employment status, n (%)				χ^2_5=1.5	.83
Full-time employed	29 (40)	17 (45)	12 (35)	—	—
Part-time employed	19 (26)	10 (26)	9 (27)	—	—
Unemployed	14 (19)	6 (16)	8 (23)	—	—
On disability	7 (10)	3 (78)	4 (12)	—	—
Prefer not to say	3 (4)	2 (5)	1 (3)	—	—
Current conditions, n (%)					
Major depressive disorder	37 (51)	20 (53)	17 (50)	Fisher exact test	.99
Persistent depressive disorder	18 (25)	10 (26)	8 (23)	Fisher exact test	.99

Characteristics	Total ITT[a] (N=72)	Immediate iDBT[b] (n=38)	Delayed iDBT (n=34)	Group comparison statistical value	Group comparison *P* value
Bipolar I or II disorder	6 (8)	4 (10)	2 (6)	Fisher exact test	.68
Generalized anxiety disorder	37 (51)	21 (55)	16 (47)	Fisher exact test	.64
Social anxiety disorder	22 (31)	10 (26)	12 (35)	Fisher exact test	.45
Posttraumatic stress disorder	18 (25)	10 (26)	8 (23)	Fisher exact test	.99
Other anxiety disorder	11 (15)	8 (21)	3 (9)	Fisher exact test	.20
Alcohol use disorder	47 (65)	23 (60)	24 (71)	Fisher exact test	.46
Any substance use disorder	40 (56)	24 (63)	16 (47)	Fisher exact test	.24
Cannabis use disorder	24 (33)	12 (32)	12 (35)	Fisher exact test	.81
Nicotine use disorder	15 (21)	9 (24)	6 (18)	Fisher exact test	.57
Stimulant use disorder	9 (12)	7 (18)	2 (6)	Fisher exact test	.16
SH or A[e] use disorder	1 (1)	1 (3)	0 (0)	Fisher exact test	.99

[a]ITT: intent-to-treat (ie, completed baseline procedures and randomized to condition).

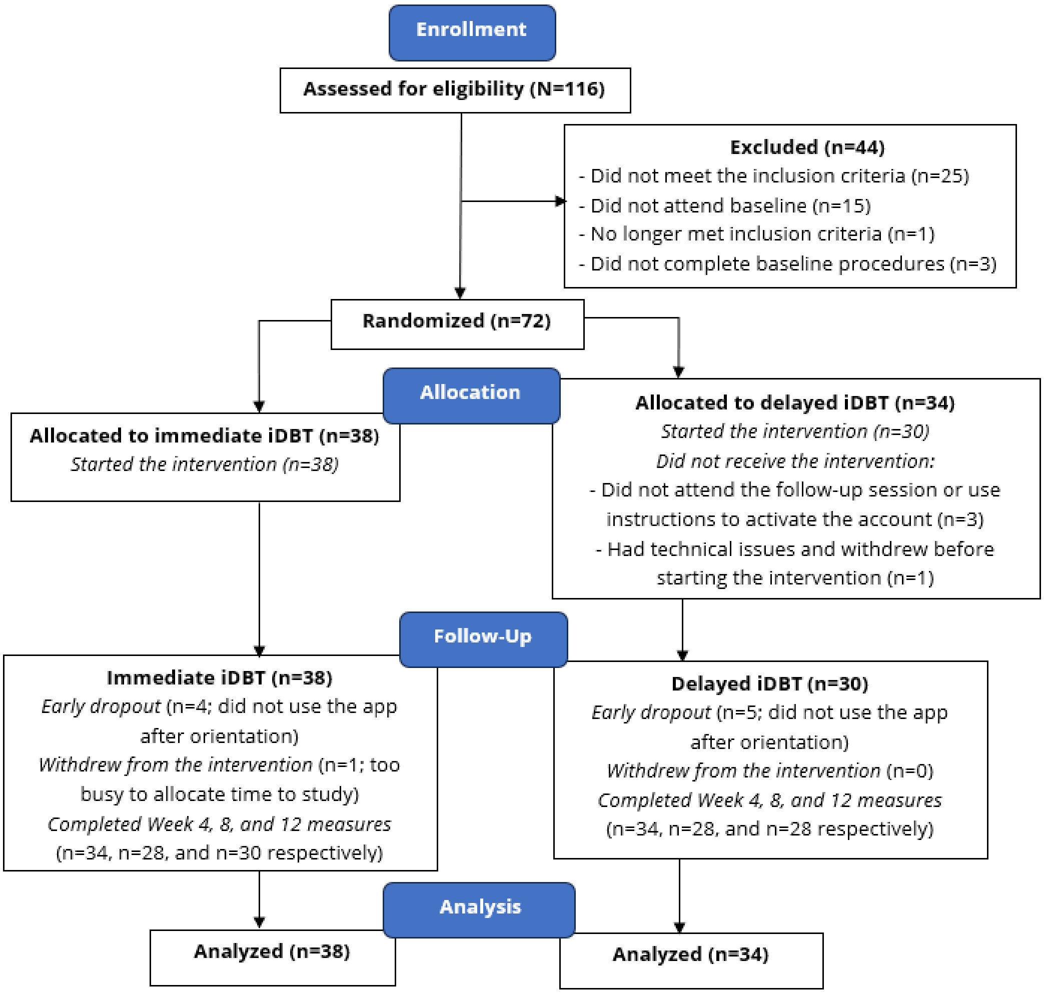
[b]iDBT: internet-delivered dialectical behavioral therapy.

[c]Some cells are empty because we report the group comparison statistic for the overall category above.

[d]Participants could select multiple options.

[e]SH or A: sedative, hypnotic, or anxiolytic.

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) diagram depicting the participant flow through the study. iDBT: internet-delivered dialectical behavior therapy.



Feasibility and Credibility

Moreover, 94% (68/72) of the randomized participants started the intervention. Three participants in the delayed iDBT group did not attend the follow-up session and never connected to the intervention following attempts to reschedule and instructions provided via email. One additional participant withdrew from the delayed group owing to technical issues and being unable to sign in and therefore did not access the intervention. One participant asked to withdraw from the immediate group because of lack of time to spend on the intervention. In addition, 87% (59/68) of those who started iDBT used it for longer than the initial sign-in day (early dropouts: 9/68, 13%), 50% (34/68) recorded at least 1 iDBT activity after 4 weeks, and 35% (24/68) used it according to the recommended dose of 8 days within the first month. Finally, 63% (43/68) of the participants used it for at least 1 hour in total.

Perceptions of intervention credibility at baseline were positive (mean 74%, SD 16%; range 44%-100%). Participants thought it was logical and it was likely to raise their quality of functioning, and they were confident in recommending it to another person. Participants also estimated a mean 58% (SD 21%; range 10%-90%) improvement in symptoms.

Acceptability and Engagement

Treatment acceptability was similar in the immediate iDBT group (mean 36.7, SD 3.8) compared with the delayed iDBT group (mean 35.9, SD 5.3) at week 12; there was no difference between groups (t_{56}=0.66; *P*=.51). A summary of the engagement metrics is provided in Table 3. On average, participants used the app for 2 hours and 24 minutes over the course of 43 days during the study, with 10 unique sign-in days. Participants also recorded an average of 543 meaningful interactions with the website. A breakdown of engagement by module is also provided in Table 3. All metrics tended to be higher on average in the immediate versus delayed iDBT group, consistent with the waitlist control design. Metrics improved further after removing 9 dropout participants who did not use the app after the first day. These individuals appeared to abide closer to the recommendation of using the resource for 8 days within the first month.

Table 3. Pocket Skills engagement metrics across groups.

	Started intervention (n=68), mean (SD; range)	Immediate iDBT[a] (n=38), mean (SD; range)	Delayed iDBT (n=30), mean (SD; range)	Continued intervention after the first day (n=59), mean (SD; range)
Overall engagement				
Total time	2 h 24 min (2 h 45 min; 0 min to 16 h 30 min)	2 h 49 min (3 h 6 min; 0 min to 16 h 30 min)	1 h 51 min (2 h 12 min; 0 min to 8 h 56 min)	2 h 45 min (2 h 47 min; 3 min to 16 h 30 min)
Interactions (clicks, page views, and inputs)	542.78 (569.46; 0-2830)	627.47 (576.22; 0-2468)	429.17 (560.28; 2-2830)	621.85 (571.36; 43-2830)
Unique days of log-in	10.24 (10.81; 1-64)	11.66 (12.68; 1-64)	8.45 (7.80; 1-36)	11.64 (10.94; 2-64)
Days spread	43.41 (32.65; 1-141)	50.95 (37.64; 1-141)	33.69 (22.42; 1-64)	49.88 (30.16; 2-141)
Days in the first 4 wk	6.69 (5.51; 1-29)	7.08 (5.88; 1-29)	6.17 (5.15; 1-21)	7.56 (5.41; 1-29)
Time in the first 4 wk	1 h 54 min (2 h 17 min; 0 min to 13 h 10 min)	2 h 9 min (2 h 30 min; 0 min to 13 h 10 min)	1 h 32 min (2 h 0 min; 1 min to 8 h 30 min)	2 h 10 min (2 h 20 min; 0 min to 13 h 10 min)
Web interactions in the first 4 wk	422.82 (474.87; 0-2566)	463.74 (455.99; 0-2143)	362.83 (507.54; 2-2566)	483.59 (481.66; 0-2566)
Module engagement				
General	38 min 19 s (1 h 21 min 35 s; 0 to 10 h 43 min)	48 min 50 s (1 h 43 min, 5 s; 0 to 10 h 36 min)	25 min 24 s (39 min 11 s; 0 min 52 s to 3 h 26 min)	43 min 38 s (1 h 26 min 25 s; 2 min 25 s to 3 h 26 min)
Mindfulness	46 min 17 s (41 min 43 s; 0 to 2 h 44 min)	53 min 46 s (43 min 35 s; 0 to 2 h 42 min)	36 min 58 s (38 min 29 s; 0 to 2 h 44 min)	53 min 21 s (40 min 21 s; 0 to 2 h 44 min)
Distress tolerance	11 min 0 s (20 min 38 s; 0 to 1 h 34 min)	13 min 06 s (20 min 54 s; 0 to 1 h 13 min)	8 min 34 s (20 min 40 s; 0 to 1 h 34 min)	12 min 41 s (21 min 41 s; 0 to 1 h 34 min)
Emotion regulation	32 min 38 s (46 min 56 s; 0 to 2 h 23 min)	36 min 31 s (51 min 31 s; 0 to 2 h 23 min)	28 min 39 s (41 min, 8 s; 0 to 2 h 10 min)	37 min 37 s (48 min 31 s; 0 to 2 h 23 min)
Interpersonal effectiveness	6 min 35 s (12 min 53 s; 0 to 48 min 39 s)	8 min 05 s (14 min 24 s; 0-48 min 39 s)	3 min 49 s (9 min 41 s; 0 to 34 min 33 s)	7 min 35 s (13 min 34 s; 0 to 48 min 39 s)
Addiction	9 min 18 s (19 min 3 s; 0 to 1 h 19 min)	8 min 52 s (17 min 42 s; 0 to 1 h 12 min)	8 min 19 s (19 min 31 s; 0 to 1 h 19 min)	10 min 43 s (20 min 6 s; 1 h to 19 min)

[a]iDBT: internet-delivered dialectical behavioral therapy.

Hypothesis 2: Were Improvements Greater in the Immediate Versus Delayed iDBT Group?

Overview

The unadjusted means, SDs, and the number of participants for each continuous variable are presented for each group and assessment point in Table 4, with between- and within-group effect size estimates presented in Table 5. The unadjusted values for each ordinal variable are provided in Multimedia Appendix 2.

Table 4. Unadjusted means (and SDs) by group and time point for continuous outcome measures.

Group and time	Continuous outcomes[a]							
	SDS[b]	PHQ-9[c]	GAD-7[d]	SBQ[e]	DERS-16[f]	MAAS[g]	WHODAS[h]	DBT-WCCL[i]
Immediate iDBT[j] (n=38)								
Week 0								
Values, n (%)[k]	38 (100)	38 (100)	38 (100)	38 (100)	38 (100)	38 (100)	38 (100)	38 (100)
Values, mean (SD)	8.7 (3.5)	12.9 (5.9)	11.9 (5.4)	8.6 (3.6)	51.8 (13.4)	3.6 (0.9)	16.5 (9.0)	1.8 (0.4)
Week 4								
Values, n (%)	34 (89)	34 (89)	34 (89)	34 (89)	33 (87)	31 (82)	31 (82)	31 (82)
Values, mean (SD)	6.5 (4.3)	9.0 (4.9)	9.25 (4.5)	7.9 (3.3)	47.6 (13.5)	3.6 (0.8)	14.9 (8.6)	1.8 (0.5)
Week 8								
Values, n (%)	28 (74)	28 (74)	28 (74)	28 (74)	28 (74)	27 (71)	27 (71)	27 (71)
Values, mean (SD)	6.3 (3.9)	9.5 (5.4)	8.7 (5.0)	7.8 (2.7)	44.1 (12.4)	4.0 (0.9)	12.4 (7.3)	1.9 (0.5)
Week 12								
Values, n (%)	30 (79)	30 (79)	30 (79)	30 (79)	30 (79)	29 (76)	29 (76)	29 (76)
Values, mean (SD)	6.2 (4.1)	8.5 (4.6)	8.5 (4.3)	7.6 (3.1)	44.4 (10.8)	4.0 (0.8)	12.2 (6.9)	1.9 (0.5)
Delayed iDBT (n=34)								
Week 0								
Values, n (%)	34 (100)	34 (100)	34 (100)	34 (100)	34 (100)	34 (100)	34 (100)	34 (100)
Values, mean (SD)	6.8 (3.5)	11.1 (6.7)	8.9 (6.1)	6.9 (3.5)	50.5 (12.1)	3.5 (1.0)	15.5 (9.2)	1.8 (0.4)
Week 4								
Values, n (%)	34 (100)	34 (100)	34 (100)	34 (100)	34 (100)	33 (97)	33 (97)	33 (97)
Values, mean (SD)	5.6 (3.9)	10.1 (6.6)	8.6 (6.1)	6.9 (3.4)	49.9 (13.2)	3.7 (1.0)	14.3 (9.0)	1.7 (0.4)
Week 8								
Values, n (%)	28 (82)	28 (82)	28 (82)	28 (82)	28 (82)	28 (82)	28 (82)	28 (82)
Values, mean (SD)	5.4 (3.9)	7.6 (6.1)	7.6 (5.4)	5.9 (3.3)	44.6 (13.5)	3.8 (0.9)	11.8 (9.3)	1.9 (0.5)
Week 12								
Values, n (%)	28 (82)	28 (82)	28 (82)	28 (82)	28 (82)	28 (82)	28 (82)	28 (82)
Values, mean (SD)	5.3 (3.9)	7.3 (5.3)	7.1 (4.6)	6.1 (3.3)	43.1 (12.9)	4.1 (1.1)	11.1 (9.2)	2.0 (0.4)

[a]Descriptive statistics for variables with ordinal values (all secondary outcome variables) are presented in Multimedia Appendix 2.
[b]SDS: Substance Dependence Scale.
[c]PHQ-9: Patient Health Questionnaire-9.
[d]GAD-7: Generalized Anxiety Disorder-7.
[e]SBQ: Suicidal Behaviors Questionnaire.
[f]DERS-16: Difficulties in Emotion Regulation Scale-16 item.
[g]MAAS: Mindful Attention Awareness Scale.
[h]WHODAS: World Health Organization Disability Assessment Schedule.
[i]DBT-WCCL: Dialectical Behavior Therapy Ways of Coping Checklist.
[j]iDBT: internet-delivered dialectical behavioral therapy.
[k]n=number of participants contributing to the calculations.

Table 5. Effect sizes for continuous outcome measures.

Outcome	Within-group effect sizes, Cohen *d*[a]				Between-group effect sizes, Cohen *d*[b]	
	Immediate iDBT[c]		Delayed iDBT		Immediate iDBT vs delayed iDBT	
	Week 4	Week 12	Week 4[d]	Week 12[e]	Week 4 vs baseline	Week 12 vs baseline
SDS[f]	−0.58	−0.95	−0.73	−0.84	−0.32	−0.32
PHQ-9[g]	−0.66	−0.87	−0.42	−1.05	−0.48	−0.04
GAD-7[h]	−1.70	−1.52	−0.27	−1.07	−0.40	−0.21
SBQ[i]	−0.27	−0.56	0.02	−0.45	−0.18	−0.01
DERS-16[j]	−0.58	−0.68	−0.11	−1.23	−0.27	0.01
MAAS[k]	−0.02	3.78	0.78	2.34	−0.22	−0.21
WHODAS[l]	−0.28	−0.76	−0.51	−3.42	−0.04	0.03
DBT-WCCL[m]	0.16	0.25	−0.38	0.48	0.22	−0.22

[a]We calculated Cohen repeated measures *d*, with a pooled SD (refer to the study by Lakens [61], formula 8) with values ≥0.20=small, ≥0.50=medium or moderate, and ≥0.80=large effect [62].

[b]Uses the Klauer method, where effect size, Cohen *d* for both groups was calculated and then subtracted from each other. This allowed for the correction of different sample sizes and baseline values.

[c]iDBT: internet-delivered dialectical behavioral therapy.

[d]This effect size technically captures the repeated baseline effect.

[e]This effect size technically captures 8 weeks following the start of the intervention.

[f]SDS: Substance Dependence Scale.

[g]PHQ-9: Patient Health Questionnaire-9.

[h]GAD-7: Generalized Anxiety Disorder-7.

[i]SBQ: Suicidal Behaviors Questionnaire.

[j]DERS-16: Difficulties in Emotion Regulation Scale-16 item.

[k]MAAS: Mindful Attention Awareness Scale.

[l]WHODAS: World Health Organization Disability Assessment Schedule.

[m]DBT-WCCL: Dialectical Behavior Therapy Ways of Coping Checklist.

Primary Outcome

Contrary to our hypotheses, we did not find any significant group×time interactions for the severity of substance dependence at week 4 or week 12.

Secondary Outcomes

Consistent with the hypotheses, the results supported greater benefits for the immediate versus delayed iDBT group for several secondary outcomes. At week 4, there were significant group×time interactions for depression and anxiety, where the immediate access group reported fewer depressive (b=−2.46; SE 1.05; 95% CI −4.51 to −0.40; *P*=.02) and anxiety symptoms (b=−2.22; SE 0.96; 95% CI −4.09 to −0.34; *P*=.02) compared with the delayed iDBT group. At week 12, there were significant group×time interactions for standard alcoholic drinks per day (b=−2.00; SE 0.83; 95% CI −3.64 to −0.36; *P*=.02) and nonalcoholic substance use (b=−3.74; SE 1.47; 95% CI −6.63 to −0.85; *P*=.01), where the immediate group had lower frequencies of both over the course of the study compared with the delayed group. Contrary to the expectations, there were no significant group×time interactions for all other outcomes: emotion dysregulation, suicidality, DBT skills acquisition, dispositional mindfulness, functional disability, and risky impulsive behaviors.

Hypothesis 3: Did iDBT Produce Improvements Regardless of Group?

Primary Outcome

There were significant main effects of time at week 4 (b=−1.73; SE 0.34; 95% CI −2.40 to −1.07; *P*<.001) and week 12 (b=−2.09; SE 0.36; 95% CI −2.80 to −1.39; *P*<.001), indicating a significant decrease in substance dependence for both groups, with no differences between groups (*P*=.25).

Secondary Outcomes

There were several findings supporting the benefits of the intervention in the follow-up phase of the study (no other significant main effects emerged at week 4). At week 12, there were significant main effects of time for depression (b=−2.95; SE 0.79; 95% CI −4.50 to −1.39; *P*<.001), anxiety (b=−1.57; SE 0.73; 95% CI −2.99 to −0.14; *P*=.03), suicidality (b=−0.70; SE 0.24; 95% CI −1.12 to −0.28; *P*=.001), emotion dysregulation (b=−6.56; SE 1.20; 95% CI −8.90 to −4.21; *P*<.001), functional disability (b=−3.64; SE 0.77; 95% CI −5.15 to −2.14; *P*<.001), dispositional mindfulness (b=0.44; SE 0.09;

95% CI 0.27-0.62; *P*<.001), and DBT skill acquisition (b=0.14; SE=0.05; 95% CI 0.04-0.23; *P*=.005), indicating that both groups saw significant improvements from baseline over the study duration, with no differences between groups (all *P*>.25). There were no main effects of time for risky impulsive behaviors and no difference between groups (*P*>.23).

Discussion

Summary

This study is unique in that it delivered high-quality iDBT in a self-guided format that participants could use through any internet browser on a computer, tablet, or smartphone. Here, we evaluated the feasibility, acceptability, and potential efficacy of iDBT in a sample of treatment-seeking individuals with SUDs often presenting with additional mental health symptoms. In this study, *Pocket Skills* 2.0 garnered some meaningful support as a potential intervention for those with SUDs and other mental health concerns. We also discuss some caveats and limitations in the following sections.

Feasibility and Acceptability

The intervention was deemed credible and potentially helpful by participants. In terms of treatment initiation, we found that 94% of the randomized participants started *Pocket Skills* compared with 98% in a previous remote iDBT intervention study [30]. For reference, 88% of the participants started the intervention in the study by van Spijker et al [28] and only 39% of the participants started self-guided iDBT in the study by Simon et al [31]. In this study, not initiating iDBT was mostly because of participants not attending a follow-up session after 4 weeks of being on the waitlist, and in 1 case, owing to technical issues. Thus, the feasibility of deploying iDBT remains high with few technical compatibility issues. These results support the feasibility of adapting DBT for delivery in internet-delivered formats in this context [25,30,34].

Of those who started iDBT, 13% (9/68) were early dropouts, defined as those who did not attempt the intervention after the first day of use, and 50% (34/68) continued to use the app after 4 weeks. Comparatively, Wilks et al [30] recorded a dropout rate of 19%, although different dropout criteria were used (eg, stopped attempting or completing the intervention for 3 weeks in a row). This study differs in that our participants had unrestricted access to all content, whereas the previous study used a week-to-week module approach; thus, we had different definitions of dropout by virtue of study design. There was an overall dropout rate of approximately 10% in the intervention arm in the study by van Spijker et al [28], whereas <9% went beyond the introduction section in the study by Simon et al [31,32]. The rate can also be compared with internet-based psychological treatments more broadly (31%), in-person delivered DBT for different clinical conditions (28%), and in-person psychological treatments for SUDs (30%) [63-65]. Although definitions of dropout vary across trials, this study provides some promise regarding the potential uptake and adherence to iDBT by individuals with SUDs. An analysis of the predictors of dropout from the 2018 study [30] indicated that technological barriers and low perceptions of usefulness emerged as significant [33]. Although we focused on the current hypotheses, we intend to examine predictors of treatment outcomes (including dropout) in future research.

The time spent on the iDBT intervention varied widely. Only 35% (24/68) of the participants completed the recommended dose of spending 8 days in the first 4 weeks, which improved to 41% (24/59) when early dropouts were omitted. These findings can be contextualized by the limited support and self-guided nature of the intervention. Comparatively, 42% of the participants in the study by Wilks et al [30] completed half of the iDBT content in the same time frame (1 month), which included considerably more support, such as daily reminders, homework assignments, and phone calls regarding suicide risk. Approximately half of the participants in the intervention arm completed ≥3 of 6 sessions in the study by van Spijker et al [28], and a similar proportion finished a 15-session course of DBT delivered by email [27]. While we did offer text message reminders to facilitate encouragement, few participants asked for additional meetings or followed our suggested guide. However, even with limited support, a sizeable proportion of participants used the app over several days within the first month, totaling >1 hour of use. These findings suggest potential benefits of additional meetings or coaching sessions, which may improve adherence and engagement to the intervention and improve clinical outcomes overall.

There were relatively high ratings for the content, suitability, and trustworthiness of *Pocket Skills* using an established measure of treatment acceptability, extending 2 earlier studies [30,34]. Participants recorded most of their engagement within the first 4 weeks of the intervention and in that time, averaged 2 hours of interaction. In a previous 4-week trial of an earlier version of *Pocket Skills*, which was conducted in patients concurrently completing in-person DBT, participants used the app for 14 out of 28 days and spent 2.25 hours on the app during that time [34]. This equated to approximately 4 minutes of activity per person per day. Given that our study was largely self-guided without additional support or check-ins, this newer version of *Pocket Skills* saw comparative engagement with respect to total time. One self-guided skills training intervention reported 10.5 hours of use on average or approximately 15 minutes per day over 6 weeks [28]. Examining iDBT as an adjunct to in-person (or videoconference) DBT to improve engagement even further may be warranted.

Potential Efficacy

With regard to our second hypothesis evaluating the waitlist control design of the study, we found little evidence that the immediate iDBT group benefited more in the acute period of 4 weeks. We did not find a difference between groups in our primary outcome at this time point, where we expected it. We found interactions at week 4 for 2 secondary outcomes (ie, depression and anxiety) in favor of the hypothesis. We found additional interactions during the follow-up period (week 12) for decreased standard alcoholic drinks per day and overall substance use frequency in favor of the immediate iDBT group. Notably, we originally planned a 16-week trial with an equal immediate and waitlist period of 8 weeks. However, in a pilot study, we found attrition and lack of engagement to be greater than in this study, which contributed to the revised study design.

Most studies on app- and internet-based interventions use one month as an acute test of the intervention with another month as a follow-up assessment period owing to increasing attrition after 7 to 8 weeks [35-37].

Our third hypothesis regarding overall improvement was more consistently supported, with favorable improvements in our primary and secondary outcome measures in both groups by week 12. There were medium to large effect size improvements in our primary outcome of substance dependence, both in the acute (week 4) and follow-up phase, consistent with the literature supporting in-person DBT for alcohol use disorders and SUDs [16,21-24]. There were many medium to large effect size improvements to our secondary outcomes by week 12, including depression, anxiety, suicidal behavior, emotion dysregulation, and functional disability and these effects did not differ significantly by group once accounting for baseline differences. These findings are consistent with in-person DBT improvements in suicidal behavior, depression, anxiety, and emotion dysregulation [23,66-69] and extend a previous stand-alone iDBT study [30]. The intervention also improved DBT skill acquisition and dispositional mindfulness in both groups by week 12, as seen by small-to-large positive effect size values and in line with the literature findings on face-to-face DBT [52,70,71]. Whether DBT skill acquisition mediated treatment outcomes in this study, as implied by previous research [68,69], is a hypothesis that could be examined in a follow-up analysis.

Limitations

Our ability to detect interactions during the acute phase of treatment was limited given that differences would have had to be medium or large to detect using our unbalanced waitlist design. The use of a waitlist control may have led some individuals to consider other treatment options or drop out of the study without connecting to the intervention, especially given its unblinded nature following randomization. Alternatively, the waitlist condition may have supported a nocebo effect as participants reported favorable changes in most outcomes despite a lack of treatment [72]. Although the randomization procedure was conducted in a blind manner, following the baseline session, all other follow-ups and contact with participants were unblinded, which may have introduced potential experimenter bias. The primary and most secondary outcomes were participant rated (ie, self-report), which is also less robust to bias than a blinded outcomes assessor.

In terms of feasibility metrics, although we saw that roughly half of the participants remained active on iDBT after 4 weeks, it is not clear how consistently active they were. Because we did not restrict access to other interventions, it remains unclear how specific the intervention effects were tied to iDBT in this study. We attempted to mitigate concerns about this by asking questions about different services that participants were receiving at each follow-up and found little to no increase in psychotropic medications, community support groups, and additional treatments. More analyses are needed to understand the dose-response relationship between the intervention and treatment outcome, which will be addressed in future work.

Pocket Skills 2.0 omits several aspects of in-person formats of DBT that may improve engagement and adherence, such as a significant group therapy component; more robust tracking of thoughts, emotions, and behaviors using diary cards; and handouts and worksheets for homework. These implementation differences compared with standard in-person DBT may have influenced treatment outcomes. As a technical limitation, most participants accessed iDBT on a home computer or laptop. Although we discussed the ability to sign in on mobile devices with participants, we typically asked them to sign in for the first time on a computer or laptop based on experiences during the piloting phase of this study (ie, there was more difficulty signing in on the mobile iOS platform). This procedural issue may have introduced a barrier to using iDBT on mobile devices; however, we could have spent more time ensuring that the intervention was working on both participants' smartphones and computer devices. Future implementations could include subsequent meetings with participants (eg, check-ins) to address any technological and compatibility issues more quickly.

Despite efforts to recruit a diverse sample, our sample was predominantly White, female, heterosexual, and largely aged between 18 and 45 years. Future research should attempt to replicate the outcomes of iDBT across more diverse samples, such as sexual, gender, and ethnoracial minority groups (refer to the study by Harned et al [73] for review). Our sample was heterogeneous with respect to their endorsement of alcohol or nonalcoholic substance difficulties. Measuring the frequency and severity of multiple substances in an efficient way is challenging. Owing to the use of different measures and rating scales to assess alcohol and nonalcoholic substances (as well as risky and impulsive behaviors), our raw data required rescaling to create new ordinal scales that approximately modeled the same frequency and severity across multiple scales. Future research could adhere to more standardized approaches that allow for responses to be collected and then coded more reliability. For example, the Timeline Follow Back interview has been used to self-report alcohol and substance use as well as risky sexual behaviors [74,75].

Conclusions

Notwithstanding the limitations of this study, our iDBT intervention *Pocket Skills* 2.0 was supported as a feasible and acceptable intervention for those with SUDs and other mental health concerns. However, methods to improve engagement should be further evaluated. The intervention not only showed potential effectiveness for substance dependence but also demonstrated positive effects across various mental health symptoms, affirming its clinical utility. These findings add to the sparse literature on internet-based DBT and internet-delivered psychological interventions for SUDs. This format has the potential to increase accessibility and reduce the costs and resources required for in-person DBT. Several research priorities were identified to potentially improve engagement and optimize treatment outcomes as well as understand how our iDBT intervention can be integrated into the larger landscape of treatment options for SUDs and other conditions.

Acknowledgments

This study was supported by a Canadian Institutes for Health Research Fellowship Award (#458772) awarded to ARD. ARD would like to thank Daniel J Tse for help in organizing and creating formulas to summarize the website analytic data and Megan Brndjar for help in organizing and tabulating the outcome data. The authors sincerely thank Geraldine "Geri" Rodriguez and Psychwire for permitting us to use their training videos featuring Dr Linehan.

Data Availability

Deidentified data may be requested from ARD.

Conflicts of Interest

CRW receives consultation fees from Mindstrong Health, Click Therapeutics, and Behavioral Tech Research. CRW is one of the developers of Pocket Skills. EHP served, in addition to other duties, as a technical support during the study and piloted the intervention. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 8678 KB - mental_v11i1e50399_app1.pdf]

Multimedia Appendix 2
Supplementary results and tables.
[DOCX File , 45 KB - mental_v11i1e50399_app2.docx]

References

1. Degenhardt L, Whiteford HA, Ferrari AJ, Baxter AJ, Charlson FJ, Hall WD, et al. Global burden of disease attributable to illicit drug use and dependence: findings from the Global Burden of Disease Study 2010. Lancet 2013 Nov 09;382(9904):1564-1574. [doi: 10.1016/S0140-6736(13)61530-5] [Medline: 23993281]
2. Rehm J. The risks associated with alcohol use and alcoholism. Alcohol Res Health 2011;34(2):135-143 [FREE Full text] [Medline: 22330211]
3. Schulte MT, Hser YI. Substance use and associated health conditions throughout the lifespan. Public Health Rev 2014;35(2) [FREE Full text] [doi: 10.1007/BF03391702] [Medline: 28366975]
4. Rush B, Koegl CJ. Prevalence and profile of people with co-occurring mental and substance use disorders within a comprehensive mental health system. Can J Psychiatry 2008 Dec;53(12):810-821. [doi: 10.1177/070674370805301207] [Medline: 19087479]
5. Alcohol and drug use in Canada, 2019. Statistics Canada. 2021 Dec 20. URL: https://www150.statcan.gc.ca/n1/daily-quotidien/211220/dq211220c-eng.htm [accessed 2023-12-12]
6. Roberts A, Rogers J, Mason R, Siriwardena AN, Hogue T, Whitley GA, et al. Alcohol and other substance use during the COVID-19 pandemic: a systematic review. Drug Alcohol Depend 2021 Dec 01;229(Pt A):109150 [FREE Full text] [doi: 10.1016/j.drugalcdep.2021.109150] [Medline: 34749198]
7. Satre DD, Meacham MC, Asarnow LD, Fisher WS, Fortuna LR, Iturralde E. Opportunities to integrate mobile app-based interventions into mental health and substance use disorder treatment services in the wake of COVID-19. Am J Health Promot 2021 Nov;35(8):1178-1183 [FREE Full text] [doi: 10.1177/08901171211055314] [Medline: 34652971]
8. Carreiro S, Newcomb M, Leach R, Ostrowski S, Boudreaux ED, Amante D. Current reporting of usability and impact of mHealth interventions for substance use disorder: a systematic review. Drug Alcohol Depend 2020 Oct 01;215:108201 [FREE Full text] [doi: 10.1016/j.drugalcdep.2020.108201] [Medline: 32777691]
9. Klein JW. Pharmacotherapy for substance use disorders. Med Clin North Am 2016 Jul;100(4):891-910. [doi: 10.1016/j.mcna.2016.03.011] [Medline: 27235620]
10. Magill M, Ray L, Kiluk B, Hoadley A, Bernstein M, Tonigan JS, et al. A meta-analysis of cognitive-behavioral therapy for alcohol or other drug use disorders: treatment efficacy by contrast condition. J Consult Clin Psychol 2019 Dec;87(12):1093-1105 [FREE Full text] [doi: 10.1037/ccp0000447] [Medline: 31599606]
11. Dellazizzo L, Potvin S, Giguère S, Landry C, Léveillé N, Dumais A. Meta-review on the efficacy of psychological therapies for the treatment of substance use disorders. Psychiatry Res 2023 Aug;326:115318. [doi: 10.1016/j.psychres.2023.115318] [Medline: 37356250]
12. Lieb R. Epidemiological perspectives on comorbidity between substance use disorders and other mental disorders. In: Dom G, Moggi F, editors. Co-occurring Addictive and Psychiatric Disorders: A Practice-Based Handbook from a European Perspective. Berlin, Germany: Springer; 2015:3-12.

13. Stellern J, Xiao KB, Grennell E, Sanches M, Gowin JL, Sloan ME. Emotion regulation in substance use disorders: a systematic review and meta-analysis. Addiction 2023 Jan;118(1):30-47 [FREE Full text] [doi: 10.1111/add.16001] [Medline: 35851975]
14. Fernandez KC, Jazaieri H, Gross JJ. Emotion regulation: a transdiagnostic perspective on a new RDoC domain. Cognit Ther Res 2016 Jun;40(3):426-440 [FREE Full text] [doi: 10.1007/s10608-016-9772-2] [Medline: 27524846]
15. Cludius B, Mennin D, Ehring T. Emotion regulation as a transdiagnostic process. Emotion 2020 Feb;20(1):37-42. [doi: 10.1037/emo0000646] [Medline: 31961175]
16. Cavicchioli M, Movalli M, Ramella P, Vassena G, Prudenziati F, Maffei C. Feasibility of dialectical behavior therapy skills training as an outpatient program in treating alcohol use disorder: the role of difficulties with emotion regulation and experiential avoidance. Addic Res Theory 2019 Apr 11;28(2):103-115. [doi: 10.1080/16066359.2019.1590558]
17. Moos RH. Theory-based active ingredients of effective treatments for substance use disorders. Drug Alcohol Depend 2007 May 11;88(2-3):109-121 [FREE Full text] [doi: 10.1016/j.drugalcdep.2006.10.010] [Medline: 17129682]
18. Linehan MM. Skills Training Manual for Treating Borderline Personality Disorder: Diagnosis and Treatment of Mental Disorders. New York, NY: The Guilford Press; 1993.
19. Linehan MM. DBT Skills Training Manual. 2nd edition. New York, NY: The Guilford Press; 2014.
20. Neacsiu AD, Bohus M, Linehan MM. Dialectical behavior therapy: an intervention for emotion dysregulation. In: Gross JJ, editor. Handbook of Emotion Regulation. New York, NY: The Guilford Press; 2014:491-507.
21. Dimeff LA, Linehan MM. Dialectical behavior therapy for substance abusers. Addict Sci Clin Pract 2008 Jun;4(2):39-47 [FREE Full text] [doi: 10.1151/ascp084239] [Medline: 18497717]
22. Lee NK, Cameron J, Jenner L. A systematic review of interventions for co-occurring substance use and borderline personality disorders. Drug Alcohol Rev 2015 Nov;34(6):663-672. [doi: 10.1111/dar.12267] [Medline: 25919396]
23. Valentine SE, Smith AM, Stewart K. A review of the empirical evidence for DBT skills training as a stand-alone intervention. In: Bedics J, editor. The Handbook of Dialectical Behavior Therapy: Theory, Research, and Evaluation. Cambridge, MA: Academic Press; 2020:325-358.
24. Warner N, Murphy M. Dialectical behaviour therapy skills training for individuals with substance use disorder: a systematic review. Drug Alcohol Rev 2022 Feb;41(2):501-516. [doi: 10.1111/dar.13362] [Medline: 34337811]
25. Lakeman R, King P, Hurley J, Tranter R, Leggett A, Campbell K, et al. Towards online delivery of dialectical behaviour therapy: a scoping review. Int J Ment Health Nurs 2022 Aug;31(4):843-856 [FREE Full text] [doi: 10.1111/inm.12976] [Medline: 35048482]
26. Lopez A, Rothberg B, Reaser E, Schwenk S, Griffin R. Therapeutic groups via video teleconferencing and the impact on group cohesion. Mhealth 2020 Aug;6:13 [FREE Full text] [doi: 10.21037/mhealth.2019.11.04] [Medline: 32270005]
27. Alavi N, Stephenson C, Rivera M. Effectiveness of delivering dialectical behavioral therapy techniques by email in patients with borderline personality disorder: nonrandomized controlled trial. JMIR Ment Health 2021 Apr 30;8(4):e27308 [FREE Full text] [doi: 10.2196/27308] [Medline: 33835936]
28. van Spijker BA, van Straten A, Kerkhof AJ. Effectiveness of online self-help for suicidal thoughts: results of a randomised controlled trial. PLoS One 2014;9(2):e90118 [FREE Full text] [doi: 10.1371/journal.pone.0090118] [Medline: 24587233]
29. Whiteside U, Richards J, Huh D, Hidalgo R, Nordhauser R, Wong AJ, et al. Development and evaluation of a web-based resource for suicidal thoughts: NowMattersNow.org. J Med Internet Res 2019 May 02;21(5):e13183 [FREE Full text] [doi: 10.2196/13183] [Medline: 31045498]
30. Wilks CR, Lungu A, Ang SY, Matsumiya B, Yin Q, Linehan MM. A randomized controlled trial of an internet delivered dialectical behavior therapy skills training for suicidal and heavy episodic drinkers. J Affect Disord 2018 May;232:219-228 [FREE Full text] [doi: 10.1016/j.jad.2018.02.053] [Medline: 29499504]
31. Simon GE, Shortreed SM, Rossom RC, Beck A, Clarke GN, Whiteside U, et al. Effect of offering care management or online dialectical behavior therapy skills training vs usual care on self-harm among adult outpatients with suicidal ideation: a randomized clinical trial. JAMA 2022 Feb 15;327(7):630-638 [FREE Full text] [doi: 10.1001/jama.2022.0423] [Medline: 35166800]
32. Stanley B, Dixon L. Health system-based low-intensity interventions to prevent self-harm among patients with suicidal ideation: disentangling the effects of implementation strategies and interventions. JAMA 2022 Feb 15;327(7):626-628. [doi: 10.1001/jama.2022.0770] [Medline: 35166818]
33. Wilks CR, Yin Q, Zuromski KL. User experience affects dropout from internet-delivered dialectical behavior therapy. Telemed J E Health 2020 Jun;26(6):794-797 [FREE Full text] [doi: 10.1089/tmj.2019.0124] [Medline: 31502945]
34. Schroeder J, Wilks C, Rowan K, Toledo A, Paradiso A, Czerwinski MP, et al. Pocket skills: a conversational mobile web app to support dialectical behavioral therapy. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018 Presented at: CHI '18; April 21-26, 2018; Montreal, QC p. 1-15 URL: https://dl.acm.org/doi/10.1145/3173574.3173972 [doi: 10.1145/3173574.3173972]
35. Miner A, Kuhn E, Hoffman JE, Owen JE, Ruzek JI, Taylor CB. Feasibility, acceptability, and potential efficacy of the PTSD coach app: a pilot randomized controlled trial with community trauma survivors. Psychol Trauma 2016 May;8(3):384-392. [doi: 10.1037/tra0000092] [Medline: 27046668]

36. Mohr DC, Schueller SM, Tomasino KN, Kaiser SM, Alam N, Karr C, et al. Comparison of the effects of coaching and receipt of app recommendations on depression, anxiety, and engagement in the IntelliCare platform: factorial randomized controlled trial. J Med Internet Res 2019 Aug 28;21(8):e13609 [FREE Full text] [doi: 10.2196/13609] [Medline: 31464192]
37. Saad A, Bruno D, Camara B, D'Agostino J, Bolea-Alamanac B. Self-directed technology-based therapeutic methods for adult patients receiving mental health services: systematic review. JMIR Ment Health 2021 Nov 26;8(11):e27404 [FREE Full text] [doi: 10.2196/27404] [Medline: 34842556]
38. Biener L, Abrams DB. The Contemplation Ladder: validation of a measure of readiness to consider smoking cessation. Health Psychol 1991;10(5):360-365. [doi: 10.1037//0278-6133.10.5.360] [Medline: 1935872]
39. McCabe RE, Milosevic I, Rowa K, Shnaider P, Pawluk EJ, Antony MM. Diagnostic assessment research tool (DART). APA PsycTests. 2017. URL: https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft81500-000 [accessed 2023-12-12]
40. American Psychiatric Association. Diagnostic And Statistical Manual Of Mental Disorders. 5th Edition. New York, NY: American Psychiatric Association; 2013.
41. Devilly GJ, Borkovec TD. Psychometric properties of the credibility/expectancy questionnaire. J Behav Ther Exp Psychiatry 2000 Jun;31(2):73-86. [doi: 10.1016/s0005-7916(00)00012-4] [Medline: 11132119]
42. Thompson-Hollands J, Bentley KH, Gallagher MW, Boswell JF, Barlow DH. Credibility and outcome expectancy in the unified protocol: relationship to outcomes. J Exp Psychopathol 2014 Mar 30;5(1):72-82 [FREE Full text] [doi: 10.5127/jep.033712]
43. Hunsley J. Development of the treatment acceptability questionnaire. J Psychopathol Behav Assess 1992 Mar;14(1):55-64 [FREE Full text] [doi: 10.1007/bf00960091]
44. Gossop M, Darke S, Griffiths P, Hando J, Powis B, Hall W, et al. The Severity of Dependence Scale (SDS): psychometric properties of the SDS in English and Australian samples of heroin, cocaine and amphetamine users. Addiction 1995 May;90(5):607-614. [doi: 10.1046/j.1360-0443.1995.9056072.x] [Medline: 7795497]
45. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
46. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
47. Osman A, Bagge CL, Gutierrez PM, Konick LC, Kopper BA, Barrios FX. The Suicidal Behaviors Questionnaire-Revised (SBQ-R): validation with clinical and nonclinical samples. Assessment 2001 Dec;8(4):443-454. [doi: 10.1177/107319110100800409] [Medline: 11785588]
48. Üstün TB, Kostanjsek N, Chatterji S, Rehm J, World Health Organization. Measuring health and disability: manual for WHO Disability Assessment Schedule (WHODAS 2.0) . World Health Organization. 2010. URL: https://iris.who.int/bitstream/handle/10665/43974/9789241547598_eng.pdf?sequence=1 [accessed 2023-12-12]
49. Kaufman EA, Xia M, Fosco G, Yaptangco M, Skidmore CR, Crowell SE. The difficulties in emotion regulation scale short form (DERS-SF): validation and replication in adolescent and adult samples. J Psychopathol Behav Assess 2015 Nov 23;38(3):443-455 [FREE Full text] [doi: 10.1007/s10862-015-9529-3]
50. Gratz KL, Roemer L. Multidimensional assessment of emotion regulation and dysregulation: development, factor structure, and initial validation of the difficulties in emotion regulation scale. J Psychopathol Behav Assess 2004 Mar;26(1):41-54 [FREE Full text] [doi: 10.1023/B:JOBA.0000007455.08539.94]
51. Brown KW, Ryan RM. The benefits of being present: mindfulness and its role in psychological well-being. J Pers Soc Psychol 2003 Apr;84(4):822-848. [doi: 10.1037/0022-3514.84.4.822] [Medline: 12703651]
52. Neacsiu AD, Rizvi SL, Vitaliano PP, Lynch TR, Linehan MM. The dialectical behavior therapy ways of coping checklist: development and psychometric properties. J Clin Psychol 2010 Jun;66(6):563-582. [doi: 10.1002/jclp.20685] [Medline: 20455249]
53. WHO ASSIST Working Group. The Alcohol, Smoking and Substance Involvement Screening Test (ASSIST): development, reliability and feasibility. Addiction 2002 Sep;97(9):1183-1194 [FREE Full text] [doi: 10.1046/j.1360-0443.2002.00185.x] [Medline: 12199834]
54. Collins RL, Parks GA, Marlatt GA. Social determinants of alcohol consumption: the effects of social interaction and model status on the self-administration of alcohol. J Consult Clin Psychol 1985 Apr;53(2):189-200. [doi: 10.1037//0022-006x.53.2.189] [Medline: 3998247]
55. Sadeh N, Baskin-Sommers A. Risky, impulsive, and self-destructive behavior questionnaire (RISQ): a validation study. Assessment 2017 Dec;24(8):1080-1094. [doi: 10.1177/1073191116640356] [Medline: 27002123]
56. McDonald RP. Test Theory: A Unified Treatment. Mahwah, NJ: Lawrence Erlbaum Associates; 1999.
57. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Soft 2015;67(1):1-48 [FREE Full text] [doi: 10.18637/jss.v067.i01]
58. Twisk J, Bosman L, Hoekstra T, Rijnhart J, Welten M, Heymans M. Different ways to estimate treatment effects in randomised controlled trials. Contemp Clin Trials Commun 2018 Jun;10:80-85 [FREE Full text] [doi: 10.1016/j.conctc.2018.03.008] [Medline: 29696162]
59. Kuznetsova A, Brockhoff PB, Christensen RH. lmerTest package: tests in linear mixed effects models. J Stat Soft 2017;82(13):1-26 [FREE Full text] [doi: 10.18637/jss.v082.i13]

60. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
61. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 2013 Nov 26;4:863 [FREE Full text] [doi: 10.3389/fpsyg.2013.00863] [Medline: 24324449]
62. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd edition. Oxfordshire, UK: Routledge; 1988.
63. Melville KM, Casey LM, Kavanagh DJ. Dropout from Internet-based treatment for psychological disorders. Br J Clin Psychol 2010 Nov;49(Pt 4):455-471. [doi: 10.1348/014466509X472138] [Medline: 19799804]
64. Dixon LJ, Linardon J. A systematic review and meta-analysis of dropout rates from dialectical behaviour therapy in randomized controlled trials. Cogn Behav Ther 2020 May;49(3):181-196. [doi: 10.1080/16506073.2019.1620324] [Medline: 31204902]
65. Lappan SN, Brown AW, Hendricks PS. Dropout rates of in-person psychosocial substance use disorder treatments: a systematic review and meta-analysis. Addiction 2020 Feb;115(2):201-217. [doi: 10.1111/add.14793] [Medline: 31454123]
66. Harley R, Sprich S, Safren S, Jacobo M, Fava M. Adaptation of dialectical behavior therapy skills training group for treatment-resistant depression. J Nerv Ment Dis 2008 Feb;196(2):136-143. [doi: 10.1097/NMD.0b013e318162aa3f] [Medline: 18277222]
67. McCauley E, Berk MS, Asarnow JR, Adrian M, Cohen J, Korslund K, et al. Efficacy of dialectical behavior therapy for adolescents at high risk for suicide: a randomized clinical trial. JAMA Psychiatry 2018 Aug 01;75(8):777-785 [FREE Full text] [doi: 10.1001/jamapsychiatry.2018.1109] [Medline: 29926087]
68. Neacsiu AD, Eberle JW, Kramer R, Wiesmann T, Linehan MM. Dialectical behavior therapy skills for transdiagnostic emotion dysregulation: a pilot randomized controlled trial. Behav Res Ther 2014 Aug;59:40-51. [doi: 10.1016/j.brat.2014.05.005] [Medline: 24974307]
69. Southward MW, Eberle JW, Neacsiu AD. Multilevel associations of daily skill use and effectiveness with anxiety, depression, and stress in a transdiagnostic sample undergoing dialectical behavior therapy skills training. Cogn Behav Ther 2022 Mar;51(2):114-129. [doi: 10.1080/16506073.2021.1907614] [Medline: 33881386]
70. Perroud N, Nicastro R, Jermann F, Huguelet P. Mindfulness skills in borderline personality disorder patients during dialectical behavior therapy: preliminary results. Int J Psychiatry Clin Pract 2012 Sep;16(3):189-196. [doi: 10.3109/13651501.2012.674531] [Medline: 22443984]
71. Zeifman RJ, Boritz T, Barnhart R, Labrish C, McMain SF. The independent roles of mindfulness and distress tolerance in treatment outcomes in dialectical behavior therapy skills training. Personal Disord 2020 May;11(3):181-190. [doi: 10.1037/per0000368] [Medline: 31647267]
72. Furukawa TA, Noma H, Caldwell DM, Honyashiki M, Shinohara K, Imai H, et al. Waiting list may be a nocebo condition in psychotherapy trials: a contribution from network meta-analysis. Acta Psychiatr Scand 2014 Sep;130(3):181-192. [doi: 10.1111/acps.12275] [Medline: 24697518]
73. Harned MS, Coyle TN, Garcia NM. The inclusion of ethnoracial, sexual, and gender minority groups in randomized controlled trials of dialectical behavior therapy: a systematic review of the literature. Clin Psychol Sci Pract 2022 Jun;29(2):83-93. [doi: 10.1037/cps0000059]
74. Hoeppner BB, Stout RL, Jackson KM, Barnett NP. How good is fine-grained timeline follow-back data? Comparing 30-day TLFB and repeated 7-day TLFB alcohol consumption reports on the person and daily level. Addict Behav 2010 Dec;35(12):1138-1143 [FREE Full text] [doi: 10.1016/j.addbeh.2010.08.013] [Medline: 20822852]
75. Carey MP, Carey KB, Maisto SA, Gordon CM, Weinhardt LS. Assessing sexual risk behaviour with the Timeline Followback (TLFB) approach: continued development and psychometric evaluation with psychiatric outpatients. Int J STD AIDS 2001 Jun;12(6):365-375 [FREE Full text] [doi: 10.1258/0956462011923309] [Medline: 11368817]

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
DBT: dialectical behavior therapy
iDBT: internet-delivered dialectical behavior therapy
RCT: randomized controlled trial
REDCap: Research Electronic Data Capture
SUD: substance use disorder

Edited by J Torous; submitted 26.08.23; peer-reviewed by W Zhang, L Asarnow; comments to author 06.10.23; revised version received 02.11.23; accepted 24.11.23; published 16.01.24.

Please cite as:
Daros AR, Guimond TH, Yager C, Palermo EH, Wilks CR, Quilty LC
Feasibility, Acceptability, and Potential Efficacy of a Self-Guided Internet-Delivered Dialectical Behavior Therapy Intervention for Substance Use Disorders: Randomized Controlled Trial
JMIR Ment Health 2024;11:e50399
URL: https://mental.jmir.org/2024/1/e50399
doi:10.2196/50399
PMID:38227362

Examining a Fully Automated Mobile-Based Behavioral Activation Intervention in Depression: Randomized Controlled Trial

Nicholas Santopetro[1], MS; Danielle Jones[1], BS; Andrew Garron[1], BS; Alexandria Meyer[2], PhD; Keanan Joyner[3], PhD; Greg Hajcak[2], PhD

1

2

3

Corresponding Author:
Nicholas Santopetro, MS

Abstract

Background: Despite significant progress in our understanding of depression, prevalence rates have substantially increased in recent years. Thus, there is an imperative need for more cost-effective and scalable mental health treatment options, including digital interventions that minimize therapist burden.

Objective: This study focuses on a fully automated digital implementation of behavioral activation (BA)—a core behavioral component of cognitive behavioral therapy for depression. We examine the efficacy of a 1-month fully automated SMS text message–based BA intervention for reducing depressive symptoms and anhedonia.

Methods: To this end, adults reporting at least moderate current depressive symptoms (8-item Patient Health Questionnaire score ≥10) were recruited online across the United States and randomized to one of three conditions: enjoyable activities (ie, BA), healthy activities (ie, an active control condition), and passive control (ie, no contact). Participants randomized to enjoyable and healthy activities received daily SMS text messages prompting them to complete 2 activities per day; participants also provided a daily report on the number and enjoyment of activities completed the prior day.

Results: A total of 126 adults (mean age 32.46, SD 7.41 years) with current moderate depressive symptoms (mean score 16.53, SD 3.90) were recruited. Participants in the enjoyable activities condition (BA; n=39) experienced significantly greater reductions in depressive symptoms compared to participants in the passive condition (n=46). Participants in both active conditions—enjoyable activities and healthy activities (n=41)—reported reduced symptoms of anxiety compared to those in the control condition.

Conclusions: These findings provide preliminary evidence regarding the efficacy of a fully automated digital BA intervention for depression and anxiety symptoms. Moreover, reminders to complete healthy activities may be a promising intervention for reducing anxiety symptoms.

Trial Registration: ClincalTrials.gov NCT06492824; https://clinicaltrials.gov/study/NCT06492824

(JMIR Ment Health 2024;11:e54252) doi:10.2196/54252

KEYWORDS

digital intervention; digital health; digital application; digital applications; mobile health; mHealth; automation; automate; automated; behavioral activation; BA; BA intervention; depression; depressed; depressive; depressive symptoms; anhedonia; anhedonia symptoms; anxiety; anxious; anxiety symptoms; adults; adult; psychiatry; psych; psychology; major depressive disorder; MDD

Introduction

Depressive disorders remain one of the most prevalent and burdensome mental health disorders worldwide [1-3]. Individuals with depression are more at risk for suicide and substance abuse, experience significant occupational impairments (eg, missing work), and have reduced life expectancy [4,5]. Research has also highlighted that anhedonia, a core depressive symptom characterized by a reduced ability to experience pleasure or enjoyment, is particularly pernicious as it is linked to worse treatment responsiveness and depression remission [6,7].

Despite advancements in our understanding and identification of depression, prevalence rates of depressive disorders have generally remained constant in the general population over the past decades [8,9]. However, recent studies examining the impact of the COVID-19 pandemic on internalizing disorders suggest substantial increases in rates of depressive disorders [10-12]. Although evidence-based treatments such as cognitive behavioral therapy (CBT) and behavioral activation (BA) have been extensively underscored as effective psychotherapy interventions for depression, current shortages of professionals in the United States mental health care system are significantly compounding the burden of these rising depression rates. More specifically, in the United States in 2018, approximately 115 million people reported living in a mental health professional shortage area, and approximately 25% of adults with a current mental health disorder did not receive treatment [13]. Thus, depressive disorders remain a persistent and taxing global public health issue, and there is a crucial need for briefer, more accessible, and more affordable evidence-based mental health interventions for depression to supplement these shortages in mental health services.

Digital mental health interventions are one possible solution for making evidence-based psychotherapy treatments more accessible and less expensive. Digital interventions involve using digital technology, such as smart devices, the internet, or mobile apps, to foster or support behavior change. For example, more recent research efforts have examined the effectiveness of chatbot-based interventions in increasing positive affect [14]. Research examining the effectiveness of digital interventions for the treatment of mental health problems has increased rapidly over the past three decades: recent reviews and meta-analyses suggest that digital interventions appear to be effective at reducing symptoms of depression [15-17]. There was a medium overall effect size of digital interventions for depression when compared to control conditions (Hedges g=0.52), and the vast majority (74%) of digital interventions used to treat depressive disorders rely on the tenets of CBT as their theoretical foundation [16]. Importantly, this meta-analysis did not find differences in treatment effect sizes between digital interventions and more traditional in-person psychotherapy for depression.

Various meta-analytic reviews have concluded that BA, a specific aspect of CBT for depression, is itself a highly effective treatment for depressive disorders [18,19]. From the perspective of BA, depressive symptoms are a result of few positive experiences coincident with increases in negative experiences, which then lead to a diminished pursuit for further positive experiences [20]. BA treatment therefore attempts to disrupt this negative feedback loop by purposefully increasing an individual's participation in enjoyable or rewarding (ie, positive) activities to increase positive mood and reduce negative affect [21].

Past studies have examined the effectiveness of digital BA interventions for depression, and results indicate that in nonclinical settings, guided internet-based BA interventions were equivalent to other forms of behavioral therapy and superior to waitlist conditions at reducing depression and anxiety outcomes [22]. Despite these promising findings, existing digital BA interventions still require a relatively high level of therapist guidance or interaction from the study team. For example, Ly and colleagues [23] administered an 8-week smartphone-based BA intervention that required a trained therapist to contact each participant weekly for no longer than 20 minutes as part of the standard intervention protocol (ie, up to 2 h and 40 min of therapist contact per participant). Similarly, other digital BA interventions have required a trained therapist to review completed assignments and provide personalized feedback for each participant [24,25].

Although these digital interventions require less therapist time than conventional psychotherapy, the ability to deliver digital BA interventions at scale would be increased if interventions did not require human interaction (ie, if they were fully automated). As of now, the existing digital BA protocols investigated in the literature would not effectively assist individuals living in mental health professional shortage areas as there is still a reliance on trained mental health professionals to deliver these interventions, restricting access to services. To our knowledge, a fully automated digital BA protocol for depression has not been examined previously. Furthermore, and in keeping with recent calls for brief and low-intensity mental health interventions in the context of rising mental health issues due to the COVID-19 pandemic [26], the current digital BA intervention focused specifically on the core mechanism of change in BA: increasing enjoyable experiences. Thus, the primary goal of this study was to examine if a fully automated digital intervention designed to promote and increase positive experiences among individuals with elevated depression would result in significant reductions in depressive symptoms over 1 month. The second goal of this study was to assess the impact of the BA digital interventions on symptoms of anhedonia—a symptom of depression that is directly related to positive experiences [6,7,22].

Participants with higher levels of depression (ie, moderate or higher depressive symptoms based on self-report on the 8-item Patient Health Questionnaire [PHQ-8]) were randomized to an automated digital (ie, SMS text message based) intervention prompting individuals to complete daily enjoyable activities for 1 month. We compared this digital BA intervention to a similar automated digital (ie, SMS text message based) intervention that prompted participants to complete healthy living activities (eg, bathing, brushing teeth). Finally, we included a passive control condition that did not prompt any behavior changes. Based on past meta-analytic studies on the effectiveness of both traditional and digital BA treatments, we predict that participants in the enjoyable activities (ie, BA) condition would experience greater decreases in depressive symptoms and anhedonia compared to participants in the healthy activities and passive control conditions. As an exploratory aim, we also investigated if these interventions significantly influenced changes in anxiety and stress over the 4 weeks.

Methods

Ethical Considerations

The project was approved by the Florida State University Institutional Review Board (00003451) and was retrospectively registered as a randomized clinical trial (NCT06492824). Participants provided informed consent before participating and were compensated for their time (US $20 for completing the pre- and postsurveys, and an additional US $30 if participants completed at least 90% of their daily checklists over the 4-week time frame, if applicable).

Participants

Participants were recruited online from the United States via Facebook advertisements from November 2022 to February 2023 (advertisements included in Multimedia Appendix 1).

Randomization was conducted during the initial baseline survey using Qualtrics (version XM) software and coding. More specifically, if the interested participants passed the eligibility criteria (discussed in more detail below), the survey was coded to randomize them to either the BA, healthy activity, or passive control conditions.

The inclusion criteria for this study were being 18 years or older, having a current PHQ-8 total score ≥10 (ie, moderate depressive symptoms or greater), and successfully signing up for the SlickText messaging service (ie, providing phone number and following prompts sent to that number to enroll in the service). More specifically, interested participants were first asked to consent to all study procedures and complete the PHQ-8 during the baseline survey. Interested participants who either did not consent to the study or scored lower than a 10 on the PHQ-8, which resulted in the survey terminating, would not be randomized to an experimental condition.

Measures

8-Item Patient Health Questionnaire

The PHQ-8 is an 8-item measure of current depressive symptoms that assesses 8 of the 9 core symptoms that are the criteria for depression, excluding symptoms of suicidal ideations and behaviors [27]. The questionnaire was administered only during the baseline survey and assessed depressive symptoms over the last 2 weeks to assess eligibility for this study. Each item is scored on a 4-point scale (ranging from 0 to 3; 0=not at all, 1=several days, 2=more than half the days, and 3=nearly every day). Therefore, PHQ-8 total scores range from 0 to 24, and a score ≥10 is considered a validated threshold for clinical levels of depression [27], which is why this measure was used as a screening tool for this study. In this sample, the PHQ-8 total score demonstrated good internal consistency at the baseline assessment (Cronbach α=0.78).

21-Item Depression, Anxiety, and Stress Scale

The 21-item Depression, Anxiety, and Stress Scale (DASS-21) comprises 21 total items; 7 items that assess depressive symptoms, 7 items that assess anxiety symptoms, and 7 items that assess stress symptoms over the past week. Each item is scored on a 4-point scale (ranging from 0 to 3; 0=did not apply to me at all, 1=applied to me to some degree or some of the time, 2=applied to me to a considerable degree or a substantial part of the time, and 3=applied to me very much or most of the time). Participants in this study completed the DASS-21 before and after the 4-week intervention or control period. Past research suggests that the DASS-21 accurately distinguishes features of depression, physical arousal, and psychological tension and agitation. Additionally, these psychometric studies observe acceptable-to-excellent internal consistency and concurrent validity [28]. In this sample, Cronbach α showed good-to-excellent internal consistencies at both the pre- and postassessment depression scores (pre=0.90; post=0.89), anxiety scores (pre=0.88; post=0.86), and stress scores (pre=0.87; post=0.86).

Personality Inventory for Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition): Anhedonia Subscale

The anhedonia subscale of the Personality Inventory for *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition; PID-5) consists of eight items rated on a 4-point Likert scale. Participants in this study completed the PID-5 subscale before and after the 4 weeks, reporting symptoms from over the past 2 weeks. Past research involving this subscale of the PID-5 suggests good internal consistency (Cronbach α ranging from 0.87 to 0.89) in various adult samples [29,30]. In line with these past studies, the PID-5 anhedonia subscale scores in this sample demonstrated good internal consistency according to Cronbach α at the pre- and postintervention assessments (0.75 and 0.85, respectively).

Procedures

Enjoyable Activities, Healthy Activities, and Passive Control Conditions

The active interventional conditions (ie, enjoyable activities [BA] and healthy activities) were delivered automatically via daily SMS text messages using the service SlickText. SlickText provides an SMS text message product that enabled participants to receive automated communications from the research group. Participants randomized to the enjoyable activities and healthy activities conditions were instructed to enroll in SlickText after consenting to this study and provided their phone number, email address, and initials; no other personal information was collected/shared with SlickText.

Participants in the enjoyable activities condition watched an introduction video (approximately 3 minutes in length) that discussed the rationale for the enjoyable activities intervention in combating depressive symptoms during the initial online assessment (video script provided in Multimedia Appendix 2). Additionally, during the initial assessment, enjoyable activities participants were asked to choose 5 enjoyable activities from a large list of various activities from several domains such as social, soothing/fun, physical, and religious (full list provided in Multimedia Appendix 3) that they wanted to do more frequently over the next 4 weeks. Some examples of enjoyable activities from the list were go for a walk, stargaze, go for a drive, go to a service, do some gardening, play with your pet, and text or call one of your friends. Over the next 4 weeks, daily SMS text message reminders were sent, via SlickText, each morning. These SMS text messages reminded participants of the 5 enjoyable activities that they wanted to do more of and encouraged them to complete at least 2 of the enjoyable activities that day. Participants were asked to try to complete at least 2 out of their 5 selected activities each day, though they were told that they could complete more if they wanted. Additionally, participants received a link to complete a daily checklist (ie, short survey) where they reported if they had completed any of their 5 activities on the previous day (example SMS text messages in Multimedia Appendix 4). Next, participants were asked to rate how much they enjoyed doing their completed activities (0 - 10 rating) from the previous day on the checklist (short survey). Participants were also asked to complete short weekly questionnaires asking about their mood for the preceding week. At the end of the 4 weeks, enjoyable activities participants completed a postintervention online survey consisting of the same measures from the preintervention survey.

Participants in the healthy activities condition completed most of the same tasks as those participants in the enjoyable activities condition: they completed an initial online preintervention survey and chose 5 enjoyable activities from the same list. However, healthy activities participants were not explicitly asked to complete these activities for their intervention. Instead, participants in the healthy activities condition were encouraged to complete the following 5 activities each day: drinking more water, going to bed early, brushing their teeth, showering or bathing that day, and eating well-balanced meals. Healthy activities condition participants also watched an introduction video (approximately 3 minutes in length) that discussed the rationale for the healthy activities intervention in combating depressive symptoms during the initial online assessment (video script provided in Multimedia Appendix 5).

Daily SMS text reminders were sent, via SlickText, regarding these 5 healthy living activities each morning. In that same message, healthy activities participants received a link to a short online checklist (survey) where they reported which healthy living activities they completed the previous day and how much they enjoyed completing those healthy activities. Consistent with the enjoyable activities condition, healthy activities participants were asked to try to complete at least 2 of the 5 healthy activities each day but could do more if they desired. They were also asked to complete a short weekly questionnaire asking about their mood during the preceding week. At the end of the month, they completed an online postintervention survey consisting of the same measures from the preintervention assessment. Participants in this condition were given the option to complete the enjoyable activities digital intervention after completing the healthy activities condition if they desired.

Control participants completed an initial online survey that also included an introduction video (approximately 1 and a half minutes long; video script provided in Multimedia Appendix 6) explaining the rationale for this specific condition and that researchers wanted to understand the naturalistic course of mood and behavior over 1 month. However, passive control participants were not explicitly asked to complete any enjoyable activities or healthy living activities during the subsequent 4 weeks. Passive control participants were not sent daily text reminders or checklists to complete over the 4 weeks. They were asked to complete the same short weekly questionnaire asking about their mood during the preceding week. At the end of the month, they also completed a postintervention online survey consisting of the same measures from the initial survey. Participants in this condition also had the option to complete the enjoyable activities digital intervention after completing the passive control condition if they desired.

Statistical Analyses

All data analyses were performed using SPSS Statistics software (version 27.0; IBM Corp). First, demographic and clinical data were analyzed between the enjoyable activities, healthy activities, and passive control groups using 1-way ANOVAs, independent samples *t* tests, and χ^2 tests. Next, a total of 4 univariate ANOVA models were conducted to determine the effect of group status on changes in symptoms of depression, anhedonia, anxiety, and stress over 4 weeks. Specifically, the experimental condition (ie, enjoyable activities [BA], healthy activities, and passive control) was used as a fixed factor predicting postintervention symptoms; the respective preintervention symptoms were used as covariates in each ANOVA model. Bonferroni corrections were used for all post hoc procedures in the ANOVA models. Considering that multiple models are being conducted, the criterion of statistical significance for all analyses was adjusted and set at .013 (ie, .05 divided by 4). Lastly, demographic and clinical data were analyzed between participants who completed the treatment procedures entirely (ie, returned for follow-up and engaged with their respective interventions according to the completion of daily checklists) and participants who did not complete the treatment procedures or return for the follow-up using independent samples *t* tests and χ^2 tests to assess attrition bias.

Power Analysis

The G*Power (Version 3.1; Erdfelder, Faul, and Buchner) application was used to conduct the a priori power analysis [31]. Based on meta-analytic findings, the estimated effect size between BA treatment effectiveness on depression compared to controls that did not receive treatment is expected to be large (Hedges g=0.83) [19]. Using 80% power and a large estimated effect size (f=0.40 or η^2=0.14), the recommended total sample size for a 3-group analysis of covariance is 64 participants, and therefore the current sample size is adequately powered to detect the presence of the BA treatment effect on depressive symptoms.

Results

The full sample initially recruited and randomly allocated to the experimental conditions comprised 126 eligible participants: 39 randomized to the enjoyable activities condition, 41 randomized to the healthy activities condition, and 46 randomized to the control condition. Of these 126 randomized participants, 1 participant did not engage with the BA intervention, and 4 participants did not engage with the healthy activities intervention over the month as evidenced by the average number of daily completed activities recorded in their checklists (mean activities 0.28, SD 0.29). Figure 1 displays the CONSORT (Consolidated Standards of Reporting Trials) diagram, which further details the number of participants during initial engagement and recruitment, allocation to the experimental conditions, follow-up data collected, and final analyses for this study.

A total of 97 participants (enjoyable activities: n=31; healthy activities: n=23; passive control: n=43) completed pre- and postintervention measures of depressive, anxiety, and stress symptoms (ie, DASS-21), and 98 participants (enjoyable activities: n=27; healthy activities: n=28; passive control: n=43) completed pre- and postintervention measures of anhedonia (ie, PID-5 anhedonia subscale). The average age of the total sample was 32.20 (SD 7.14) years, and 28% (n=28) of the sample identified as female. In terms of race, half (n=49, 50%) of the participants identified as White, whereas 41% (n=40) of participants identified as African American or Black, 4% (n=3) identified as Asian, 3% (n=3) identified as American Indian or Alaskan Native, 1% (n=1) identified as Pacific Islander, and 1% (n=1) of participants identified as another race. Furthermore, 17% (n=17) identified as Hispanic. Additionally, 34% (n=33) of the sample reported currently taking psychotropic medications (eg, selective serotonin reuptake inhibitor) for symptoms of depression. Table 1 has more detailed information regarding this information as a function of the three conditions.

There were no significant age, gender, or racial composition differences between the enjoyable activities, healthy activities, and passive control participants in this study (all *P* values >.07). Moreover, there were no significant differences in preintervention symptoms of depression (PHQ-8 or DASS-21), anxiety, stress, or anhedonia between the three groups (all *P* values >.33). On average, participants in this study reported moderate-to-severe levels of depressive symptoms at the initial assessment according to both the PHQ-8 and DASS-21. Additionally, there were no differences in current psychotropic medication status between the groups (*P*=.12). There was a significant difference in the average number of daily activities completed (*P*=.002) between the two active intervention conditions such that participants in the healthy activity group completed more activities each day compared to individuals in the enjoyable activities condition. Surprisingly, there were no differences in average enjoyment ratings of these completed daily activities (*P*=.59) between participants in the enjoyable activities and healthy activities conditions; on average, participants in these conditions rated their completed activities as being highly enjoyable (approximately 7 on a 0-10 scale).

To assess the impact of the intervention on changes in depressive symptoms, experimental condition (ie, enjoyable activities [BA], healthy activities, and passive control) was entered as a fixed factor in a univariate ANOVA predicting the post–DASS-21 depression score; the pre–DASS-21 depression score was also used as a covariate in the model. The overall ANOVA model was significant ($F_{3,98}$=3.06; *P*=.03). Experimental condition emerged as a significant predictor of postintervention depressive symptoms ($F_{2,98}$=4.53; *P*=.01; η_p^2=0.09). Further, Bonferroni follow-up comparisons revealed that individuals in the BA condition had significantly lower depressive scores after 1 month than individuals in the passive control condition (mean difference −7.27, SE 2.42, 95% CI –13.18 to –1.36; *P*=.01). There were no significant differences in depressive scores for individuals in the BA condition compared to individuals in the healthy activity condition (mean difference −3.55, SE 2.78, 95% CI −10.33 to 3.22; *P*=.61). There was also no difference in postintervention depressive symptoms between individuals in the healthy activity condition and the passive control condition (mean difference −3.72, SE 2.62, 95% CI −10.09 to 2.66; *P*=.48). Lastly, preintervention DASS-21 depressive scores were not significantly associated with postinterevention DASS-21 depressive scores in the model ($F_{1,98}$=0.50; *P*=.48; η_p^2=0.01). Figure 2A displays changes in depressive symptoms from pre- to postintervention in all three groups.

To further examine the impact of the intervention specifically on symptoms of anhedonia, experimental condition (ie, enjoyable activities [BA], healthy activities, and passive control) was entered as a fixed factor in a univariate ANOVA predicting postintervention PID-5 anhedonia scores; preintervention PID-5 anhedonia scores were also used as a covariate in the model. The overall ANOVA model was not significant ($F_{3,98}$=1.87; *P*=.14). Experimental condition also did not significantly predict postintervention anhedonia symptoms ($F_{2,98}$=2.52; *P*=.09; η_p^2=0.05). Preintervention PID-5 anhedonia scores were not significantly associated with postintervention PID-5 anhedonia scores in the model ($F_{1,98}$=0.22; *P*=.64; η_p^2=0.00). Figure 2B displays changes in anhedonia symptoms from pre- to postintervention in all three groups.

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram. BA: behavioral activation; DASS-21: 21-item Depression, Anxiety and Stress Scale; PID-5: Personality Inventory for *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition).

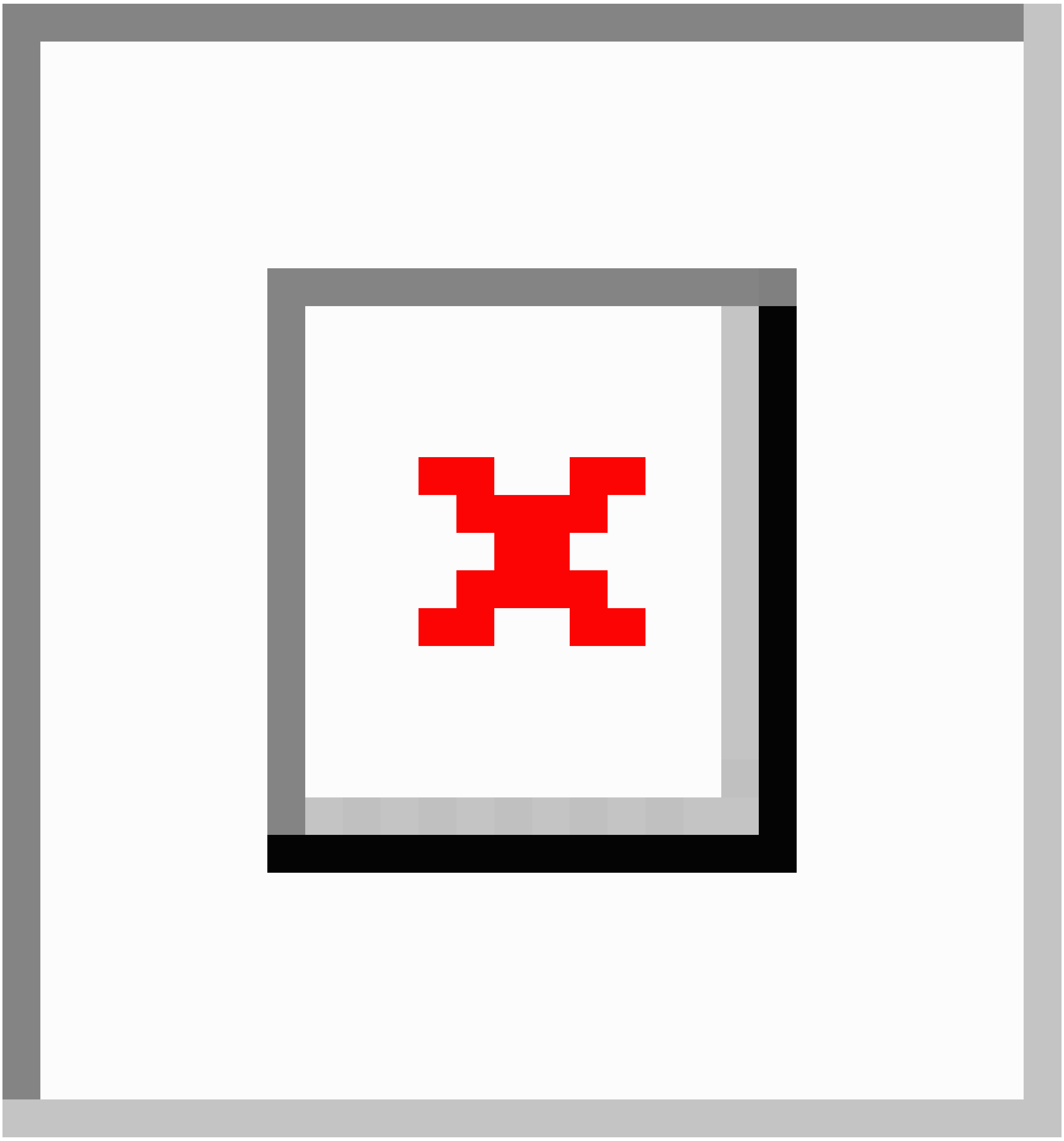


Table . Demographic and clinical measures in enjoyable activities, healthy activities, and passive control groups.

	Enjoyable activities (n=31)	Healthy activities (n=28)	Passive control (n=43)	*P* value
Demographics and intervention				
Age (years), mean (SD)	32.48 (7.41)	32.53 (7.63)	31.74 (6.70)	.87
Gender (female), n (%)	10 (32)	11 (39)	7 (16)	.06
Race, n (%)				.59
White	12 (39)	15 (54)	22 (51)	
Black/African American	14 (45)	11 (39)	15 (35)	
Asian	0 (0)	1 (4)	2 (5)	
Current psychotropic medication, n (%)	15 (48)	9 (31)	11 (26)	.12
Number of completed daily activities, mean (SD)	2.86 (0.78)	3.50 (0.77)	—[a]	.002
Enjoyment rating of activities (0 - 10), mean (SD)	7.40 (1.73)	7.18 (1.56)	—	.59
Clinical, mean (SD)				
Pre–PHQ-8[b]	16.81 (3.98)	17.18 (3.58)	15.88 (4.05)	.33
Pre–DASS-21[c] depression	24.90 (10.00)	25.33 (12.74)	27.58 (9.07)	.49
Pre–DASS-21 anxiety	20.13 (10.18)	21.92 (12.41)	23.03 (9.88)	.51
Pre–DASS-21 stress	25.61 (9.56)	23.86 (11.25)	26.34 (8.35)	.59
Pre–PID-5[d] anhedonia	14.63 (5.26)	14.86 (2.98)	13.70 (3.72)	.45
Post–DASS-21 depression	10.84 (10.52)	13.71 (10.53)	17.91 (9.53)	.01
Post–DASS-21 anxiety	10.97 (10.17)	12.41 (8.59)	19.44 (8.97)	<.001
Post–DASS-21 stress	13.87 (10.97)	15.94 (8.98)	17.01 (10.52)	.43
Post–PID-5 anhedonia	8.94 (5.55)	10.19 (5.19)	11.61 (4.21)	.07

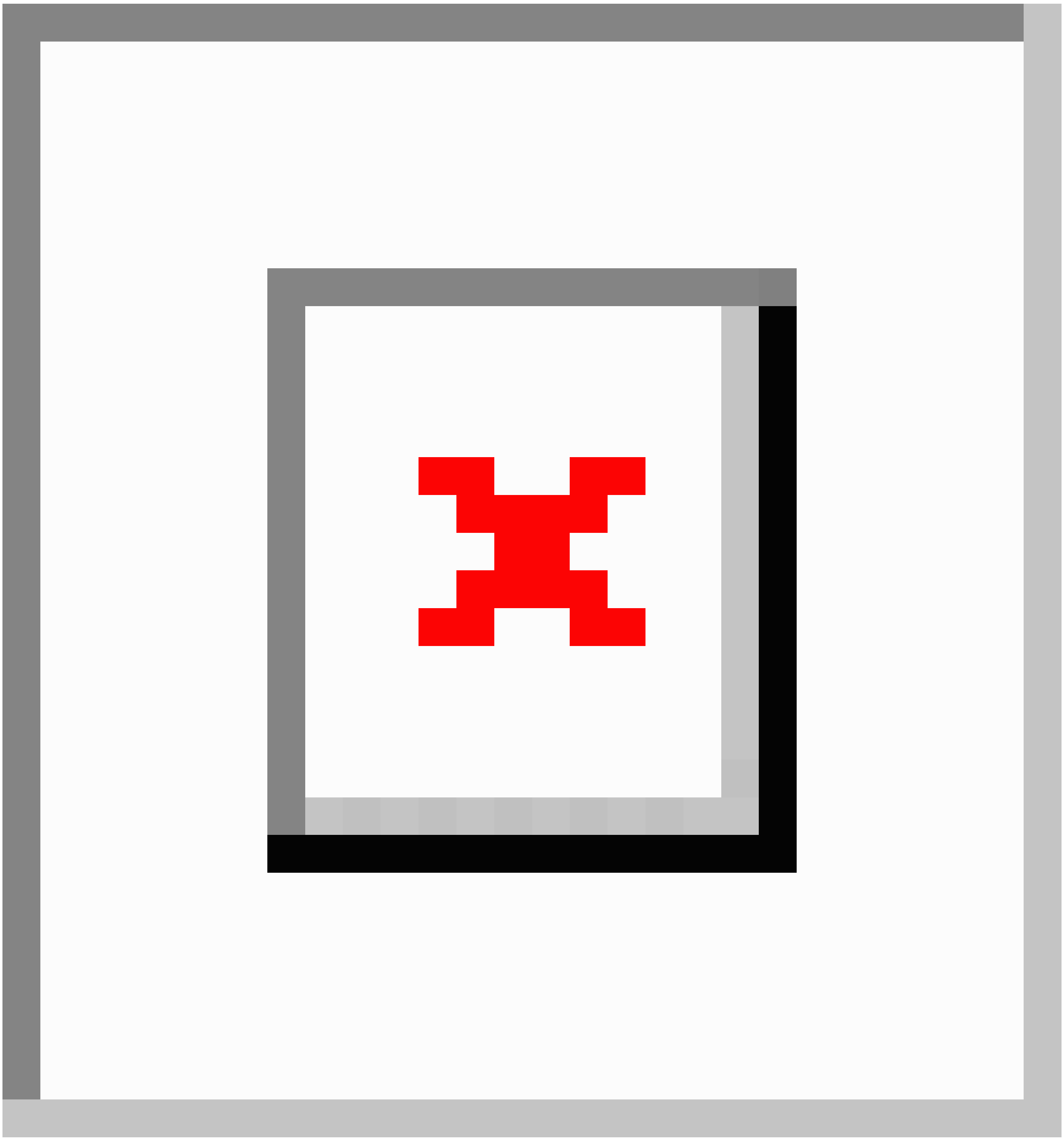
[a]Not applicable.

[b]PHQ-8: 8-item Patient Health Questionnaire.

[c]DASS-21: 21-item Depression, Anxiety, and Stress Scale.

[d]PID-5: Personality Inventory for *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition).

Figure 2. Changes in (A) depressive symptoms, (B) anhedonic symptoms, and (C) anxiety symptoms over 4 weeks by condition. BA: behavioral activation; DASS-21: 21-item Depression, Anxiety, and Stress Scale; PID-5: Personality Inventory for *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition).



A similar ANOVA model was conducted to examine if either active condition modulated changes in anxiety symptoms. Experimental condition (ie, enjoyable activities [BA], healthy activities, and passive control) was entered as a fixed factor in a univariate ANOVA predicting the postintervention DASS-21 anxiety score; the preintervention DASS-21 anxiety score was also used as a covariate in the model. The overall ANOVA model was significant ($F_{3,98}$=5.48; *P*=.002). Experimental condition emerged as a significant predictor of postintervention anxiety symptoms ($F_{2,98}$=8.18; *P*=.001; η_p^2=0.15). Bonferroni follow-up comparisons revealed that individuals in the BA condition had significantly lower anxiety scores after 1 month than individuals in the passive control condition (mean difference −8.51, SE 2.20, 95% CI −13.88 to −3.14; *P*=.001). There were no significant differences in postintervention anxiety scores between individuals in the BA condition compared to individuals in the healthy activity condition (mean difference −2.35, SE 2.53, 95% CI −3.82 to 8.51; *P*>.99). Additionally, individuals in the healthy activity condition reported significantly lower postintervention anxiety symptoms than individuals in the passive control condition (mean difference −6.17, SE 2.37, 95% CI −11.94 to −0.39; *P*=.03). Lastly, preintervention DASS-21 anxiety scores were not significantly associated with postintervention DASS-21 anxiety scores in the model ($F_{1,98}$=0.02; P=.88; η_p^2=0.00). Figure 2C displays changes in anxiety symptoms from pre- to postintervention in all three groups.

A final ANOVA model was conducted to examine if participants in either of the active conditions experienced greater changes in symptoms related to stress. Consistent with the previous ANOVA models, experimental condition (ie, enjoyable activities [BA], healthy activities, and passive control) was entered as a fixed factor in a univariate ANOVA predicting postintervention DASS-21 stress scores; preintervention DASS-21 stress scores were also used as a covariate in the model. The overall ANOVA model was not significant ($F_{3,98}$=0.70; P=.55). Experimental condition also did not significantly predict postintervention stress symptoms ($F_{2,98}$=0.92; P=.40; η_p^2=0.02). Lastly, preintervention DASS-21 stress scores were not significantly associated with postintervention DASS-21 stress scores in the model ($F_{1,98}$=0.26; P=.61; η_p^2=0.00).

There were no significant age, gender, or racial composition differences between participants who completed the intervention procedures and participants who did not complete the study (all P values >.05). Additionally, there were no significant differences in preintervention symptoms of depression, anxiety, stress, or anhedonia between the two groups (all P values >.06). Lastly, there were no differences in current psychotropic medication status between the groups (P=.14).

Discussion

This study examined the efficacy of a fully automated digital BA intervention at reducing depressive symptoms and anhedonia over 1 month in individuals with moderate or greater current symptoms of depression. Compared to a fully automated digital active control condition in which individuals were prompted to increase healthy activities and a control condition in which individuals were not prompted to make any lifestyle changes, individuals who completed the fully automated digital BA intervention experienced greater reductions in their depressive symptoms compared to the control condition; there were no significant differences in depressive symptoms changes between the healthy activities and control condition. However, there did not appear to be specific reductions in anhedonia related to the digital BA intervention. Moreover, individuals that completed either the active digital intervention (ie, enjoyable activities [BA] or healthy activities) also reported significant decreases in symptoms of anxiety over the 4 weeks compared to the passive control condition. Neither digital intervention appeared to be effective at reducing symptoms attributed to stress.

These findings are directly in line with past research that has demonstrated BA as a frontline treatment for depressive disorders [18,19] and further supports the digital implementations of BA treatment for reducing symptoms of depression [22]. Unlike past digital BA projects, this study implemented a completely automated treatment protocol requiring no human intervention. Considering the rising prevalence rates of depression compounded by limited access to evidence-based mental health services and experts, there is an imperative need to create scalable effective mental health interventions that can be widely disseminated [26].

Individuals who completed the digital BA intervention did not experience a significant decrease in symptoms of anhedonia compared to individuals assigned to the other two experimental conditions. Anhedonia is a particularly deleterious symptom of depression in that it is associated with negative outcomes such as a prolonged disease course, worse long-term prognosis, and higher suicide rates, making it a difficult symptom to modulate [32,33]. The current results are at odds with other studies that have observed that traditional (ie, nondigital) BA treatment can alleviate this particular core symptom of depression [34] and neuroimaging studies that further suggest that successful BA leads to increases in brain systems associated with reward processes, such as increases in activation of the dorsal striatum linked to reward anticipation, which is typically reduced in individuals experiencing heightened depression and anhedonia [35,36]. Although the current digital BA intervention did not specifically reduce symptoms of anhedonia, it is still plausible that other digital versions of BA treatment could likely incite the same mechanism of change seen in traditional BA in that the intervention successfully increases positive more rewarding experiences, which in turn decreases negative affect and perhaps more specifically anhedonia.

Participants who completed both the BA and healthy activities interventions also experienced significant decreases in anxiety symptoms over 1 month. Despite both conditions prompting daily lifestyle changes, reductions in anxiety symptoms seen in the BA and healthy activities conditions appear to be independent of one another. Past research has demonstrated that BA treatment can be effective at treating symptoms of anxiety [37,38], which most researchers attribute to disruptions of avoidant behaviors and behavioral inhibition, which are core features of anxiety [37]. Regarding healthy activities such as eating more well-balanced meals, bathing regularly, and going to bed early, past research has demonstrated that individuals with elevated anxiety are more likely to report disruptions of these psychosocial rhythms and poor general health even after controlling for current depressive symptoms [39]. Factors such as poor nutrition and impaired sleep share a bidirectional relationship with elevated symptoms of anxiety [40,41]. Therefore, promoting more consistent participation in these healthy activities daily should affect anxiety symptoms. Thus, it is possible that BA and healthy activities reduced anxiety symptoms via different mechanisms of change—though this possibility requires further evidence.

Surprisingly, there were no significant differences in the average enjoyment ratings of completed activities between the BA (ie, enjoyable activities) and healthy activities conditions. Moreover, individuals in the healthy activity condition completed more activities over the month compared to individuals in the BA condition. It is possible that although participants reported similar levels of enjoyment on average in their completed activities, there are likely other key variables that were not assessed in this study regarding these experiences that significantly alleviated low mood. For example, it is plausible that the amount of time spent doing the activity influences its effectiveness on positive affect and therefore negative affect.

Additionally, it is possible that completing activities from different domains might be more effective at increasing positive affect (eg, social activities compared to individual activities, physical activities compared to more sedentary hobbies). Future studies should expand on the activity information that is gathered from participants during BA interventions to further elucidate the mechanisms underlying reductions in depressive symptoms seen in BA treatment.

Lastly, we observed that almost half of the sample identified as Black or African American in this randomized controlled trial. Although this number is disproportionate to the percentage of Black or African Americans in the United States in general, this percentage might be more proportionate when considering differences in ethnic and racial prevalence rates of depression in the United States. Studies posit that rates of depression are higher for individuals identifying as Black or African American [42-44], which might be reflected here considering recruitment for the study was aimed toward individuals with low mood. Relatedly, it is commonly observed that Black or African American individuals seek traditional mental health help at much lower rates than White Americans [45]. It is plausible that the unique virtual and automated design of the present intervention was more attractive for individuals as it was more convenient and posed fewer barriers compared to other interventions with in-person aspects.

This study has limitations worth noting. Participants in this study were not administered clinical interviews to confirm the diagnosis of a depressive disorder. Although higher PHQ-8 scores, specifically total scores ≥10, suggest an increased likelihood of a depressive disorder [46], this was not confirmed by clinical interview, so the clinical status of our participants is unknown. More detailed information regarding depressive disorders is typically collected in diagnostic interviews, such as number of past episodes, age of onset, and the duration of the current episode; these are all crucial factors that have been associated with worse treatment response [47] that could be examined in future studies. Relatedly, overall depressive symptoms improved for participants in each condition over 1 month in this study. This is likely accounted for by regression toward the mean. More specifically, it is possible that depressive symptoms reported at the initial session reflected extreme symptoms and would therefore be a motivating factor to participate in this study. Therefore, reassessing these symptoms again a month later could more naturally reflect their average depressive ratings regardless of experimental condition. Additionally, a positive placebo response is often seen in clinical trials for depression [48]. However, despite these possibilities, there is still evidence to suggest that the BA intervention particularly led to more significant decreases in depression compared to the passive control group. Additionally, it would have been important to understand how other comorbid mental health disorders, such as anxiety disorders or substance use disorders, might have influenced responsiveness to the current digital interventions. Although current psychotropic medication information was collected from participants in the study and examined between the three experimental conditions, we did not assess for detailed information regarding current psychotherapy, which limits our findings. It is plausible that there might have been differences between the three groups regarding current therapy enrollment outside of the current interventions, which could have influenced the effectiveness of the digital BA intervention.

The current randomized controlled trial was not preregistered and used Facebook to recruit participants. Although Facebook allowed the research team to conveniently reach a large number of interested diverse participants from around the United States in a short amount of time, the overall quality of these participants could not be established. However, the sound psychometric properties (ie, Cronbach α ranging from 0.75 to 0.90) of the various clinical measures used as outcomes suggest the validity and reliability of data collected from the sample. Future studies might replicate these results using other online platforms such as Amazon Mechanical Turk, which allows researchers to recruit from more "vetted" registered participant pools [49]. This study also had significant strengths that are worth noting. As discussed earlier, this study is the first to implement a fully automated digital BA treatment for depression that requires no human intervention, which is highly scalable and could be widely accessible. Additionally, this study did not predominately involve White participants, which is a chronic issue in psychological research that typically hinders the generalizability of results [50].

In summary, this study demonstrated the efficacy of a fully automated BA intervention (ie, promoting daily enjoyable activities) for decreasing depressive symptoms over 1 month. Additionally, both automated digital intervention protocols (ie, BA and healthy activities) significantly reduced symptoms of anxiety over 4 weeks. Neither digital intervention had a significant effect on anhedonia or stress symptoms.

Data Availability

Data or other materials are available through correspondence with the authors.

Conflicts of Interest

None declared.

Editorial Notice

This randomized study was only retrospectively registered. Oversight and the aim of the study to test the feasibility of a fully automated intervention protocol were why the study was not prospectively registered. The editor granted an exception from the

International Committee of Medical Journal Editors rules mandating prospective registration of randomized trials as the risk of bias appears low.

Multimedia Appendix 1
Advertisements.
[DOCX File, 16 KB - mental_v11i1e54252_app1.docx]

Multimedia Appendix 2
Enjoyable activities video script.
[DOCX File, 14 KB - mental_v11i1e54252_app2.docx]

Multimedia Appendix 3
List of enjoyable activities.
[DOCX File, 14 KB - mental_v11i1e54252_app3.docx]

Multimedia Appendix 4
Example of SMS text messages.
[DOCX File, 13 KB - mental_v11i1e54252_app4.docx]

Multimedia Appendix 5
Healthy activities video script.
[DOCX File, 14 KB - mental_v11i1e54252_app5.docx]

Multimedia Appendix 6
Passive condition video script.
[DOCX File, 13 KB - mental_v11i1e54252_app6.docx]

Checklist 1
CONSORT-EHEALTH checklist (V 1.6.1)
[PDF File, 2803 KB - mental_v11i1e54252_app7.pdf]

References

1. Goodwin RD, Dierker LC, Wu M, Galea S, Hoven CW, Weinberger AH. Trends in U.S. depression prevalence from 2015 to 2020: the widening treatment gap. Am J Prev Med 2022 Nov;63(5):726-733. [doi: 10.1016/j.amepre.2022.05.014] [Medline: 36272761]
2. Lim GY, Tam WW, Lu Y, Ho CS, Zhang MW, Ho RC. Prevalence of depression in the community from 30 countries between 1994 and 2014. Sci Rep 2018 Feb 12;8(1):2861. [doi: 10.1038/s41598-018-21243-x] [Medline: 29434331]
3. Ferrari AJ, Somerville AJ, Baxter AJ, et al. Global variation in the prevalence and incidence of major depressive disorder: a systematic review of the epidemiological literature. Psychol Med 2013 Mar;43(3):471-481. [doi: 10.1017/S0033291712001511] [Medline: 22831756]
4. GBD 2013 DALYs and HALE Collaborators, Murray CJL, Barber RM, et al. Global, regional, and national disability-adjusted life years (DALYS) for 306 diseases and injuries and healthy life expectancy (HALE) for 188 countries, 1990-2013: quantifying the epidemiological transition. Lancet 2015 Nov 28;386(10009):2145-2191. [doi: 10.1016/S0140-6736(15)61340-X] [Medline: 26321261]
5. Lépine JP, Briley M. The increasing burden of depression. Neuropsychiatr Dis Treat 2011;7(Suppl 1):3-7. [doi: 10.2147/NDT.S19617] [Medline: 21750622]
6. Vrieze E, Demyttenaere K, Bruffaerts R, et al. Dimensions in major depressive disorder and their relevance for treatment outcome. J Affect Disord 2014 Feb;155:35-41. [doi: 10.1016/j.jad.2013.10.020] [Medline: 24210628]
7. Khazanov GK, Xu C, Dunn BD, Cohen ZD, DeRubeis RJ, Hollon SD. Distress and anhedonia as predictors of depression treatment outcome: a secondary analysis of a randomized clinical trial. Behav Res Ther 2020 Feb;125:103507. [doi: 10.1016/j.brat.2019.103507] [Medline: 31896529]
8. Mojtabai R, Jorm AF. Trends in psychological distress, depressive episodes and mental health treatment-seeking in the United States: 2001-2012. J Affect Disord 2015 Mar 15;174:556-561. [doi: 10.1016/j.jad.2014.12.039] [Medline: 25556674]
9. Baxter AJ, Scott KM, Ferrari AJ, Norman RE, Vos T, Whiteford HA. Challenging the myth of an “epidemic” of common mental disorders: trends in the global prevalence of anxiety and depression between 1990 and 2010. Depress Anxiety 2014 Jun;31(6):506-516. [doi: 10.1002/da.22230] [Medline: 24448889]

10. Mahmud S, Mohsin M, Dewan MN, Muyeed A. The global prevalence of depression, anxiety, stress, and insomnia among general population during COVID-19 pandemic: a systematic review and meta-analysis. Trends Psychol 2022 Jan 4;31(1):143-170. [doi: 10.1007/s43076-021-00116-9]
11. Schafer KM, Lieberman A, Sever AC, Joiner T. Prevalence rates of anxiety, depressive, and eating pathology symptoms between the pre- and peri-COVID-19 eras: a meta-analysis. J Affect Disord 2022 Feb 1;298(Pt A):364-372. [doi: 10.1016/j.jad.2021.10.115] [Medline: 34740748]
12. Bueno-Notivol J, Gracia-García P, Olaya B, Lasheras I, López-Antón R, Santabárbara J. Prevalence of depression during the COVID-19 outbreak: a meta-analysis of community-based studies. Int J Clin Health Psychol 2021;21(1):100196. [doi: 10.1016/j.ijchp.2020.07.007] [Medline: 32904715]
13. Ku BS, Li J, Lally C, Compton MT, Druss BG. Associations between mental health shortage areas and county-level suicide rates among adults aged 25 and older in the USA, 2010 to 2018. Gen Hosp Psychiatry 2021;70:44-50. [doi: 10.1016/j.genhosppsych.2021.02.001] [Medline: 33714795]
14. Liu I, Liu F, Xiao Y, Huang Y, Wu S, Ni S. Investigating the key success factors of chatbot-based positive psychology intervention with retrieval- and generative pre-trained transformer (GPT)-based chatbots. Int J Hum Computer Interaction 2024 Jan 8:1-12. [doi: 10.1080/10447318.2023.2300015]
15. Wilhelm S, Weingarden H, Ladis I, Braddick V, Shin J, Jacobson NC. Cognitive-behavioral therapy in the digital age: presidential address. Behav Ther 2020 Jan;51(1):1-14. [doi: 10.1016/j.beth.2019.08.001] [Medline: 32005328]
16. Moshe I, Terhorst Y, Philippi P, et al. Digital interventions for the treatment of depression: a meta-analytic review. Psychol Bull 2021 Aug;147(8):749-786. [doi: 10.1037/bul0000334] [Medline: 34898233]
17. Pauley D, Cuijpers P, Papola D, Miguel C, Karyotaki E. Two decades of digital interventions for anxiety disorders: a systematic review and meta-analysis of treatment effectiveness. Psychol Med 2023 Jan;53(2):567-579. [doi: 10.1017/S0033291721001999] [Medline: 34047264]
18. Cuijpers P, Reijnders M, Huibers MJH. The role of common factors in psychotherapy outcomes. Annu Rev Clin Psychol 2019 May 7;15:207-231. [doi: 10.1146/annurev-clinpsy-050718-095424] [Medline: 30550721]
19. Stein AT, Carl E, Cuijpers P, Karyotaki E, Smits JAJ. Looking beyond depression: a meta-analysis of the effect of behavioral activation on depression, anxiety, and activation. Psychol Med 2021 Jul;51(9):1491-1504. [doi: 10.1017/S0033291720000239] [Medline: 32138802]
20. Ferster CB. A functional anlysis of depression. Am Psychol 1973 Oct;28(10):857-870. [doi: 10.1037/h0035605] [Medline: 4753644]
21. Martell CR, Dimidjian S, Herman-Dunn R. Behavioral Activation for Depression: A Clinician's Guide: Guilford Publications; 2021.
22. Huguet A, Miller A, Kisely S, Rao S, Saadat N, McGrath PJ. A systematic review and meta-analysis on the efficacy of internet-delivered behavioral activation. J Affect Disord 2018 Aug 1;235:27-38. [doi: 10.1016/j.jad.2018.02.073] [Medline: 29649708]
23. Ly KH, Trüschel A, Jarl L, et al. Behavioural activation versus mindfulness-based guided self-help treatment administered through a smartphone application: a randomised controlled trial. BMJ Open 2014 Jan 9;4(1):e003440. [doi: 10.1136/bmjopen-2013-003440] [Medline: 24413342]
24. Eisma MC, Boelen PA, van den Bout J, et al. Internet-based exposure and behavioral activation for complicated grief and rumination: a randomized controlled trial. Behav Ther 2015 Nov;46(6):729-748. [doi: 10.1016/j.beth.2015.05.007] [Medline: 26520217]
25. Carlbring P, Hägglund M, Luthström A, et al. Internet-based behavioral activation and acceptance-based treatment for depression: a randomized controlled trial. J Affect Disord 2013 Jun;148(2-3):331-337. [doi: 10.1016/j.jad.2012.12.020] [Medline: 23357657]
26. Gruber J, Prinstein MJ, Clark LA, et al. Mental health and clinical psychological science in the time of COVID-19: challenges, opportunities, and a call to action. Am Psychol 2021 Apr;76(3):409-426. [doi: 10.1037/amp0000707] [Medline: 32772538]
27. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
28. Antony MM, Bieling PJ, Cox BJ, Enns MW, Swinson RP. Psychometric properties of the 42-item and 21-item versions of the Depression Anxiety Stress Scales in clinical groups and a community sample. Psychol Assess 1998;10(2):176-181. [doi: 10.1037/1040-3590.10.2.176]
29. Fossati A, Krueger RF, Markon KE, Borroni S, Maffei C. Reliability and validity of the Personality Inventory for DSM-5 (PID-5): predicting DSM-IV personality disorders and psychopathy in community-dwelling Italian adults. Assessment 2013 Dec;20(6):689-708. [doi: 10.1177/1073191113504984] [Medline: 24065702]
30. Quilty LC, Ayearst L, Chmielewski M, Pollock BG, Bagby RM. The psychometric properties of the personality inventory for DSM-5 in an APA DSM-5 field trial sample. Assessment 2013 Jun;20(3):362-369. [doi: 10.1177/1073191113486183] [Medline: 23588687]
31. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]

32. Lynch CJ, Gunning FM, Liston C. Causes and consequences of diagnostic heterogeneity in depression: paths to discovering novel biological depression subtypes. Biol Psychiatry 2020 Jul 1;88(1):83-94. [doi: 10.1016/j.biopsych.2020.01.012] [Medline: 32171465]
33. Spijker J, Bijl RV, de Graaf R, Nolen WA. Determinants of poor 1-year outcome of DSM-III-R major depression in the general population: results of the Netherlands Mental Health Survey and Incidence Study (NEMESIS). Acta Psychiatr Scand 2001 Feb;103(2):122-130. [doi: 10.1034/j.1600-0447.2001.103002122.x] [Medline: 11167315]
34. Alsayednasser B, Widnall E, O'Mahen H, et al. How well do cognitive behavioural therapy and behavioural activation for depression repair anhedonia? A secondary analysis of the COBRA randomized controlled trial. Behav Res Ther 2022 Dec;159:104185. [doi: 10.1016/j.brat.2022.104185] [Medline: 36371903]
35. Dichter GS, Felder JN, Petty C, Bizzell J, Ernst M, Smoski MJ. The effects of psychotherapy on neural responses to rewards in major depression. Biol Psychiatry 2009 Nov 1;66(9):886-897. [doi: 10.1016/j.biopsych.2009.06.021] [Medline: 19726030]
36. Borsini A, Wallis ASJ, Zunszain P, Pariante CM, Kempton MJ. Characterizing anhedonia: a systematic review of neuroimaging across the subtypes of reward processing deficits in depression. Cogn Affect Behav Neurosci 2020 Aug;20(4):816-841. [doi: 10.3758/s13415-020-00804-6] [Medline: 32472419]
37. Boswell JF, Iles BR, Gallagher MW, Farchione TJ. Behavioral activation strategies in cognitive-behavioral therapy for anxiety disorders. Psychotherapy (Chic) 2017 Sep;54(3):231-236. [doi: 10.1037/pst0000119] [Medline: 28922003]
38. Hopko DR, Robertson SMC, Lejuez CW. Behavioral activation for anxiety disorders. Behav Analyst Today 2006;7(2):212-232. [doi: 10.1037/h0100084]
39. Strine TW, Chapman DP, Kobau R, Balluz L. Associations of self-reported anxiety symptoms with health-related quality of life and health behaviors. Soc Psychiatry Psychiatr Epidemiol 2005 Jun;40(6):432-438. [doi: 10.1007/s00127-005-0914-1] [Medline: 16003592]
40. Kris-Etherton PM, Petersen KS, Hibbeln JR, et al. Nutrition and behavioral health disorders: depression and anxiety. Nutr Rev 2021 Feb 11;79(3):247-260. [doi: 10.1093/nutrit/nuaa025] [Medline: 32447382]
41. Alvaro PK, Roberts RM, Harris JK. A systematic review assessing bidirectionality between sleep disturbances, anxiety, and depression. Sleep 2013 Jul 1;36(7):1059-1068. [doi: 10.5665/sleep.2810] [Medline: 23814343]
42. Dunlop DD, Song J, Lyons JS, Manheim LM, Chang RW. Racial/ethnic differences in rates of depression among preretirement adults. Am J Public Health 2003 Nov;93(11):1945-1952. [doi: 10.2105/ajph.93.11.1945] [Medline: 14600071]
43. Hudson DL, Puterman E, Bibbins-Domingo K, Matthews KA, Adler NE. Race, life course socioeconomic position, racial discrimination, depressive symptoms and self-rated health. Soc Sci Med 2013 Nov;97:7-14. [doi: 10.1016/j.socscimed.2013.07.031] [Medline: 24161083]
44. Peplinski B, McClelland R, Szklo M. Associations between socioeconomic status markers and depressive symptoms by race and gender: results from the Multi-Ethnic Study of Atherosclerosis (MESA). Ann Epidemiol 2018 Aug;28(8):535-542. [doi: 10.1016/j.annepidem.2018.05.005] [Medline: 29934242]
45. Burkett CA. Obstructed use: reconceptualizing the mental health (help-seeking) experiences of Black Americans. J Black Psychol 2017 Feb 15;43(8):813-835. [doi: 10.1177/0095798417691381]
46. Wu Y, Levis B, Riehm KE, et al. Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis. Psychol Med 2020 Jun;50(8):1368-1380. [doi: 10.1017/S0033291719001314] [Medline: 31298180]
47. Herzog DP, Wagner S, Engelmann J, et al. Early onset of depression and treatment outcome in patients with major depressive disorder. J Psychiatr Res 2021 Jul;139:150-158. [doi: 10.1016/j.jpsychires.2021.05.048] [Medline: 34058654]
48. Furukawa TA, Cipriani A, Atkinson LZ, et al. Placebo response rates in antidepressant trials: a systematic review of published and unpublished double-blind randomised controlled studies. Lancet Psychiatry 2016 Nov;3(11):1059-1066. [doi: 10.1016/S2215-0366(16)30307-8] [Medline: 27726982]
49. Rodd JM. Moving experimental psychology online: how to obtain high quality data when we can't see our participants. J Mem Lang 2024 Feb;134:104472. [doi: 10.1016/j.jml.2023.104472]
50. Roberts SO, Bareket-Shavit C, Dollins FA, Goldie PD, Mortenson E. Racial inequality in psychological research: trends of the past and recommendations for the future. Perspect Psychol Sci 2020 Nov;15(6):1295-1309. [doi: 10.1177/1745691620927709] [Medline: 32578504]

Abbreviations

BA: behavioral activation
CBT: cognitive behavioral therapy
CONSORT: Consolidated Standards of Reporting Trials
DASS-21: 21-item Depression, Anxiety, and Stress Scale
PHQ-8: 8-item Patient Health Questionnaire
PID-5: Personality Inventory for *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition)

Edited by J Torous; submitted 02.11.23; peer-reviewed by C Simons, I Liu, N Lau; revised version received 15.05.24; accepted 09.06.24; published 30.08.24.

Please cite as:
Santopetro N, Jones D, Garron A, Meyer A, Joyner K, Hajcak G
Examining a Fully Automated Mobile-Based Behavioral Activation Intervention in Depression: Randomized Controlled Trial
JMIR Ment Health 2024;11:e54252
URL: https://mental.jmir.org/2024/1/e54252
doi:10.2196/54252

Natural Language Processing for Depression Prediction on Sina Weibo: Method Study and Analysis

Zhenwen Zhang, MEng; Jianghong Zhu, MSc; Zhihua Guo, MSc; Yu Zhang, MSc; Zepeng Li, PhD; Bin Hu, PhD

Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Lanzhou, China

Corresponding Author:
Zepeng Li, PhD

Abstract

Background: Depression represents a pressing global public health concern, impacting the physical and mental well-being of hundreds of millions worldwide. Notwithstanding advances in clinical practice, an alarming number of individuals at risk for depression continue to face significant barriers to timely diagnosis and effective treatment, thereby exacerbating a burgeoning social health crisis.

Objective: This study seeks to develop a novel online depression risk detection method using natural language processing technology to identify individuals at risk of depression on the Chinese social media platform Sina Weibo.

Methods: First, we collected approximately 527,333 posts publicly shared over 1 year from 1600 individuals with depression and 1600 individuals without depression on the Sina Weibo platform. We then developed a hierarchical transformer network for learning user-level semantic representations, which consists of 3 primary components: a word-level encoder, a post-level encoder, and a semantic aggregation encoder. The word-level encoder learns semantic embeddings from individual posts, while the post-level encoder explores features in user post sequences. The semantic aggregation encoder aggregates post sequence semantics to generate a user-level semantic representation that can be classified as depressed or nondepressed. Next, a classifier is employed to predict the risk of depression. Finally, we conducted statistical and linguistic analyses of the post content from individuals with and without depression using the Chinese Linguistic Inquiry and Word Count.

Results: We divided the original data set into training, validation, and test sets. The training set consisted of 1000 individuals with depression and 1000 individuals without depression. Similarly, each validation and test set comprised 600 users, with 300 individuals from both cohorts (depression and nondepression). Our method achieved an accuracy of 84.62%, precision of 84.43%, recall of 84.50%, and F_1-score of 84.32% on the test set without employing sampling techniques. However, by applying our proposed retrieval-based sampling strategy, we observed significant improvements in performance: an accuracy of 95.46%, precision of 95.30%, recall of 95.70%, and F_1-score of 95.43%. These outstanding results clearly demonstrate the effectiveness and superiority of our proposed depression risk detection model and retrieval-based sampling technique. This breakthrough provides new insights for large-scale depression detection through social media. Through language behavior analysis, we discovered that individuals with depression are more likely to use negation words (the value of "swear" is 0.001253). This may indicate the presence of negative emotions, rejection, doubt, disagreement, or aversion in individuals with depression. Additionally, our analysis revealed that individuals with depression tend to use negative emotional vocabulary in their expressions ("NegEmo": 0.022306; "Anx": 0.003829; "Anger": 0.004327; "Sad": 0.005740), which may reflect their internal negative emotions and psychological state. This frequent use of negative vocabulary could be a way for individuals with depression to express negative feelings toward life, themselves, or their surrounding environment.

Conclusions: The research results indicate the feasibility and effectiveness of using deep learning methods to detect the risk of depression. These findings provide insights into the potential for large-scale, automated, and noninvasive prediction of depression among online social media users.

(JMIR Ment Health 2024;11:e58259) doi:10.2196/58259

KEYWORDS

depression; social media; natural language processing; deep learning; mental health; statistical analysis; linguistic analysis; Sina Weibo; risk prediction; mood analysis

Introduction

Background

Depression is a global mental illness that is affecting the physical and mental health of an increasing number of people worldwide. In recent years, despite the World Health Organization and national governments introducing relevant policies for the diagnosis and treatment of depression, the significant challenge remains in early detection and timely treatment for a larger number of potential patients with depression [1,2]. Researchers have been exploring the potential application of clinical assessments [3,4], biological markers [5-9], and imaging techniques [10-12] in detecting depression, but there is still a lack of widely accepted and validated objective biological markers or imaging techniques for clinical diagnosis. Therefore, the diagnosis of clinical depression still heavily relies on clinical assessments and subjective symptom reports. The rapid proliferation of mobile internet technology has encouraged more individuals to share their lives and emotions on social media platforms. Meanwhile, the accumulation of vast amounts of user-generated content has sparked researchers' interest in studying the mental health of social media users within the academic community [13-16].

Challenges

Early studies primarily relied on feature-based statistical methods to learn the differences between individuals with depression and those without. Several statistical features, such as emotional words [17], language style [18], and social behavior [19] were widely used. Although these features played a crucial role in studying the differences between depressed and nondepressed groups at the time, they did not support more in-depth research and further exploration. Additionally, due to the limitations of early data collection technologies, conclusions drawn from small-scale data sets may not generalize well to larger user populations. With the rapid development of natural language processing (NLP) and deep learning, many scholars have explored applying these technologies to depression detection tasks on social media [20]. Some popular neural network models, such as convolutional neural networks (CNNs) and recurrent neural networks are widely used to encode user posts to obtain a user-level semantic representation [21-26].

Existing research treats depression detection as a long text classification task, where user posts are concatenated into a long text and then encoded through neural networks. However, these methods face several significant challenges. (1) The concatenated long text loses the fine-grained emotional information expressed in different posts and faces challenges in terms of computational speed and computing resources. (2) The existing research uses all collected user posts to train the model, which is worth discussing. Not all posts from a user express symptoms, emotions, or thoughts related to depression. (3) Previous studies have mainly focused on English social media, and the findings of these studies lack adaptability and generalizability to Chinese social media.

Contributions

To address the above challenges, we first constructed a depression detection data set based on Sina Weibo, containing 527,333 posts from 1600 users with depression and 1600 users without depression. We propose a hierarchical transformer network (HTN) model to obtain a high-quality user-level semantic representation. The model mainly consists of a 2-level transformer structure that focuses on learning semantic representations at the post level and the user level. For each user, the model first uses a transformer encoder to encode each post and obtain post-level semantic representations. Then, these post embeddings are further encoded by another transformer encoder and aggregated through a long short-term memory (LSTM) with attention to obtaining user-level semantic representations. This structure not only effectively considers the sequential evolutionary relationships of user emotional changes but also dynamically evaluates the importance of different posts. In addition, we also propose a retrieval-based post sampling strategy to mitigate the impact of noise on the model training process. Specifically, we construct a depression-related dictionary to match user posts with relevant content for model training. Experimental results demonstrate that the model and sampling strategy proposed in this paper achieve promising results on the constructed depression detection data set. This fully illustrates the sophistication and effectiveness of the proposed model and sampling strategy. Our methodology provides strong support for identifying users at risk of depression through online social media data in Chinese communities, which is important for public health and social harmony.

Our contributions can be summarized as follows: (1) we propose a hierarchical transformer-based model that can effectively capture both local and global semantic information from user posts; (2) we propose a retrieval-based post sampling strategy that effectively reduces noise in user post data and improves the quality of user-level semantic representations; and (3) we construct a depression detection data set consisting of 3200 online social media users, with over 527,333 posts collected from 1600 users with depression and 1600 users without depression.

Related Works

With the rapid growth of mobile social media, an increasing number of people are sharing their daily lives and emotional states online. As a result, researchers have become interested in using artificial intelligence technology to detect mental health issues, particularly depression, from social media data [27]. Early studies, however, were limited by small data sets and the development of NLP. These studies primarily focused on detecting depression using feature-based statistical methods, examining features such as emotional words, social engagement, and language style [19,22,28]. Researchers also explored the use of depressive-related expressions on Twitter, finding that individuals with depression tend to use more negative language in their online posts compared to those without depression [23]. Additionally, they established an evaluation task using NLP to identify individuals with depression and posttraumatic stress disorder on social media by building a data set of approximately

1800 individuals from Twitter [24]. Furthermore, they investigated the linguistic disparities between individuals with and individuals without depression by analyzing discussions of depression-related topics on social media platforms.

With the advent of deep learning and neural network technologies, there has been a significant breakthrough in detecting depression through social media. These technologies have enhanced feature extraction capabilities, allowing for the automatic capture of complex semantic information from user-generated content. They excel in semantic understanding and sentiment analysis, particularly in accurately identifying users' emotional states using attention mechanisms and recurrent neural networks [10]. They used machine learning methods to analyze photos from 166 Instagram users, suggesting that color analysis, metadata components, and algorithmic facial detection may serve as effective markers for detecting depression in photos [14]. They built a depression data set based on Reddit self-reported depression diagnosis(RSDD) and suggested using CNN to learn embedded representations for each post [17]. They proposed an integrated multi-classifier depression detection method, revealing the effectiveness of ensemble learning on depression detection tasks [26], constructed a depression detection data set based on Twitter, and proposed a method that integrates multiple semantic representations to detect depressive individuals [29]. They introduced a collaborative representation model based on reinforcement learning, which automatically selects depression-related posts and images from user-generated data to enhance depression detection performance [30]. They proposed an attention-based feature fusion model, which achieved good predictive performance on small-scale data sets [31], and a multimodal depression recognition framework that combines deep convolutional networks (DCNNs) and deep neural networks (DNNs). DCNNs are used to learn local feature representations for each modality, while DNNs integrate various features for final prediction [32]. They integrated tweet and user behavioral features, encoding user tweets using a hierarchical attention network [33], and investigated the depression classification capability of 3 bidirectional encoder representation from transformer (BERT) variants and 4 combinations of BERT variants on the text responses to 12 clinical interview questions. They found that ensemble methods could improve both F_1-scores and robustness [34] and proposed a multimodal fusion method for depression detection, where BERT is used to obtain the sentence representation and LSTM and CNN are employed to capture the representation of speech.

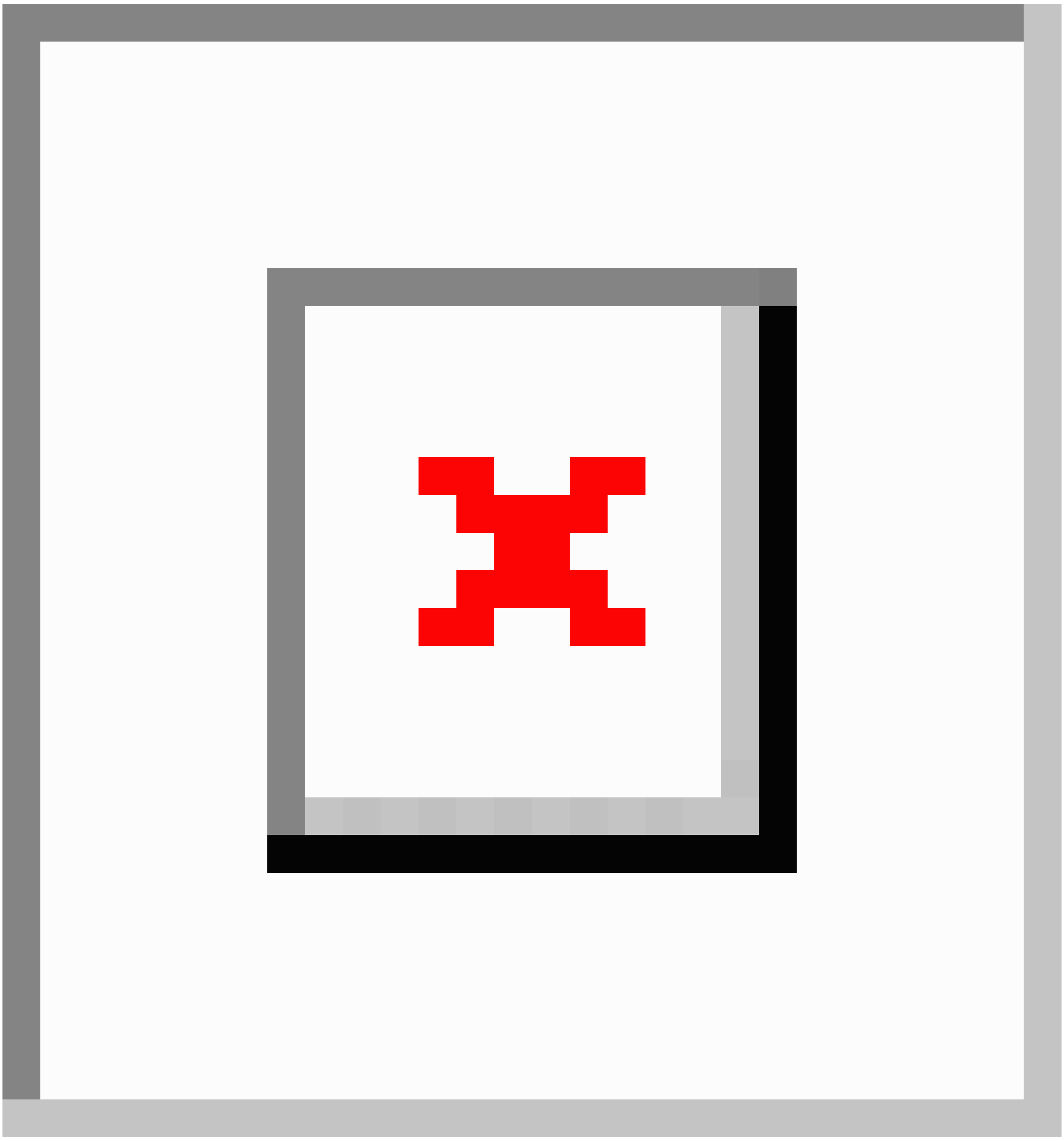
Although previous studies have explored the detection of depression using social media from the perspectives of features and encoding models and achieved significant results, there are still some issues that need to be further investigated [35-37]. User-level depression detection faces 2 key issues. First is the design of neural semantic encoders that balance performance and computational speed. Second is the quality control of user posts. Specifically, previous work has treated the classification of users with depression as a long-text classification task. User posts are concatenated into long text for encoding, which not only loses the emotional or sentiment information expressed in different posts but also creates a text length that is difficult to adapt to models like Transformer [38] and BERT [39]. It is worth noting that, despite BERT's remarkable performance improvement in many NLP tasks, it relies on pretraining knowledge from large-scale general domains. However, this general domain knowledge does not match well with the specific domain knowledge of depression. Additionally, more computational resources are strongly required in scenarios based on the BERT model. Therefore, this poses greater challenges for applying BERT to user sequence modeling.

Methods

Data Collection and Annotation

Figure 1 illustrates the workflow of constructing a user-level depression detection data set based on the Sina Weibo platform, which includes 3 steps: data collection, data preprocessing, and model training. In the following sections, we provide a detailed explanation and description of these steps.

Figure 1. The workflow for data set construction and model learning. Eval set: evaluation set; Train set: training set.



Step 1: Data Collection

Overview

We followed the annotation guidelines [29,32] proposed in studies on English social media for depression. If a user self-reported in their post that they were diagnosed with depression, then we annotated the user as depressed. For nondepressed users, if the posts they published did not clearly reveal symptoms or keywords related to depression, we annotated them as nondepression users (normal users). The detailed process is given in the sections below.

Search With Keywords

We employed 2 methods for retrieval. One method involved directly searching for “depression” as a keyword on Weibo. The other method involved using keywords such as “depression,” “symptoms,” and “medication names” within the depression supertopic on Sina Weibo.

Get Candidate Posts

We manually selected posts from individuals genuinely experiencing depression and removed posts related to popular science.

Get Candidate User ID

We obtained user IDs of candidate posts through the Weibo platform's field parsing system.

Crawl User's Posts by User ID

We used web crawling technology to scrape posts published by users on the Sina Weibo platform within a specific time period.

Cross-Annotation

Three annotators cross-annotated users based on the scraped posts, labeling them as depression or nondepression. When the decisions of the 3 annotators were consistent, we considered the user as valid and included them in either the depression or nondepression group. The determination principle for depression users was as follows: if a user voluntarily reported being diagnosed with depression in their posts, we labeled them as having depression. Additionally, we also considered expressions in the Chinese context, such as mentioning medication or suicidal thoughts. The determination principle for nondepression users was that their posts did not explicitly contain expressions related to depression.

Step 2: Data Preprocessing

The raw data collected from Sina Weibo often contains irrelevant or informal expressions, which can negatively impact the model's performance. To address this issue, we processed the raw data using the following steps: (1) user-identifiable information was removed to protect user privacy; (2) each post was segmented into a word sequence using the Jieba tokenizer for efficient processing; (3) emoticons were replaced with their corresponding emotion words for more accurate analysis; (4) numbers, URLs, and punctuation were eliminated from the posts to reduce noise; (5) automatically generated posts by Sina Weibo's robot assistant, such as birthday reminders and membership-level notifications, were filtered out; (6) duplicate posts were removed to ensure data uniqueness; and (7) posts consisting of fewer than 3 words were excluded from training to maintain quality standards.

Step 3: Model Training

Since the data set we constructed was balanced, we divided the 1600 depression and nondepression users into training, validation, and testing sets, with 1000 users for training, 300 for validation, and 300 for testing. Therefore, a total of 2000 users were used for training, 600 users for validation, and 600 users for testing.

Ethical Considerations

All data in this study were obtained from publicly shared information on Sina Weibo, and any personal information that could potentially expose user privacy was excluded from the study. Therefore, this analysis applied the standards for waiving informed consent and similar guidelines [40]. In addition, our research complied with the requirements of the Sina Weibo platform regarding the use of user data. We ensured that our study did not involve infringement of user privacy or ethical issues. Specifically, we desensitized and anonymized the collected user data, removing any information that could potentially indicate user identities during the preprocessing stage. Furthermore, since this study used a limited data set of Sina Weibo users for modeling and analysis, these conclusions may not fully generalize to all depression and nondepression users on Sina Weibo. The predictive outcomes of the model should be considered as suggested conclusions and not be regarded as definitive decisions in the real world.

Problem Definition

This study aimed to develop a depression risk prediction model using NLP and user-generated data from social media. The input to this model was each user's posts, and the output was a label indicating whether the user is depressed or not.

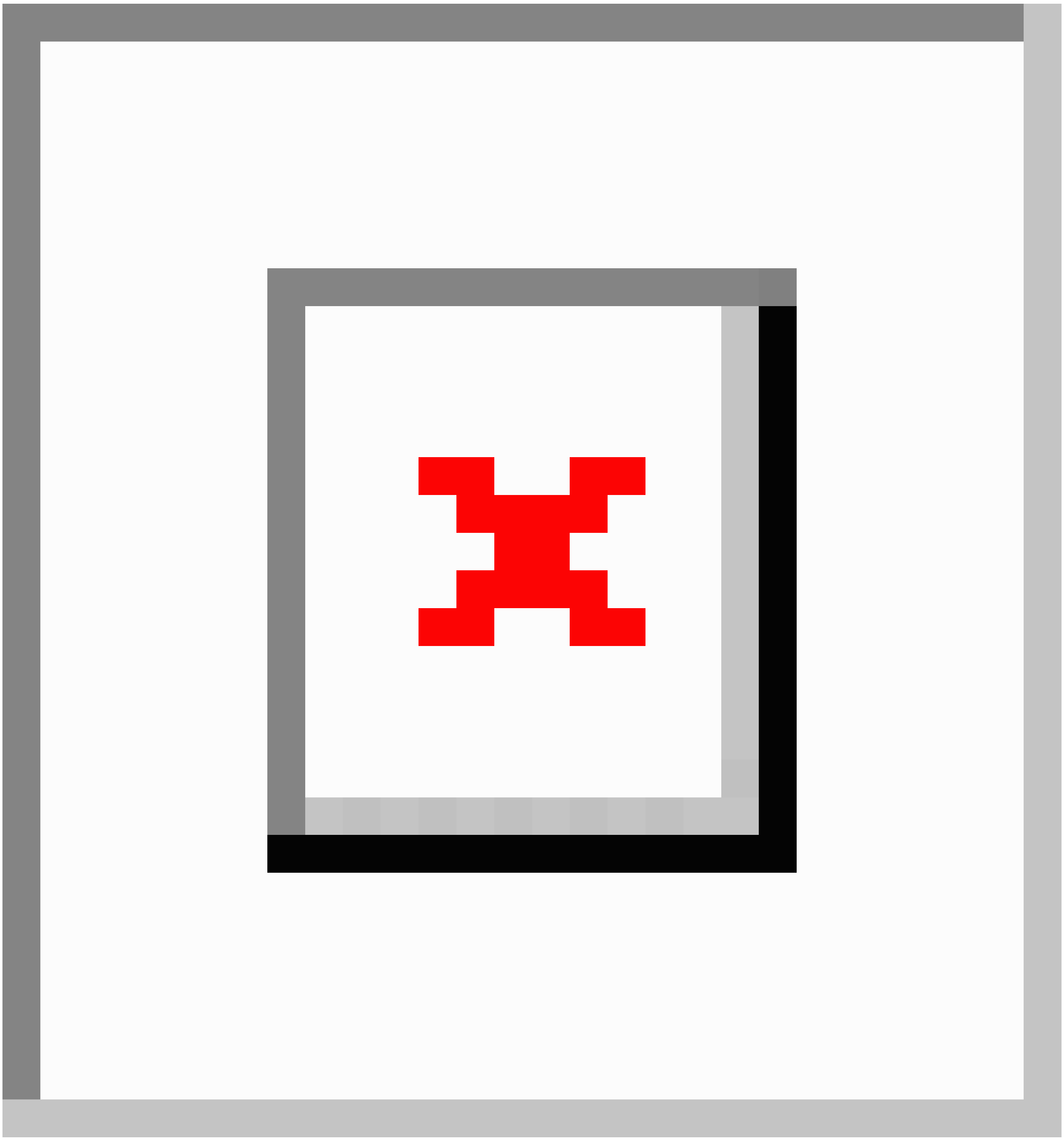
Proposed Model

Overview

Figure 2 illustrates the workflow of our proposed depression detection model, which consists of 5 steps: word embedding, post embedding, user embedding, classification, model training, and evaluation. We provide detailed insights into the development and training in the following sections.

As shown in Figure 2, we propose an HTN to study textual semantic features from users' posts. The Transformer is an attention-based neural network architecture that has gained considerable attention in recent years, particularly in NLP and computer vision. Unlike other deep learning models, the Transformer not only dynamically captures long-term dependencies but also exhibits faster computation speed. Inspired by this, we incorporated the Transformer into our model to better understand and encode behavior and intention from user posts. Our model consists of 2 levels of transformers: a word-level transformer and a post-level transformer. The word-level transformer is used to compute semantic features for each post, with word embeddings from each post as input. The sentence-level transformer is employed to calculate aggregated semantic features for all user posts, with the input being the embeddings of all user posts. After obtaining the aggregated global feature representation, we performed classification on it to predict whether the user is depressed. Since our prediction task is a binary classification task, we used a sigmoid function for prediction. The proposed model is capable of learning fine-grained feature representations at the levels of words, sentences, and documents from user posts, which is crucial for enhancing prediction accuracy.

Figure 2. The architecture of our proposed depression prediction model. FFN: feedforward neural network; LSTM-Attention: long short-term memory with attention.



Word Embedding

To obtain better word embeddings, we used Tencent's pretrained word embeddings [41] (Tencent AI Lab Embedding Corpus for Chinese Words and Phrases) to initialize the embedding representations of each word in user posts. This embedding corpus was pretrained on Wikipedia, Baidu Baike, and web text data using the Directional Skip-Gram algorithm, and it includes embeddings for 12,287,936 Chinese words. Specifically, we first employed the vocabulary from the Tencent pretrained word embedding database as external vocabulary for tokenizing each user post with the Jieba tokenizer. Then, we retrieved the embedding for each word in the user posts from the Tencent pretrained word embedding database and input them into the model for further training.

Post Embedding

After obtaining the pretrained word embeddings, we added positional encodings to each word in the posts and combined them with the word embeddings. These new embeddings were then fed into the first-level transformer encoder for encoding, where each transformer encoder consists of a multi-head self-attention mechanism and a feed-forward neural network. The self-attention mechanism allows each word to interact with other words in the sequence, while the feed-forward neural network applies independent nonlinear transformations to each word. Each sublayer uses residual connections and layer normalization to stabilize the training process. After processing through multiple layers, the contextual representation of each word is obtained, with the representation of the [CLS] token being used as the final embedding representation of the post.

User Embedding

As described above, we employed a shared transformer encoder to obtain semantic embeddings for each post. To effectively merge these post embeddings, we employed another transformer encoder along with an LSTM network equipped with an attention mechanism for deeper semantic feature extraction and aggregation of each user's posts. Specifically, the embeddings of user posts are sequentially input into the transformer encoder in the order of their posting time for deep feature extraction. Subsequently, the semantic context obtained from the transformer encoder is processed by an attention-based LSTM structure to extract and aggregate sequential information. The advantage of this model architecture is that it not only learns more effective deep semantic contextual representations but also dynamically considers the importance of different posts.

Classification

We focused on predicting whether a user is at risk of depression, thus a binary classification process was applied to the user embeddings.

Model Training and Evaluation

We divided the raw data into 3 sets: the training set, validation set, and test set. The training set consists of 1000 depressed and 1000 nondepressed users, the validation set consists of 300 depressed and 300 nondepressed users, and the test set consists of 300 depressed and 300 nondepressed users. All models were implemented using the PyTorch [42] framework on a graphics processing unit (GPU) server equipped with 2 Tesla A100 cards. For the CNN model, the convolutional kernel size was set to {2, 3, 4}, and the number of filters was set to 100. For other baselines, both the hidden size and attention size were set to 256. For our proposed model, each post was padded or truncated to 512 words. The learning rate was set to 1e-3, and the batch size was optimized from the range of {32, 64, 128}.

Comparison Baselines

To comprehensively evaluate the potential of applying deep learning for predicting depression risk on social media, we adopted 11 widely used neural network models as baselines. These included CNN, LSTM, gated recurrent unit (GRU), bidirectional GRU, and bidirectional LSTM models and attention-based methods like LSTM with attention, GRU with attention, bidirectional LSTM with attention, and bidirectional GRU with attention, BERT, and a hierarchical convolutional network model.

Evaluation Metrics

We used accuracy, macroaveraged precision, macroaveraged recall, and macroaveraged F1-score to evaluate the models presented in this study. These metrics are widely used to assess the performance of deep learning–based models.

Results

Performance Comparison

Table 1 presents the experimental results of the baseline models and our proposed model on the test set. We observed that our proposed model achieves over 80% accuracy in predicting depression risk across all scenarios. Compared with neural models without the attention mechanism, attention-based neural models demonstrate better detection performance across all sampling strategies, with particularly significant improvements observed when using the no-sampling strategy. We attribute this improvement to the attention mechanism's ability to automatically focus more on words or phrases indicative of depression, thereby facilitating a superior semantic representation of the user. The HTN model outperforms the other baseline models, with at least a 2% improvement in the retrieval strategy and more than a 5% improvement in the other conditions. This suggests that encoding a user's post data with HTN is more effective than treating it as a single long text. HTN enables the model to fully consider post interactions and intuitively fit better with human thinking. Simply treating all of a user's posts as a single long text may lead to computational and gradient challenges, limiting the model's ability to detect depression.

Table . Overall performance comparison of our proposed model and baseline models. Without: results without applying any sampling strategy; Random: results of applying random sampling strategy to sample 50% of posts; Retrieval: results based on retrieval sampling strategy.

Model		Accuracy	Precision	Recall	F_1-score
CNN [a]					
	Without	79.93	80.70	80.79	79.93
	Random	78.37	78.30	78.63	78.29
	Retrieval	93.53	93.21	93.54	93.30
LSTM [b]					
	Without	71.80	73.71	69.91	69.91
	Random	69.55	69.35	68.52	68.65
	Retrieval	88.41	88.40	88.10	88.23
GRU [c]					
	Without	78.55	78.98	77.55	77.86
	Random	77.68	78.03	76.69	76.98
	Retrieval	92.25	92.09	92.49	92.21
BiGRU [d]					
	Without	67.99	67.77	67.94	67.81
	Random	70.24	70.02	69.30	69.43
	Retrieval	91.52	91.35	91.63	91.46
BiLSTM [e]					
	Without	65.92	65.88	66.06	65.81
	Random	65.05	65.19	65.36	64.99
	Retrieval	84.95	84.95	85.35	84.90
LSTM-attention [f]					
	Without	78.55	78.34	78.15	78.23
	Random	74.05	73.77	73.90	73.82
	Retrieval	91.87	91.72	92.13	91.82
GRU-attention [g]					
	Without	82.53	82.43	82.12	82.24
	Random	80.62	80.39	80.43	80.41
	Retrieval	91.27	91.16	91.34	91.15
BiLSTM-attention [h]					
	Without	78.03	77.94	77.42	77.59
	Random	74.39	74.20	73.72	73.87
	Retrieval	91.35	91.58	90.05	91.19
BiGRU-attention [i]					
	Without	80.97	80.75	80.97	80.83
	Random	76.64	76.72	77.03	76.59
	Retrieval	92.77	92.68	92.88	92.64
BERT [j]					
	Without	81.44	80.37	80.52	80.11
	Random	79.92	78.42	78.66	78.21

Model		Accuracy	Precision	Recall	F_1-score
	Retrieval	90.21	89.48	88.71	89.05
HCN [k]					
	Without	83.33	83.19	83.84	83.41
	Random	78.62	80.66	79.39	79.77
	Retrieval	93.53	93.34	94.02	93.40
HTN [l]					
	Without	84.62	84.43	84.50	84.32
	Random	82.43	82.24	82.44	82.35
	Retrieval	95.46	95.30	95.70	95.43

[a]CNN: convolutional neural network.
[b]LSTM: long short-term memory.
[c]GRU: gated recurrent unit.
[d]BiGRU: bidirectional gated recurrent unit.
[e]BiLSTM: bidirectional long short-term memory.
[f]LSTM-attention: long short-term memory with attention.
[g]GRU-attention: gated recurrent unit with attention.
[h]BiLSTM-attention: bidirectional long short-term memory with attention.
[i]BiGRU-attention: bidirectional gated recurrent unit with attention.
[j]BERT: bidirectional encoder representation from transformer.
[k]HCN: hierarchical convolutional network.
[l]HTN: hierarchical transformer network; (best performing model).

Effectiveness of Sampling Strategy

Figure 3 illustrates the comparison of model performance before and after applying our proposed retrieval-based sampling strategy. After applying the retrieval-based sampling strategy, the proposed model's depression risk prediction accuracy exceeds 95%. These fully highlight the necessity and importance of sampling user posts. Through sampling, the computational overhead of model training can be effectively reduced, allowing the model to focus more on learning about depression. In addition, we also noticed that the random sampling strategy performed worse than the no-sampling strategy, likely due to the inherent uncertainty in the random sampling process.

Figure 4 illustrates the F_1-scores of each model under various sampling strategies and sampling ratios. It is evident that the application of effective sampling strategies can significantly enhance the depression detection capabilities of the models. Conversely, in random sampling experiments, achieving performance beyond that of the full data set (sampling rate of 1.0) is challenging when the sampling rate is less than 1.0. By employing a retrieval-based sampling strategy to select posts relevant to depression, not only is the computational complexity of the model reduced, but the model also gains a better focus on acquiring knowledge related to depression from user posts. We observed that the retrieval-based sampling strategy consistently demonstrated a stable upward trend as the sampling rate increased incrementally, unlike the random sampling strategy, which exhibited more pronounced fluctuations. We attribute this primarily to the fact that the retrieval-based sampling strategy ensures the selection of posts related to depression in each sampling iteration. Conversely, the post selection process in the random sampling strategy is probabilistic and does not guarantee the relevance of a user's post to depression in each selection.

Figure 3. Performance comparison between applying retrieval-based sampling strategy and not applying any sampling strategy. BERT: bidirectional encoder representation from transformer; BiGRU: bidirectional gated recurrent unit; BiGRU-attention: bidirectional gated recurrent unit with attention; BiLSTM: bidirectional long short-term memory; BiLSTM-attention: bidirectional long short-term memory with attention; CNN: convolutional neural network; GRU: gated recurrent unit; GRU-Attention: gated recurrent unit with attention; HCN: hierarchical convolutional network; HTN: hierarchical transformer network; LSTM: long short-term memory; LSTM-Attention: long short-term memory with attention.

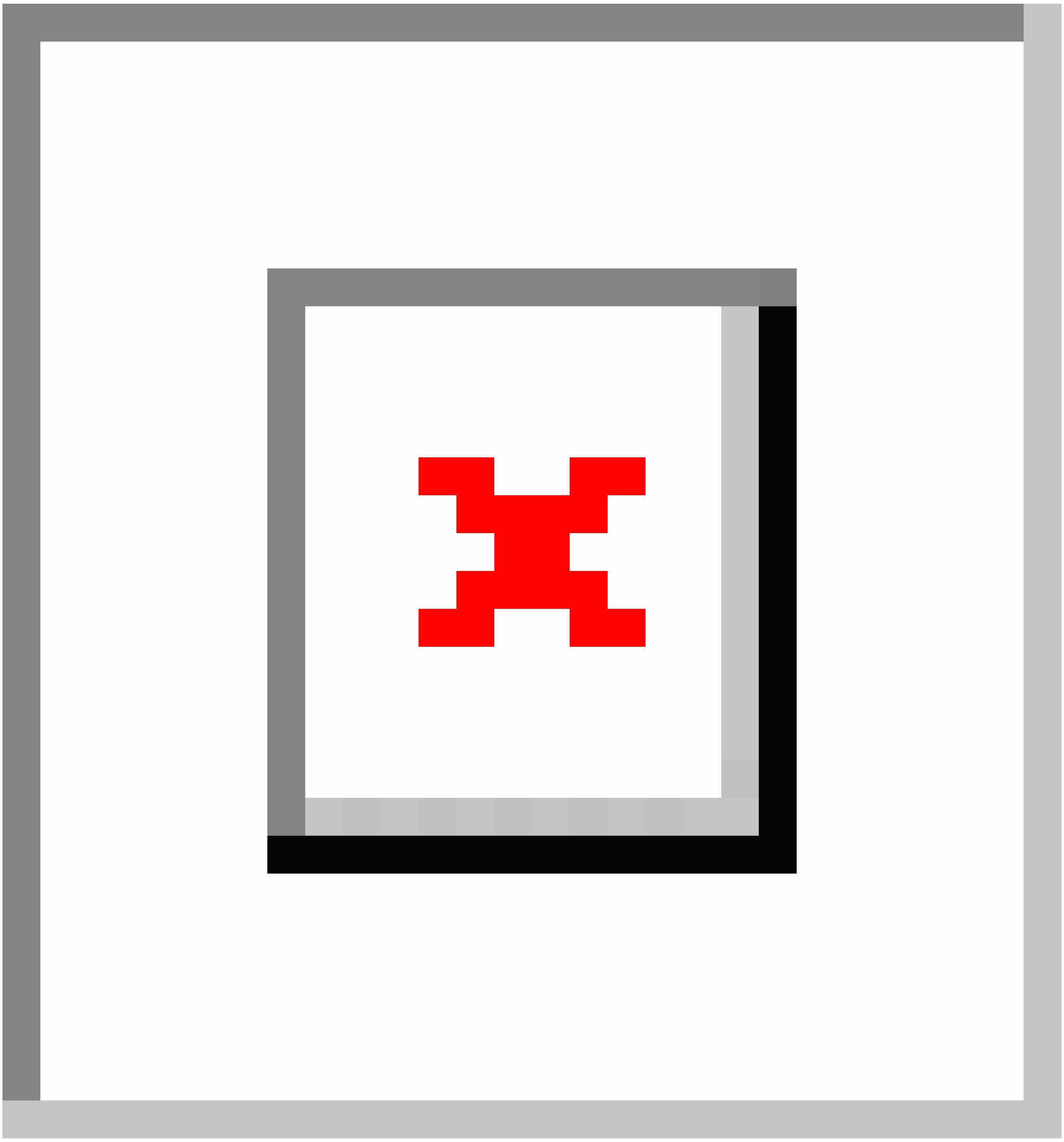
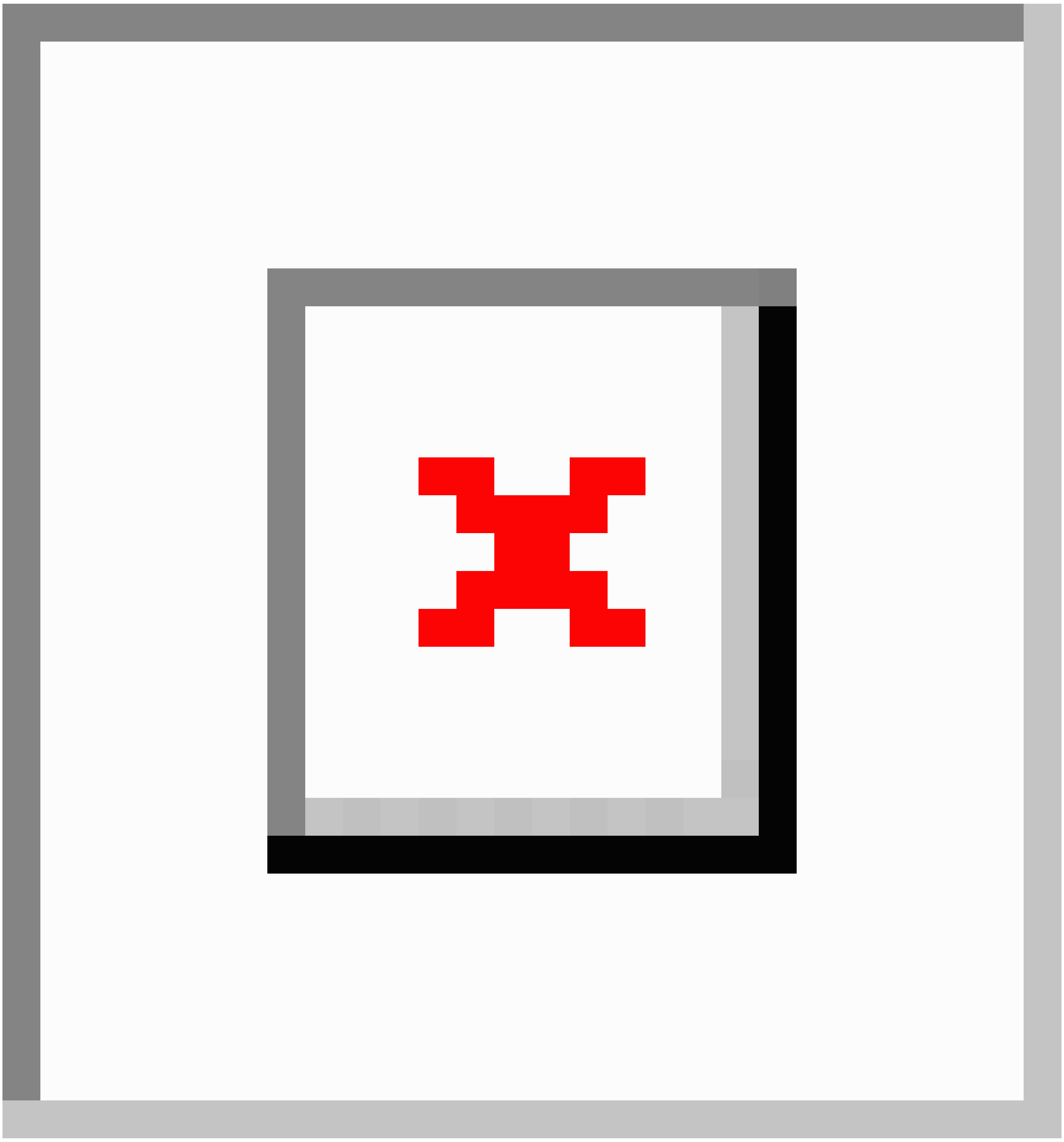


Figure 4. Comparison of model performance results with different sampling strategies and sampling ratios. BERT: bidirectional encoder representation from transformer; BiGRU: bidirectional gated recurrent unit; BiGRU-attention: bidirectional gated recurrent unit with attention; BiLSTM: bidirectional long short-term memory; BiLSTM-attention: bidirectional long short-term memory with attention; CNN: convolutional neural network; GRU: gated recurrent unit; GRU-Attention: gated recurrent unit with attention; HCN: hierarchical convolutional network; HTN: hierarchical transformer network; LSTM: long short-term memory; LSTM-Attention: long short-term memory with attention.



Linguistic and Behavior Analysis

Figure 5 compares the common differences in social behaviors between depressed and nondepressed users. We can observe that, compared with nondepressed users, depressed users have fewer posts and lower posting frequency, reflecting the less active social engagement of depressed users. In terms of pronoun use, depressed users tend to use the first-person singular (我) more frequently in their posts, while nondepressed users use the first-person plural (我们) more often. This suggests that depressed users may be more self-focused and have less interaction with others, whereas nondepressed users are more group-oriented and engage in more interactive behaviors. Additionally, depressed users are more likely to focus on depression-related topics on social media, such as discussing their condition, treatment processes, and medication, while nondepressed users mention and discuss these topics less frequently.

Figure 6 presents the comparative results of modal particle use between depressed users and nondepressed users. We can observe that the use of “的” (de) is more frequent in both depressed and nondepressed users, while “呢” (ne) is used the least frequently. The main reason is that “的” is commonly used as a modifier in almost all sentences, whereas “呢” and “吗” are often used in contexts expressing questions or uncertainties. It is worth noting that “吧” (ba) is used more frequently in the language expressions of users with depression, while “啊” (a) is used more frequently in the language expressions of nondepressed users. These 2 words are typically used at the end of sentences, with “吧” often used to modify completed events, while “啊” is typically used to modify events that are about to happen. In the expressions of users with depression, “吧” is more often expressed as “好吧” (“okay”), “行吧” (“all right”), “就这样吧” (“just like this”), “去死吧” (“go die,”), etc. On the other hand, “啊” is often combined in expressions of nondepressed users as “真开心啊” (“really happy”), “原来是这样啊” (“so that’s how it is”), and “你对我真好啊” (“you’re really good to me”).

Figure 7 illustrates the comparative results of punctuation use between depressed users and nondepressed users. We discovered that depressed users tend to use periods more frequently than nondepressed users, while nondepressed users prefer commas over those with depression. We speculate that this trend may stem from the fact that depressed users often experience low moods and slowed thinking, which could manifest in more cautious and negative expressions. A period can signify a conclusion or a clear break between ideas, possibly reflecting the psychological inclination of these individuals to conclude or avoid further communication. In contrast, nondepressed users typically exhibit active and divergent thinking patterns. They frequently employ commas to separate sentence components and convey incomplete thought processes.

Additionally, we observed that nondepressed users are more inclined to use exclamation marks (“!”), which aligns with the experimental results regarding the interjection “啊” (“a”) presented in Figure 6. Furthermore, depressed users tend to use the tilde (“~”) and ellipses more frequently. These symbols are commonly employed in the Chinese internet context to convey a sense of helplessness or resignation.

We used the Chinese Linguistic Inquiry and Word Count (LIWC) dictionary [43] to analyze the differences in language use between users with depression and nondepressed users, and Figure 8 presents the comparative results. Figure 8 reveals that users with depression are more likely to use negative vocabulary, such as “Swear,” “Affect,” “PosEmo,” “NegEmo,” “Anx,” “Anger,” “Sad,” etc, than nondepressed users. Depressed users appeared to favor discussing past and present events (“PastM,” “PresentM”), whereas nondepressed users appeared to focus more on possible future events (“FutureM”). We speculated that this difference might be attributed to the significant influence of their family of origin on many depressed users, leading them to reflect more on the impact of past events in their posts. Furthermore, we observed that depressed users exhibited relatively more negative than nondepressed users when discussing topics related to “Social,” “Family,” “Friends,” and “Home.” Additionally, we found that words such as “Bio,” “Body,” “Health,” “Death,” and “Psychology” were more frequently used in the posts of depressed users. The main reason for this is that posts by depressed users may express their intentions related to suicide or self-harm, or they may involve sharing experiences and discussions about the condition among fellow patients, encompassing the diagnosis process, physical condition, and medication.

Figure 5. Comparison of the social behaviors between depressed and nondepressed users. “Word/Post”: the average number of words per post; “Post/User”: the average number of posts per user; “Post/User/Week”: the average number of posts per user per week; “1stPerSing/Post”: the frequency of the first-person singular (我) used per post; “1stPerPlural/Post”: the frequency of the first-person singular (我们) used per post; “depression/Post”: the frequency of the keywords (抑郁症, 抑郁) used per post; “Drugs/Post”: the frequency of mentioning depression medication–related terms per post.

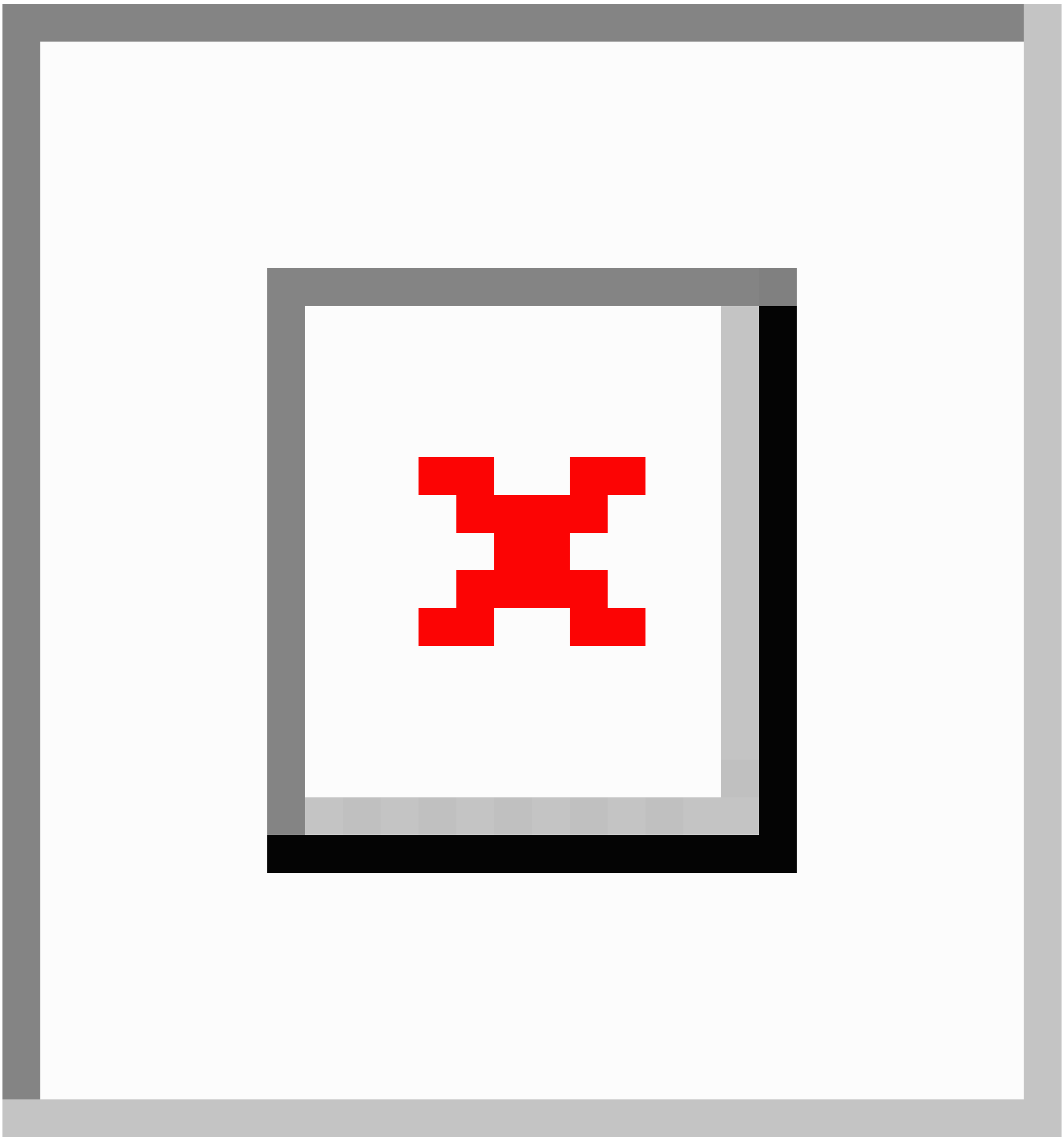


Figure 6. Comparison of the modal particle use between depressed and nondepressed users.

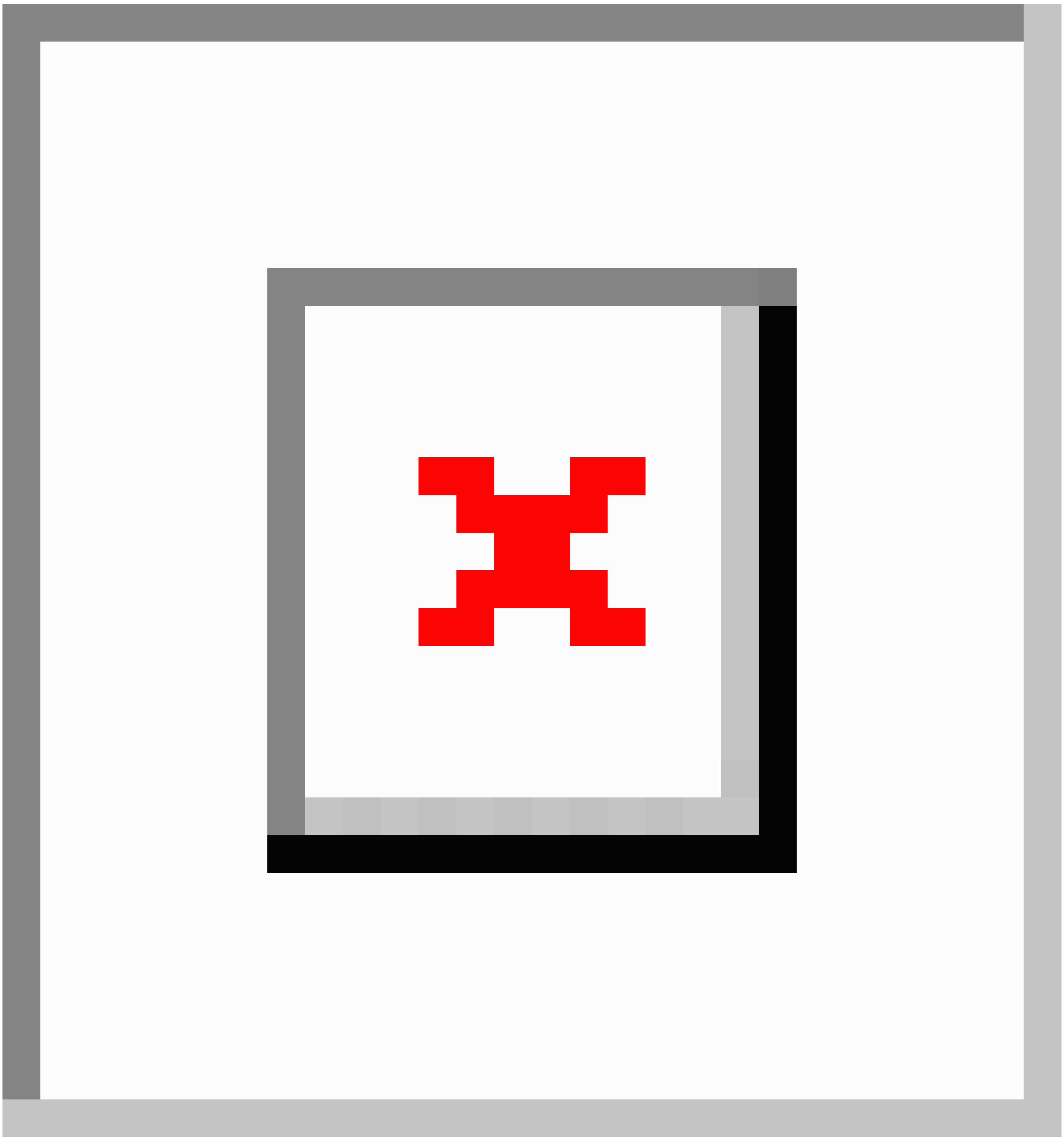


Figure 7. Comparison of the punctuation use between depressed and nondepressed users.

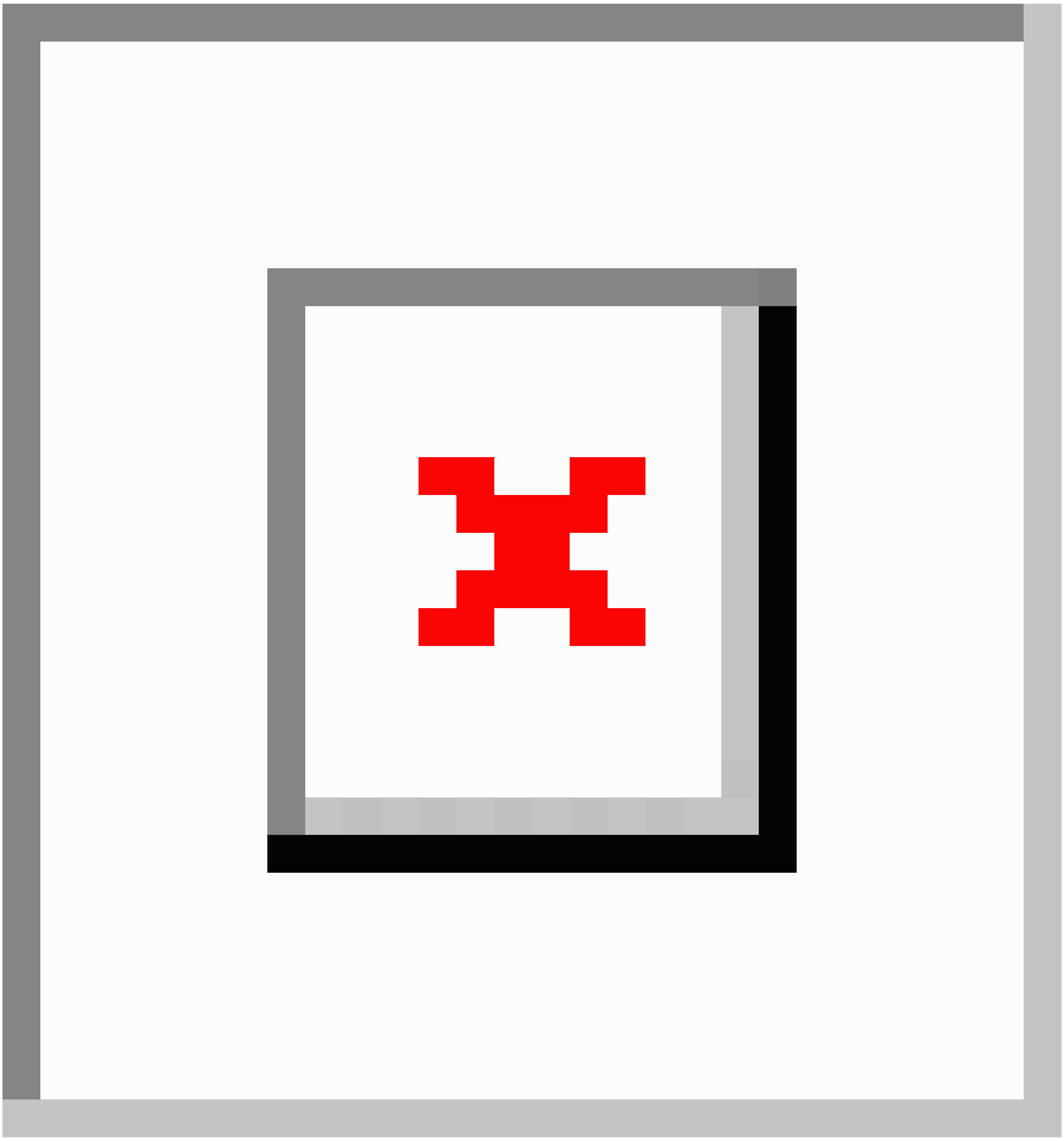
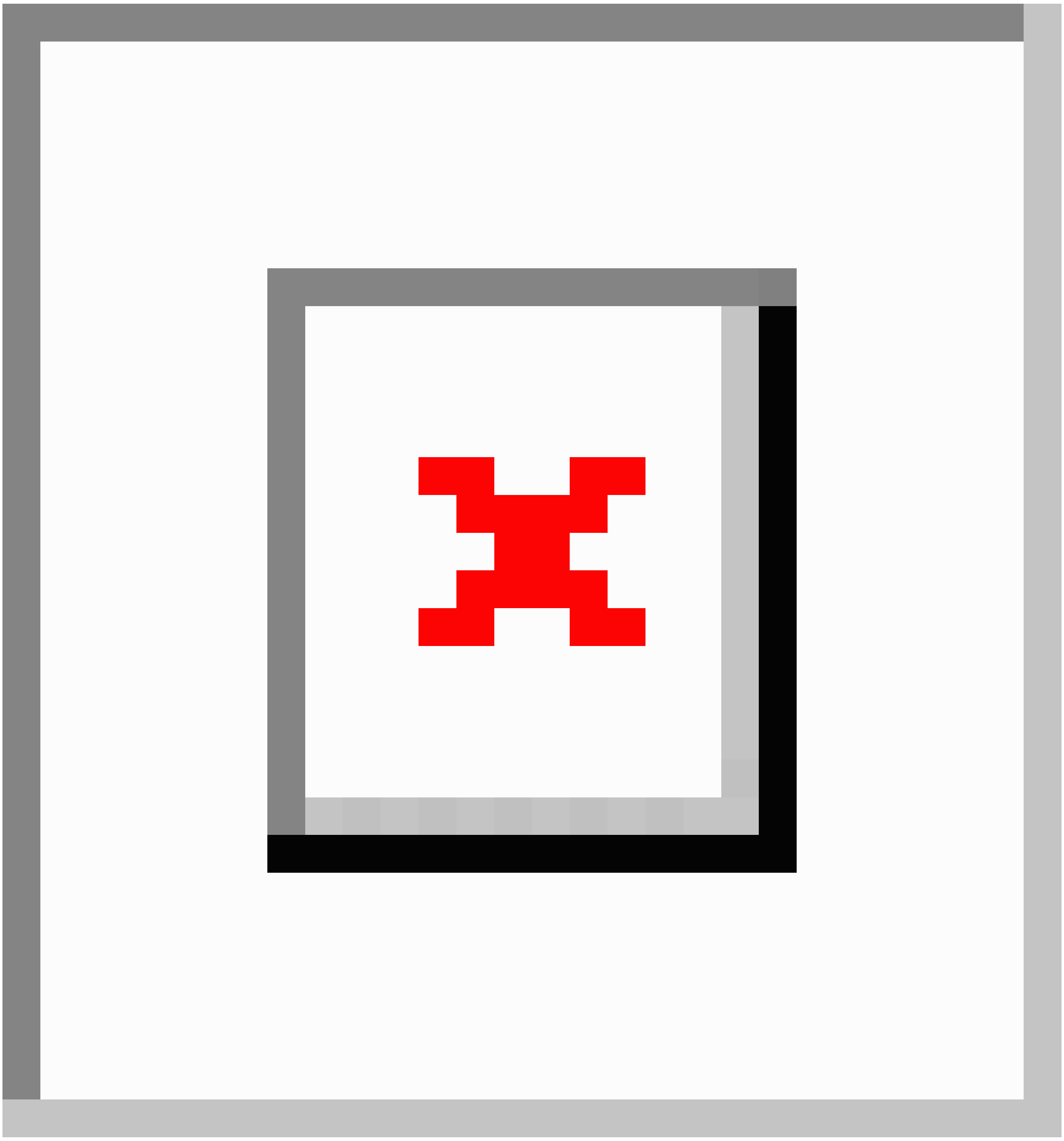


Figure 8. Comparison of significant Linguistic Inquiry and Word Count features between depressed and nondepressed users.



Discussion

Principal Results

This paper explores the automatic prediction of depression risk among users on online social media using deep learning methods. We developed and validated the model on a large-scale data set of online social media users. The research findings indicate that the proposed model exhibits significant advantages in predicting depression risk, confirming the effectiveness and advanced capabilities of deep learning for depression risk prediction. The paper carries several implications.

With the rapid development of social media technology, more and more young people are using social media to share their emotions and document their lives. Social media has become a vital platform for them to express emotions, seek support, and build social connections. However, mental health issues among young people are increasingly prominent, making them a key societal concern. Social media serves as a vital tool for them to communicate their feelings and connect with others. However, it also poses a challenge in effectively using social media data to identify and support individuals who may be facing mental health issues. More and more individuals with mental health problems, especially depression, do not actively seek help from professionals. This leads to a lack of timely treatment and support, causing them to miss optimal intervention opportunities. Furthermore, there is a growing shortage of clinical psychologists to meet the increasing mental health needs of the population. Therefore, exploring automated depression risk identification technologies based on artificial intelligence, particularly deep learning, has become a crucial and essential research topic in addressing the current societal challenges.

Furthermore, this study developed an HTN and proposed a retrieval-enhanced post sampling technique to improve the performance of depression risk detection. Experimental results indicate that our developed approach outperforms all baseline methods, achieving prediction accuracies and F_1-scores of 84% across 3 independent experiments. With the application of the retrieval sampling technique, the performance of almost all methods reaches nearly 90%. Compared with methods without sampling, there is a performance improvement of over 10% across all 4 metrics. This strongly demonstrates the effectiveness and advanced capabilities of our approach in predicting the risk of depression.

Finally, linguistic analysis revealed that depressed users exhibit more conservative and reserved social behaviors on social media compared with nondepressed users. Not only do they make fewer posts, but their posts are also shorter. This may reflect their negativity in social interactions and a tendency to avoid social engagement. Reduced social engagement could result from the loneliness, frustration, or lack of motivation commonly felt by individuals with depression. Additionally, depressed users express more negative emotions in their posts. Through linguistic sentiment analysis, we found that posts by depressed users contain more negative sentiment words, a difference more pronounced than in nondepressed users. This further highlights the psychological distress and negative emotional experiences that individuals with depression may encounter on social media. These traits offer insights into the behaviors of depressed users, providing direction for developing more accurate and personalized depression risk prediction models.

Limitations

Although our research has achieved some promising results, there are still some limitations. These limitations mainly focus on the 3 aspects given below.

Research Data

This study relies on a subset of users from the Chinese social media platform Sina Weibo, which may not fully represent the Chinese population or all users of Chinese social media. Considering the individual differences among users, the research model and results of this study may not accurately assess the depression risk of internet users. Additionally, the findings of this study may not be generalizable to users of other social media platforms or populations with different medical conditions.

Chinese LIWC

A notable limitation is that the existing Chinese LIWC dictionary covers a limited vocabulary. It may not fully capture all the emotional and semantic nuances in the texts of depressed and nondepressed users, especially as language and culture evolve and new expressions emerge, which the dictionary might not update to include in a timely manner. Another limitation is that LIWC mainly analyzes based on word frequency and lacks contextual understanding. It cannot discern the different meanings of polysemous words in various contexts, nor can it handle complex grammatical structures and sentence-level emotional expressions. Additionally, LIWC focuses on surface-level vocabulary analysis and lacks the ability to comprehend deep semantics and implied meanings. It cannot effectively handle sarcasm, metaphors, and complex emotional expressions. Therefore, when using LIWC for text analysis, we should combine it with other methods and tools to obtain more comprehensive and accurate results. We also need to remain critical of LIWC's output and consider its limitations when interpreting research conclusions.

Large Language Model

Although large language models demonstrate powerful capabilities in semantic representation, we did not explore this in our paper. Our main concerns regarding this are as follows. First, they demand high computational resources, including a large number of GPU or tensor processing unit resources as well as significant storage space. Second, due to their large number of parameters, they require longer training times, which may incur substantial time and cost. Additionally, the complexity of large language models poses a risk of overfitting, necessitating additional regularization and tuning. Furthermore, large language models have poor interpretability, making it difficult to understand and explain their internal structure and decision-making processes. Last, large language models require a large amount of training data, which may raise concerns about the use and protection of user privacy data, necessitating additional data management and security measures.

Conclusions

In this study, we explored using deep learning techniques to predict depression risk based on social media data. We collected posts from 3200 online social media users over a 1-year period in order to develop and validate a depression risk detection model. The proposed HTN demonstrated exceptional performance on the collected data, yielding a predictive accuracy of over 95% across 4 commonly employed evaluation metrics. Furthermore, we introduced a retrieval-based post sampling technique, which significantly improved our model's ability to detect the risk of depression. This research provides technical support for the automatic identification of users at risk of depression on Chinese online social media, thereby effectively supporting online platforms in engaging in societal risk management.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (number 2021YFF1201200); the Sci-Tech Innovation 2030-Major Project of Brain Science and Brain-Inspired Intelligence Technology (number 2021ZD0202003); the National Natural Science Foundation of China (62227807 and 62072219); the Fundamental Research Funds for the Central Universities (lzujbky-2023-10); and by the Supercomputing Center of Lanzhou University.

Conflicts of Interest

None declared.

References

1. Evans-Lacko S, Aguilar-Gaxiola S, Al-Hamzawi A, et al. Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: results from the WHO World Mental Health (WMH) surveys. Psychol Med 2018 Jul;48(9):1560-1571. [doi: 10.1017/S0033291717003336] [Medline: 29173244]
2. Fox AB, Smith BN, Vogt D. How and when does mental illness stigma impact treatment seeking? longitudinal examination of relationships between anticipated and internalized stigma, symptom severity, and mental health service use. Psychiatry Res 2018 Oct;268:15-20. [doi: 10.1016/j.psychres.2018.06.036] [Medline: 29986172]
3. Cosco TD, Lachance CC, Blodgett JM, et al. Latent structure of the Centre for Epidemiologic Studies Depression Scale (CES-D) in older adult populations: a systematic review. Aging Ment Health 2020 May;24(5):700-704. [doi: 10.1080/13607863.2019.1566434] [Medline: 30661386]
4. Richter P, Werner J, Heerlein A, Kraus A, Sauer H. On the validity of the Beck Depression Inventory. Psychopathology 1998 Feb;31(3):160-168. [doi: 10.1159/000066239] [Medline: 9636945]
5. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
6. Zimmerman M, Martinez JH, Young D, Chelminski I, Dalrymple K. Severity classification on the Hamilton Depression Rating Scale. J Affect Disord 2013 Sep 5;150(2):384-388. [doi: 10.1016/j.jad.2013.04.028] [Medline: 23759278]
7. Yang M, Wu Y, Tao Y, Hu X, Hu B. Trial Selection Tensor Canonical Correlation Analysis (TSTCCA) for depression recognition with facial expression and pupil diameter. IEEE J Biomed Health Inform 2023 Oct 5;PP. [doi: 10.1109/JBHI.2023.3322271] [Medline: 37796673]
8. Tao Y, Yang M, Li H, Wu Y, Hu B. DepMSTAT: multimodal spatio-temporal attentional transformer for depression detection. IEEE Trans Knowl Data Eng 2024 Jul;36(7):2956-2966. [doi: 10.1109/TKDE.2024.3350071]
9. Shao W, You Z, Liang L, et al. A multi-modal gait analysis-based detection system of the risk of depression. IEEE J Biomed Health Inform 2022 Oct;26(10):4859-4868. [doi: 10.1109/JBHI.2021.3122299] [Medline: 34699374]
10. Reece AG, Danforth CM. Instagram photos reveal predictive markers of depression. EPJ Data Sci 2017 Dec;6(1). [doi: 10.1140/epjds/s13688-017-0110-z]
11. Zheng Y, Chen X, Li D, et al. Treatment-naïve first episode depression classification based on high-order brain functional network. J Affect Disord 2019 Sep 1;256:33-41. [doi: 10.1016/j.jad.2019.05.067] [Medline: 31158714]
12. Masuccio FG, Gamberini G, Calabrese M, Solaro C. Imaging and depression in multiple sclerosis: a historical perspective. Neurol Sci 2021 Mar;42(3):835-845. [doi: 10.1007/s10072-020-04951-z] [Medline: 33411192]
13. Zhu J, Li Z, Zhang X, Zhang Z, Hu B. Public attitudes toward anxiety disorder on Sina Weibo: content analysis. J Med Internet Res 2023 Apr 4;25:e45777. [doi: 10.2196/45777] [Medline: 37014691]
14. De Choudhury M, De S. Mental health discourse on reddit: self-disclosure, social support, and anonymity. Presented at: Eighth International AAAI Conference on Weblogs and Social Media; Jul 1-4, 2014; Ann Arbor, Michigan p. 71-80. [doi: 10.1609/icwsm.v8i1.14526]
15. Naslund JA, Aschbrenner KA, Marsch LA, Bartels SJ. The future of mental health care: peer-to-peer support and social media. Epidemiol Psychiatr Sci 2016 Apr;25(2):113-122. [doi: 10.1017/S2045796015001067] [Medline: 26744309]
16. Zhang Z, Li Z, Zhu J, Guo Z, Shi B, Hu B. Enhancing user sequence representation with cross-view collaborative learning for depression detection on Sina Weibo. Knowl Based Syst 2024 Jun;293:111650. [doi: 10.1016/j.knosys.2024.111650]
17. Guo Z, Ding N, Zhai M, Zhang Z, Li Z. Leveraging domain knowledge to improve depression detection on chinese social media. IEEE Trans Comput Soc Syst 2023 Aug;10(4):1528-1536. [doi: 10.1109/TCSS.2023.3267183]
18. Guntuku SC, Klinger EV, McCalpin HJ, Ungar LH, Asch DA, Merchant RM. Social media language of healthcare super-utilizers. NPJ Digit Med 2021 Mar 25;4(1):55. [doi: 10.1038/s41746-021-00419-2] [Medline: 33767336]
19. Xu R, Zhang Q. Understanding online health groups for depression: social network and linguistic perspectives. J Med Internet Res 2016 Mar 10;18(3):e63. [doi: 10.2196/jmir.5042] [Medline: 26966078]
20. Le Glaz A, Haralambous Y, Kim-Dufor DH, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res 2021 May 4;23(5):e15708. [doi: 10.2196/15708] [Medline: 33944788]
21. Squarcina L, Villa FM, Nobile M, Grisan E, Brambilla P. Deep learning for the prediction of treatment response in depression. J Affect Disord 2021 Feb 15;281:618-622. [doi: 10.1016/j.jad.2020.11.104] [Medline: 33248809]

22. Akbari M, Hu X, Liqiang N, Chua TS. From tweets to wellness: wellness event detection from twitter streams. Presented at: Thirtieth AAAI Conference on Artificial Intelligence; Feb 12-17, 2016; Phoenix, Arizona. [doi: 10.1609/aaai.v30i1.9975]
23. Coppersmith G, Dredze M, Harman C, Hollingshead K, Mitchell M. CLPsych 2015 shared task: depression and PTSD on Twitter. Presented at: 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality; Jun 5, 2015; Denver, Colorado p. 31-39 URL: http://aclweb.org/anthology/W15-12 [accessed 2024-08-27] [doi: 10.3115/v1/W15-1204]
24. Hswen Y, Naslund JA, Brownstein JS, Hawkins JB. Online communication about depression and anxiety among Twitter users with schizophrenia: preliminary findings to inform a digital phenotype using social media. Psychiatr Q 2018 Sep;89(3):569-580. [doi: 10.1007/s11126-017-9559-y] [Medline: 29327218]
25. Coppersmith G, Dredze M, Harman C. Quantifying mental health signals in Twitter. Presented at: Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality; Jun 27, 2014; Baltimore, Maryland, USA p. 51-60 URL: http://aclweb.org/anthology/W14-32 [accessed 2024-08-27] [doi: 10.3115/v1/W14-3207]
26. Shen G, Jia J, Nie L, et al. Depression detection via harvesting social media: a multimodal dictionary learning solution. Presented at: Twenty-Sixth International Joint Conference on Artificial Intelligence; Aug 19-25, 2017; Melbourne, Australia p. 3838-3844 URL: https://www.ijcai.org/proceedings/2017 [accessed 2024-08-27] [doi: 10.24963/ijcai.2017/536]
27. Yates A, Cohan A, Goharian N. Depression and self-harm risk assessment in online forums. Presented at: 2017 Conference on Empirical Methods in Natural Language Processing; Sep 7-11, 2017; Copenhagen, Denmark URL: http://aclweb.org/anthology/D17-1 [accessed 2024-08-27] [doi: 10.18653/v1/D17-1322]
28. Park M, McDonald D, Cha M. Perception differences between the depressed and non-depressed users in Twitter. Presented at: Seventh International AAAI Conference on Weblogs and Social Media; Jul 8-11, 2013; Cambridge, Massachusetts p. 476-485. [doi: 10.1609/icwsm.v7i1.14425]
29. Gui T, Zhu L, Zhang Q, et al. Cooperative multimodal approach to depression detection in Twitter. Presented at: The Thirty-Third AAAI Conference on Artificial Intelligence, The Thirty-First Conference on Innovative Applications of Artificial Intelligence, The Ninth Symposium on Educational Advances in Artificial Intelligence; Jan 27 to Feb 1, 2019; Honolulu, Hawaii p. 110-117. [doi: 10.1609/aaai.v33i01.3301110]
30. Song H, You J, Chung JW, Park JC. Feature attention network: interpretable depression detection from social media. Presented at: 32nd Pacific Asia Conference on Language, Information and Computation; Dec 1-3, 2018; Hong Kong, China URL: https://aclanthology.org/Y18-1070.pdf [accessed 2024-08-27]
31. Yang L, Jiang D, Han W, Sahli H. DCNN and DNN based multi-modal depression recognition. Presented at: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII); Oct 23-26, 2017; San Antonio, TX p. 484-489. [doi: 10.1109/ACII.2017.8273643]
32. Zogan H, Razzak I, Jameel S, Xu G. Hierarchical Convolutional Attention Network for depression detection on social media and its impact during pandemic. IEEE J Biomed Health Inform 2023 Feb 9. [doi: 10.1109/JBHI.2023.3243249] [Medline: 37022816]
33. Senn S, Tlachac ML, Flores R, Rundensteiner E. Ensembles of BERT for depression classification. Annu Int Conf IEEE Eng Med Biol Soc 2022 Jul;2022:4691-4694. [doi: 10.1109/EMBC48229.2022.9871120] [Medline: 36085764]
34. Makiuchi MR, Warnita T, Uto K, Shinoda K. Multimodal fusion of BERT-CNN and gated CNN representations for depression detection. Presented at: The 9th International on Audio/Visual Emotion Challenge and Workshop; Oct 21, 2019; Nice, France p. 55-63 URL: https://dl.acm.org/doi/proceedings/10.1145/3347320 [accessed 2024-08-27] [doi: 10.1145/3347320.3357694]
35. Ahmed A, Aziz S, Toro CT, et al. Machine learning models to detect anxiety and depression through social media: a scoping review. Comput Methods Programs Biomed Update 2022 Sep;2:100066. [doi: 10.1016/j.cmpbup.2022.100066] [Medline: 36105318]
36. Zogan H, Razzak I, Jameel S, Xu G. DepressionNet: learning multi-modalities with user post summarization for depression detection on social media. Presented at: 44th International ACM SIGIR Conference on Research and Development in Information Retrieval; Jul 11-14, 2021; Virtual Event Canada p. 133-142 URL: https://dl.acm.org/doi/proceedings/10.1145/3404835 [accessed 2024-08-27] [doi: 10.1145/3404835.3462938]
37. Zogan H, Razzak I, Wang X, Jameel S, Xu G. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. World Wide Web 2022;25(1):281-304. [doi: 10.1007/s11280-021-00992-2] [Medline: 35106059]
38. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Adv Neural Inf Process Syst 2017 Dec:6000-6010. [doi: 10.5555/3295222.3295349]
39. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Jun 2-7, 2019; Minneapolis, Minnesota URL: http://aclweb.org/anthology/N19-1 [accessed 2024-08-27] [doi: 10.18653/v1/N19-1423]
40. Whitehead LC. Methodological and ethical issues in internet-mediated research in the field of health: an integrated review of the literature. Soc Sci Med 2007 Aug;65(4):782-791. [doi: 10.1016/j.socscimed.2007.03.005] [Medline: 17512105]

41. Song Y, Shi S, Li J, Zhang H. Directional skip-gram: explicitly distinguishing left and right context for word embeddings. Presented at: 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Jun 1-6, 2018; New Orleans, Louisiana URL: http://aclweb.org/anthology/N18-2 [accessed 2024-08-27] [doi: 10.18653/v1/N18-2028]
42. Paszke A, Gross S, Massa F, et al. Pytorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst 2019 Dec:8026-8037. [doi: 105555/34542873455008]
43. Zeng X, Yang C, Tu C, Liu Z, Sun M. Chinese LIWC lexicon expansion via hierarchical classification of word embeddings with sememe attention. Presented at: Thirty-Second AAAI Conference on Artificial Intelligence; Feb 2-7, 2018; New Orleans, Louisiana. [doi: 10.1609/aaai.v32i1.11982]

Abbreviations

BERT: bidirectional encoder representation from transformer
CNN: convolutional neural network
DCNN: deep convolutional network
DNN: deep neural network
GPU: graphics processing unit
GRU: gated recurrent unit
HTN: hierarchical transformer network
LIWC: Linguistic Inquiry and Word Count
LSTM: long short-term memory
NLP: natural language processing

Edited by J Torous; submitted 11.03.24; peer-reviewed by AA wani, AP Vit, A Hosny, M Elbattah, R Upadhyay; revised version received 02.06.24; accepted 04.06.24; published 04.09.24.

Please cite as:
Zhang Z, Zhu J, Guo Z, Zhang Y, Li Z, Hu B
Natural Language Processing for Depression Prediction on Sina Weibo: Method Study and Analysis
JMIR Ment Health 2024;11:e58259
URL: https://mental.jmir.org/2024/1/e58259
doi:10.2196/58259

Breaking Down Barriers to a Suicide Prevention Helpline: Web-Based Randomized Controlled Trial

Margot C A Van der Burgt[1,2], MSc; Saskia Mérelle[1], PhD; Willem-Paul Brinkman[3], PhD; Aartjan T F Beekman[2], MD, PhD; Renske Gilissen[1], PhD

1

2

3

Corresponding Author:
Margot C A Van der Burgt, MSc

Abstract

Background: Every month, around 3800 people complete an anonymous self-test for suicidal thoughts on the website of the Dutch suicide prevention helpline. Although 70% score high on the severity of suicidal thoughts, <10% navigate to the web page about contacting the helpline.

Objective: This study aimed to test the effectiveness of a brief barrier reduction intervention (BRI) in motivating people with severe suicidal thoughts to contact the suicide prevention helpline, specifically in high-risk groups such as men and middle-aged people.

Methods: We conducted a fully automated, web-based, randomized controlled trial. Respondents with severe suicidal thoughts and little motivation to contact the helpline were randomly allocated either to a brief BRI, in which they received a short, tailored message based on their self-reported barrier to the helpline (n=610), or a general advisory text (care as usual as the control group: n=612). Effectiveness was evaluated using both behavioral and attitudinal measurements. The primary outcome measure was the use of a direct link to contact the helpline after completing the intervention or control condition. Secondary outcomes were the self-reported likelihood of contacting the helpline and satisfaction with the received self-test.

Results: In total, 2124 website visitors completed the Suicidal Ideation Attributes Scale and the demographic questions in the entry screening questionnaire. Among them, 1222 were randomized into the intervention or control group. Eventually, 772 respondents completed the randomized controlled trial (intervention group: n=369; control group: n=403). The most selected barrier in both groups was "I don't think that my problems are serious enough." At the end of the trial, 33.1% (n=122) of the respondents in the intervention group used the direct link to the helpline. This was not significantly different from the respondents in the control group (144/403, 35.7%; odds ratio 0.87, 95% CI 0.64 - 1.18, *P*=.38). However, the respondents who received the BRI did score higher on their self-reported likelihood of contacting the helpline at a later point in time (B=0.22, 95% CI 0.12 - 0.32, *P*≤.001) and on satisfaction with the self-test (B=0.27, 95% CI 0.01 - 0.53, *P*=.04). For male and middle-aged respondents specifically, the results were comparable to that of the whole group.

Conclusions: This trial was the first time the helpline was able to connect with high-risk website visitors who were hesitant to contact the helpline. Although the BRI could not ensure that those respondents immediately used the direct link to the helpline at the end of the trial, it is encouraging that respondents indicated that they were more likely to contact the helpline at a later point in time. In addition, this low-cost intervention provided greater insight into the perceived barriers to service. Follow-up research should be focused on identifying the added value of other components (eg, video or photo material) in the BRI and increasing its effectiveness, especially for men and middle-aged people.

Trial Registration: ClinicalTrials.gov NCT05458830; https://clinicaltrials.gov/study/NCT05458830

International Registered Report Identifier (IRRID): RR2-10.2196/41078

(JMIR Ment Health 2024;11:e56396) doi:10.2196/56396

KEYWORDS

barrier reduction intervention; suicidal ideation; self-help; suicide prevention helpline; randomized controlled trial; help-seeking; suicide; RCT; self-test; effectiveness; prevention; middle-aged; behavioral; attitudinal; website visitors; website visitor; website; men; suicide prevention

Introduction

Due to its accessibility and anonymity, the internet is often the first place individuals turn to when seeking information on delicate subjects [1,2]. Each month, around 3800 people complete an anonymous self-test for suicidal thoughts on the website of the national suicide prevention helpline in the Netherlands—113 Suicide Prevention [3]. The organization offers 24/7 anonymous phone and chat support, a web-based self-help course, self-assessment tests, as well as brief web-based counseling and therapy. Since its foundation in 2009, brand awareness and service users have increased annually, with more than 151,000 chat and phone call conversations and almost 1.4 million website visits in 2022 [4]. Previous studies on the helpline provided an understanding of its visitors' profile, with the majority of helpline users being female and younger than 35 years old [5-7].

The self-test for suicidal thoughts consists of the Suicidal Ideation Attributes Scale (SIDAS) and informs the test-taker about the severity of their suicidal ideation by measuring the frequency and controllability of suicidal thoughts, the closeness to an attempt, and the distress and interference with everyday activities [8]. Even though the majority (70%) of test-takers score higher than the threshold for severe suicidal ideation (SIDAS≥21), very few of them (less than 10%) continue to the web page about contacting the helpline by phone or chat. Due to the anonymous nature of the helpline's services, it is not possible to determine to what extent test-takers follow the advice about contacting the helpline. However, while men make up around 40% of the self-test users, only around 20% of the helpline's chat users are male [6,7,9].

It is disheartening that so few test-takers go on to contact the helpline by phone or chat, especially given the seriousness of their suicidal thoughts. However, we do know from the literature that a large proportion of individuals with suicidal ideation struggle with seeking adequate help, especially in low-income countries [10,11]. Some known barriers to care are structural factors like time and finances; the lack of perceived need for services; a preference for self-management; fear of hospitalization; and stigmatizing attitudes toward suicide, mental health problems, and seeking professional treatment [10,12]. People may also not receive the care they need because there are not enough services available, they cannot afford the expense of care, or they believe the available services do not meet their needs [13,14]. The study of suicide is complex. Suicidal behavior occurs among vulnerable individuals in the context of a range of different mental illnesses, and social stresses and can be influenced by attitudes toward help-seeking and cultural norms [15]. It is crucial to better understand how to guide high-risk individuals toward professional help. Despite global progress that resulted in a 36% decrease in the age-standardized suicide rate between 2000 and 2019, and no significant increase during the COVID-19 pandemic, suicide remains among the leading causes of death worldwide, with an estimated 703,000 lives lost to suicide in 2019 [16,17]. Globally, over half (58%) of all suicides occurred before the age of 50 years, and the majority of them took place in low- and middle-income countries (77%). Men are at higher risk than women, with a 2.3 higher age-standardised suicide rate than women [17].

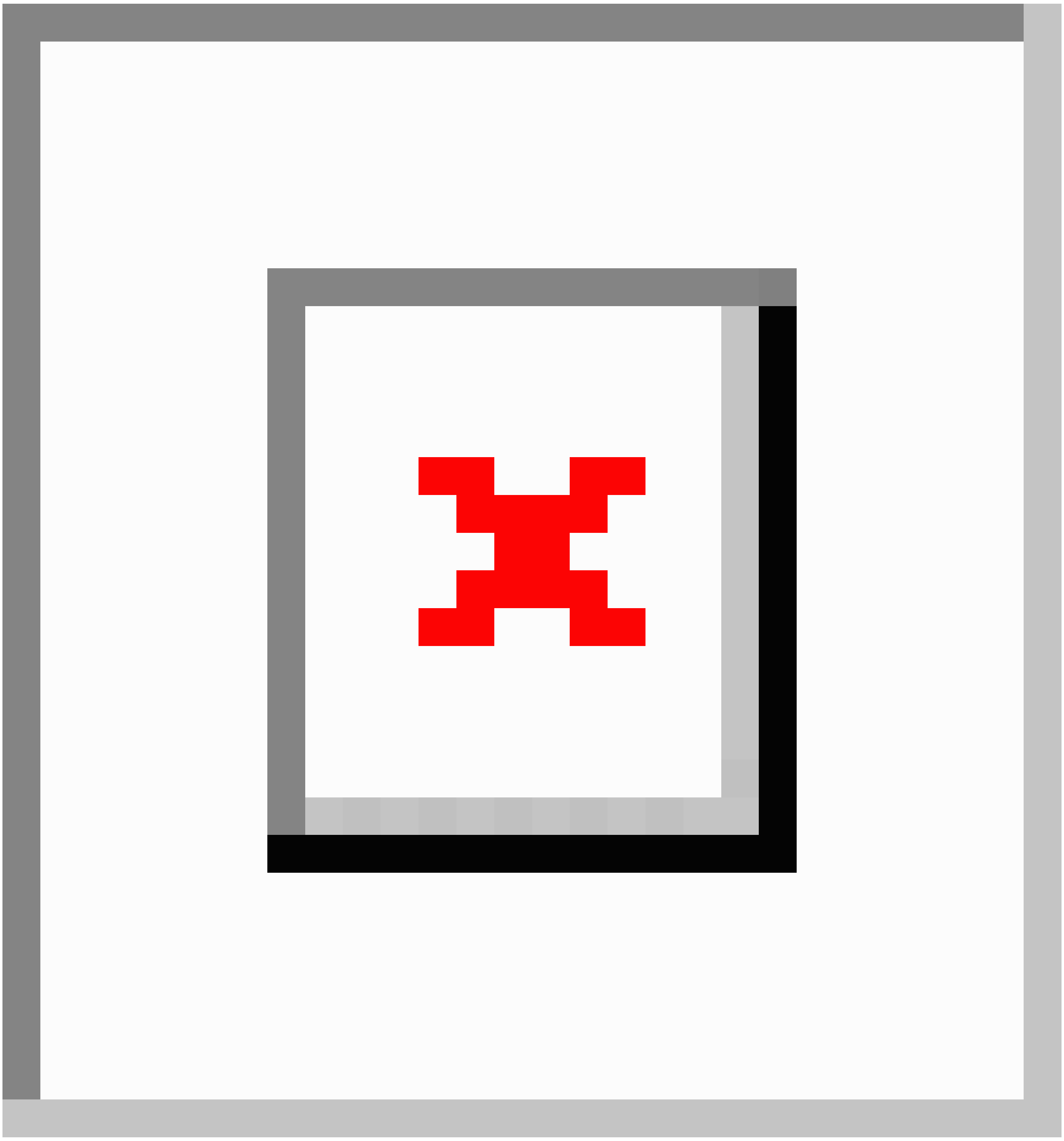
This study is focused on suicide prevention in a web-based environment. In this setting, persuasive eHealth technologies can be used to help and motivate people to reach out for help. These persuasive systems can be defined as "computerized software or information systems designed to reinforce, change or shape attitudes or behaviors or both without using coercion or deception" [18]. Little research has been done on stimulating help-seeking behavior among anonymous and high-risk internet users. Therefore, motivating reluctant high-risk individuals in a web-based environment toward professional help is still relatively uncharted territory. With this study, we intended to test whether it is possible to increase service use among individuals with severe suicidal ideation by providing more tailored information. Our study is inspired by the work of Jaroszewski et al [19], in which they evaluated a brief, automated barrier reduction intervention (BRI) designed to increase the use of crisis service referrals provided within the mental health app Koko. To the best of our knowledge, this study is the first automatic, web-based, randomized controlled trial (RCT) among people with severe suicidal ideation that aimed to reduce barriers to a suicide prevention helpline. The aim of our study is 2-fold: (1) to measure the effectiveness of a brief BRI provided in the self-test motivating people with severe suicidal thoughts to contact the Dutch suicide prevention helpline and (2) to specifically evaluate the effectiveness of the BRI in increasing helpline use by high-risk groups for suicide such as men and middle-aged people (40 - 70 years) [20].

Methods

Study Design

This study was designed as an automated, web-based, 2-arm RCT. Respondents with severe suicidal thoughts and little to no interest in contacting the helpline by phone or chat were randomly assigned to either a brief BRI or received a general advisory text (care as usual). It was intended that the intervention could be finished in less than 10 minutes to minimize the burden on our high-risk and sensitive study population. More detailed information about the study's methodology can be found in the study protocol [21]. Figure 1 displays the study's flowchart.

Figure 1. Flowchart. BRI: barrier reduction intervention; CSQ: Client Satisfaction Questionnaire; RCT: randomized controlled trial. SIDAS: Suicidal Ideation Attributes Scale.



Participants

Participants were recruited between October 7 and December 5, 2022. Anonymous visitors of the self-test for suicidal thoughts on the website of the Dutch national suicide prevention helpline were asked if they wanted to help improve the current self-test by contributing to this study. Visitors of the website could select “Yes, I will participate in the study” or “No, I just want to fill in the self-test.” People who were willing to contribute to our study were redirected to an information page and were asked for their informed consent. Those who were not willing to participate were guided to the already existing self-test on the website.

Exclusion Criteria

Participants were excluded from the study if: they were younger than 16 years old, scored below the cutoff point for severe suicidal thoughts (SIDAS score <21), or scored above the cutoff point for severe suicidal thoughts (SIDAS score≥21) and reported being likely to contact the suicide prevention helpline. They were directly transferred to the contact details of the helpline. Respondents who did not meet the requirements for inclusion were redirected to a web page thanking them for their time and encouraging them to contact the helpline in case of distress.

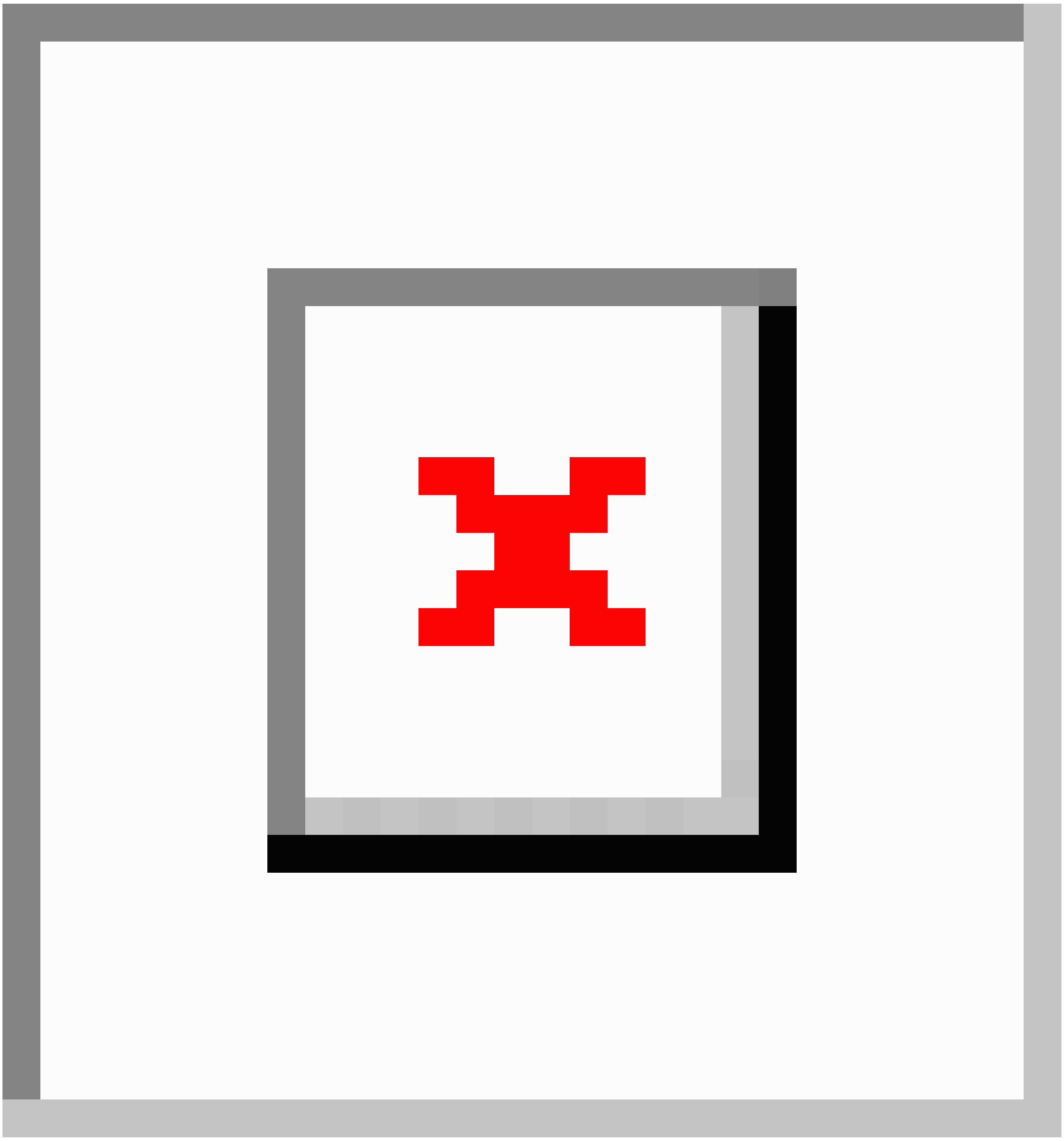
Assessment of Barriers

Both respondents in the intervention and control condition were asked the following question: “Could you indicate why you might not want to talk to one of our counselors at the moment?” Respondents could choose one of the following options: (1) “I don’t think that 113 can help me”; (2) “I’m scared to talk about my feelings”; (3) “I don’t think that my problems are serious enough”; (4) “I’m scared that people will find out”; (5) “I would rather solve it myself”; and the remaining option (6) “I have other reasons.” Answer options were based on a pilot study, more information about the pilot study can be found in the study protocol [21].

Control Group: Care as Usual

After the barrier question, the control group received a general advisory text. The text was similar to the advisory text individuals receive when they fill in the current self-test on the website of the helpline. Figure 2 shows a translated screenshot.

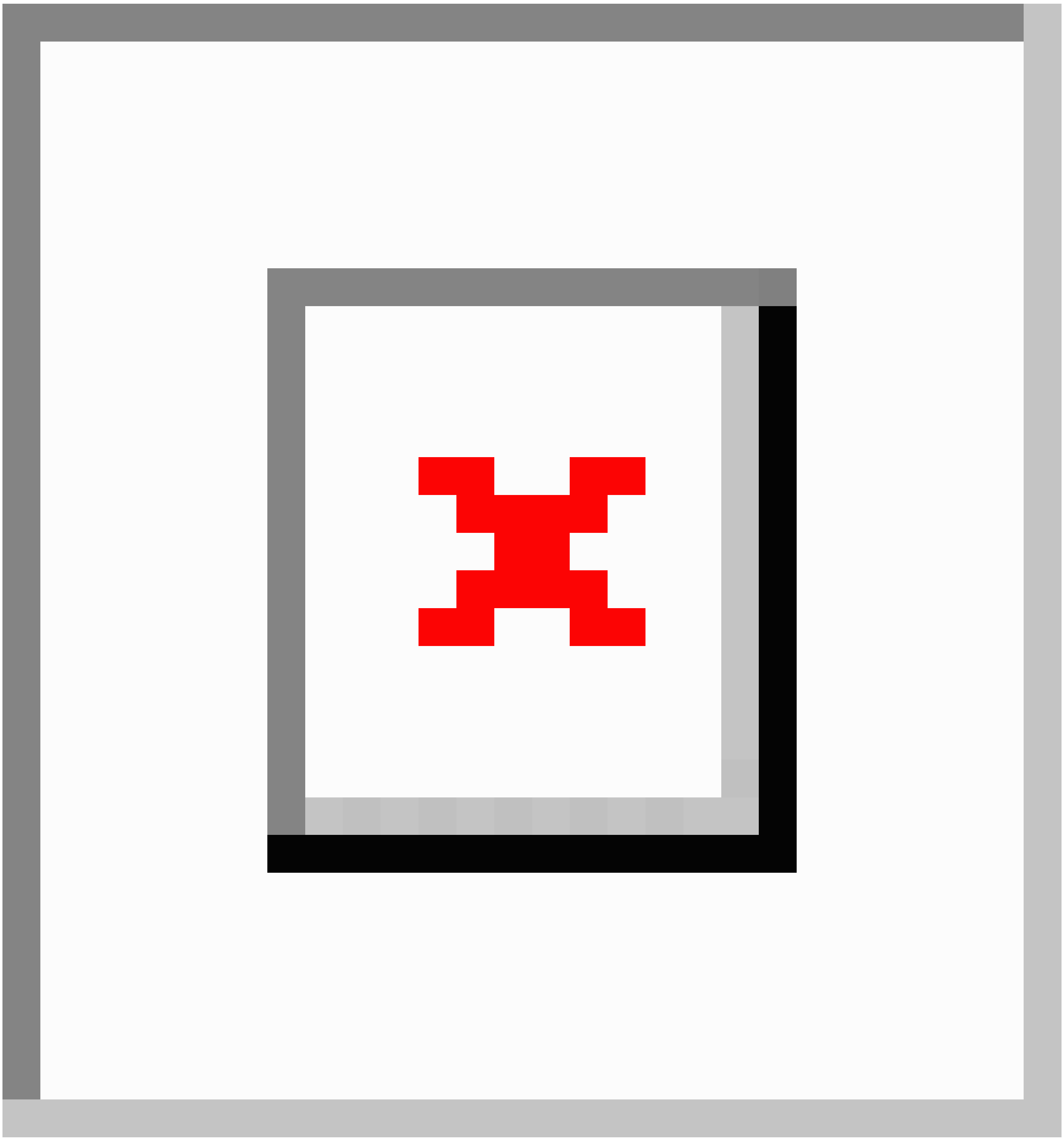
Figure 2. Translated screenshot of the barrier questionnaire (left) and plain advisory text; care as usual (right) received by the control group.



Intervention Group: BRI

Respondents in the intervention group received tailored information based on the respondents’ self-reported barrier to contacting the helpline. The tailored information aimed to address common concerns and misconceptions about the helpline. The tailored advisory text for each barrier was structured as follows: (1) a friendly and informal overall advice in the tone of voice of the helpline; (2) an anonymous quote from a help-seeker about his or her experience with help-seeking; and (3) 2 quotes from counselors with or without lived experience. To educate people that they are certainly not the only ones experiencing suicidal thoughts, each advisory text contains a figure showing the text "113 receives an average of 450 requests for help per day." As per the helpline's communication policy, each advisory text emphasizes the importance of talking about suicidal thoughts and taking the first step toward help. Figure 3 shows a translated screenshot of the BRI when selecting the barrier "I don't think that my problems are serious enough for 113."

Figure 3. Translated screenshot of the barrier reduction intervention when selecting the barrier "I don't think my problems are serious enough for 113."



Measurements

Screening and Demographic Variables

This study focused on individuals with severe suicidal ideation. This was measured by the SIDAS. This questionnaire consists of 5 items on a 10-point scale measuring the frequency and controllability of suicidal thoughts, the closeness to a suicide attempt, and the distress and interference with daily activities [8]. In terms of demographic data, respondents' self-reported gender, age group, and treatment status for mental health problems (yes, no, on a waiting list) were collected. One's likelihood of contacting the helpline before randomization was measured by the question "How likely are you to contact 113's helpline via chat or phone?" Answering options ranged from "not likely" to "very likely" on a 5-point scale.

Primary Outcome Measure: Contacting the Helpline

The use of a direct link to the helpline (yes/no) after completing the intervention or the control condition was the primary outcome measure in this trial. At the end of the RCT, respondents were given a choice between 2 buttons: "exit" or "helpline."

Secondary Outcome Measure: Self-Reported Likelihood of Contacting the Helpline

Due to the helpline's anonymous nature, it is not possible to determine if those who did not use the link directly after the intervention or control condition, did contact the helpline at a later moment in time. For that reason, a continuous outcome variable was used, that is, the respondent's self-reported likelihood of contacting the helpline. This was measured with the same likelihood question as in the entry screening questionnaire.

Tertiary Outcome Measure: Satisfaction With the Self-Test

Respondents' satisfaction with the self-test, control and intervention condition, was measured using the Dutch Client Satisfaction Questionnaire (CSQ-3) [22].

Sample Size

The study protocol specified a sample size of 775 participants. We estimated that approximately 10% of participants would drop out during the intervention. We, therefore, expected that we would need to include at least 853 participants. Furthermore, we anticipated that approximately 30% of test-takers would score below the cutoff point for high risk of suicidal behavior and that 20% of test-takers would indicate a high probability of contacting the helpline in the entry screening questionnaire. Factoring in these projections, we determined that a recruitment goal of at least 1706 respondents was necessary [21].

Missing Data

Although all questions in the RCT were mandatory to fill in, a total of 450 persons (36.8%) who were randomized did not complete their participation and closed the web page prematurely. Because demographic data were surveyed during the screening phase, before randomization, there were no missing data on gender, age group, being in treatment or not, and severity of suicidal thoughts (SIDAS score).

Statistical Analyses

The statistical analyses were conducted using R (version 4.1.1; R Foundation for Statistical Computing) and IBM SPSS Statistics (version 25.0). The first step of the data analysis was focused on verifying the randomization process and inspecting dropouts. To determine whether the control group and the intervention group, as well as the completers and noncompleters, were comparable on baseline factors (gender, age, SIDAS score, and in treatment or not) 2-tailed independent *t* tests and χ^2 tests were used. To test the hypothesis that respondents who received the brief BRI were more likely to use the direct link to contact the suicide prevention helpline at the end of the trial than respondents in the control condition, χ^2 analysis as well as multiple logistic regression analyses were used. For the secondary and tertiary outcome measures, multiple linear regression analyses were conducted.

Sensitivity Testing

As a sensitivity test, the outcome measures were also analyzed using intention-to-treat analysis. For this analysis, missing values were present in the following variables: barrier to the helpline (22/1222, 1.8%); likelihood of contacting the helpline after the intervention (342/1222, 28%); the CSQ-3 items CSQ1 (421/1222, 34.5%), CSQ2 (428/1222, 35%), and CSQ3 (437/1222, 35.8%); and the main outcome measure (390/1222, 31.9%). All missing values were imputed using the R package MICE [23], generating 10 independent data sets based on a maximum of 10 iterations each. To allow maximum imputation accuracy, we applied the classification and regression trees (cart) method for all variables and did not exclude any variables from the model a priori. Having generated 10 imputed data sets, missing values in the original data set were replaced by the most frequent imputed value for factors, and by the average imputed value for numeric variables.

Ethical Considerations

This study was reviewed and approved by the Medical Ethics Committee of the Vrije Universiteit Medical Center (registration number: 2021.0443) and is registered at ClinicalTrials.gov (NCT05458830). The study was not subject to the Research Involving Human Subjects Act (*Wet medisch-wetenschappelijk onderzoek met mensen*), as participants were not subject to procedures and were not required to follow rules of behavior. Every participant was directed to a web-based information letter and consent form. After giving consent to participate in the study and stating to be 16 years or older, participants were transferred to the web-based trial. To ensure strict anonymity, no identifying information or IP addresses were gathered.

Results

Study Sample

In just under 3 months, 1222 individuals were randomly assigned to the self-test including a brief BRI (n=610) or to the self-test without BRI (care as usual: n=612). Ultimately, 63.2% completed the RCT and answered all questions; 369 in the intervention group and 403 in the control group. The completers and noncompleters were comparable regarding gender, being in treatment for psychological problems, and SIDAS score. The chi-square test revealed a significant difference in the distribution of age groups (χ^2_6=12.65, P=.049). Table 1 displays the respondents' characteristics of those who completed the intervention. The intervention and control groups were comparable regarding age group, being in treatment for psychological problems, and SIDAS score but not on gender (χ^2_2=6.96, P=.031), with relatively more men (34.5% vs 28.5%) and fewer people who indicated having a gender other than male or female (3% vs 6.2%) in the control group than in the intervention group. The most selected barrier in both groups was "I don't think that my problems are serious enough for 113" (Table 2). A Pearson χ^2 test was conducted to assess the distribution of the barrier categories between the 2 groups. The test revealed no significant difference (χ^2_5=6.92, *P*=.23). There also appears to be no difference in the perceived barrier toward the helpline between men and women (χ^2_5=7.23, *P*=.20; Table S1 in Multimedia Appendix 1).

Table . Respondents' characteristics.

Characteristics		Intervention group (n=369)	Control group (n=403)
Gender, n (%)			
	Male	105 (28.5)	139 (34.5)
	Female	241 (65.3)	252 (62.5)
	Other	23 (6.2)	12 (3.0)
Age group (years), n (%)			
	16 - 24	178 (48.2)	196 (48.6)
	25 - 29	41 (11.1)	44 (10.9)
	30 - 39	70 (19.0)	55 (13.6)
	40 - 49	35 (9.5)	38 (9.4)
	50 - 59	30 (8.1)	41 (10.2)
	60 - 69	10 (2.7)	18 (4.5)
	≥70	5 (1.4)	11 (2.7)
Being in treatment, n (%)			
	Yes	146 (39.6)	176 (43.7)
	No	170 (46.1)	188 (46.7)
	On waiting list	53 (14.4)	39 (9.7)
SIDAS score			
	Mean (SD)	32.33 (6.99)	33.19 (7.21)

Table . Perceived barriers to the helpline per group.

	Intervention group (n=369), n (%)	Control group (n=403)[a], n (%)
I don't think that contacting 113 can help me	72 (19.5)	87 (21.6)
I'm scared to talk about my feelings	67 (18.2)	79 (19.6)
I don't think that my problems are serious enough	115 (31.2)	94 (23.3)
I'm scared that people will find out	41 (11.1)	44 (10.9)
I would rather solve it myself	42 (11.4)	53 (13.2)
I have other reasons	32 (8.7)	46 (11.4)

[a]No significant difference between the 2 groups; χ^2_5=6.92, *P*=.23.

Main Outcome Measure: The Use of a Direct Link to the Helpline

After completing the intervention, most respondents opted to "exit" the web page instead of using the direct link to the helpline. There was no significant (χ^2_1=.61, *P*=.44) difference between the control group and the intervention group, with 35.7% (n=144) of respondents in the control group using the direct link compared with 33.1% (n=122) in the intervention group. Furthermore, logistic regression analysis (Table 3) showed no differences in gender, SIDAS score, or treatment status but did show that age was a predictor for using the direct link to the helpline (odds ratio [OR] 0.88, 95% CI 0.80 - 0.96), with on average, lower odds for the older age groups (Figure S1 in Multimedia Appendix 1). One's score on the first self-reported question regarding the likelihood of contacting the helpline in the entry screening questionnaire also appears to be of influence (OR 1.23, 95% CI 1.05 - 1.43).

Table . Logistic regression analysis: using the direct link to the helpline (772 complete cases).[a]

		ln(OR)[b]	SE	OR (95% CI)	*P* value
Constant		−1.02	0.44	0.36	.02
Group (ref: control)					
	Intervention group	−0.14	0.16	0.87 (0.64 - 1.18)	.38
Age group		−0.13	0.05	0.88 (0.80 - 0.96)	.01
Gender (ref: female)					
	Male	0.08	0.17	1.08 (0.77 - 1.52)	.66
	Other	0.19	0.37	1.21 (0.59 - 2.48)	.60
SIDAS[c] score		0.01	0.01	1.01 (0.99 - 1.03)	.52
Treatment status (ref: in treatment)					
	Not in treatment	0.15	0.17	1.16 (0.83 - 1.63)	.37
	On waiting list	−0.04	0.26	0.96 (0.58 - 1.59)	.89
Likelihood contact pre-intervention		0.20	0.08	1.23 (1.05 - 1.43)	.01

[a]Nagelkerke R^2=.031.
[b]OR: odds ratio.
[c]SIDAS: Suicidal Ideation Attributes Scale.

Self-Reported Likelihood of Contacting the Helpline

In the entry screening questionnaire, both the intervention group (mean 2.41, SD 0.96) and control group (mean 2.40, SD 1.00) scored similarly on the self-reported likelihood of contacting the helpline (t_{770}=.13, *P*=.89). This scale ranges from 1 to 4, as the people who scored "very likely" in the screening phase were not included in the RCT. After the trial, the intervention group (mean 2.78, SD 1.06) scored significantly higher on the 5-point scale than the control group (mean 2.55, SD 1.09; t_{770}=2.96, *P*=.003). Additionally, the results of the multiple linear regression analysis (Table 4) indicate that an individual's baseline score on the likelihood of contacting the helpline was a significant contributing factor.

Table . Regression analysis: self-reported likelihood of contacting the helpline (772 complete cases).[a]

		B	SE	β	95% CI	*P* value
Constant		1.42	0.14	—[b]	1.14 to 1.69	<.001
Group (ref: control)						
	Intervention group	0.22	0.05	0.10	0.12 to 0.32	<.001
Age group		−0.03	0.02	−0.05	−0.06 to 0.00	.05
Gender (ref: female)						
	Male	−0.07	0.06	−0.03	−0.18 to 0.04	.19
	Other	−0.14	0.12	−0.03	−0.38 to 0.10	.25
SIDAS[c] score		0.00	0.00	0.01	0.00 to 0.01	.57
Treatment status (ref: in treatment)						
	Not in treatment	−0.05	0.06	−0.03	−0.16 to 0.05	.32
	On waiting list	00.00	0.08	0.00	−0.16 to 0.17	.96
Likelihood contact pre-intervention		0.84	0.03	0.76	0.79 to 0.89	<.001

[a]Adjusted R^2=0.598.
[b]Not applicable.
[c]SIDAS: Suicidal Ideation Attributes Scale.

Satisfaction With the Self-Test

The respondents in the intervention group scored slightly higher on satisfaction with the self-test (mean 8.55, SD 1.94) than the control group (mean 8.21, SD 1.98; t_{770}=2.36, *P*=.02). Furthermore, multiple linear regression analysis shows that, in general, the older age groups scored lower on the CSQ scale than the younger age groups. Men (mean 7.93, SD 2.17) were less satisfied with the self-test in the control and intervention conditions than women (mean 8.60, SD 1.79). The results of this regression analysis can be found in Table S5 in Multimedia Appendix 1.

Evaluating the Effectiveness of the BRI for Men and Those of Middle Age

One-third of the men in the intervention group (35/105, 33%) used the direct link to the helpline, and this is comparable to those in the control group (49/139, 35%; χ^2_1=0.098, *P*=.76). This result did not change after controlling for the different confounders. When we look at the attitudinal measure (Table 5), we see that the intervention group has a positive influence (B=0.20, SE=0.10, *P*=.04) on the self-reported likelihood of contacting the helpline.

Table . Regression analyses: self-reported likelihood of contacting the helpline for men (244 complete cases).[a]

	B	SE	β	95% CI	*P* value
Constant	0.09	0.26	—[b]	−0.42 to 0.59	.74
Group (ref: control)					
Intervention group	0.20	0.10	0.09	0.01 to 0.39	.04
Age group	−0.02	0.03	−0.04	−0.07 to 0.03	.42
SIDAS[c] score	0.01	0.01	0.05	−0.01 to 0.02	.23
Treatment status (ref: in treatment)					
Not in treatment	0.04	0.11	0.02	−0.17 to 0.25	.72
On waiting list	0.05	0.19	0.01	−0.31 to 0.42	.78
Likelihood contact pre-intervention	0.83	0.05	0.75	0.74 to 0.93	<.001

[a]Adjusted R^2=0.55.

[b]Not applicable.

[c]SIDAS: Suicidal Ideation Attributes Scale.

Among the 172 middle-aged respondents (40 - 70 years), 32% (n=24) of the intervention group and 27% (n=26) in the control group used the direct link to the helpline. This difference was too small to be significant (χ^2_1=0.55, *P*=.46) and did not change after controlling for the different confounders. On the attitudinal measure, the intervention group had a mean score of 2.63 (SD 0.91) regarding the likelihood of contacting the helpline, while the control group showed a slightly lower mean score of 2.48 (SD 0.98). When controlled for the various confounders, this difference was significant (Table 6).

Table . Regression analyses: self-reported likelihood of contacting the helpline for middle-aged respondents (172 complete cases).[a]

	B	SE	β	95% CI	*P* value
Constant	−0.06	0.29	—[b]	−0.64 to 0.51	.83
Group (ref: control)					
Intervention group	0.22	0.10	0.12	0.02 to 0.43	.03
Gender (ref: female)					
Male	−0.01	0.11	−0.01	−0.22 to 0.19	.89
Other	0.22	0.38	0.03	−0.54 to 0.98	.56
SIDAS[c] score	0.02	0.01	0.12	0.00 to 0.03	.03
Treatment status (ref: in treatment)					
Not in treatment	−0.09	0.11	−0.05	−0.32 to 0.13	.41
On waiting list	0.09	0.17	0.03	−0.24 to 0.43	.59
Likelihood contact pre-intervention	0.72	0.05	0.71	0.61 to 0.83	<.001

[a]Adjusted R^2=0.53.

[b]Not applicable.

[c]SIDAS: Suicidal Ideation Attributes Scale.

Sensitivity Test

As a sensitivity test, the outcome measures were also analyzed using intention-to-treat analysis. When using the imputed data set (n=1222) for our regression models, the outcomes were consistently confirmed. The only difference can be found in the regression analysis on self-reported likelihood of contacting the helpline for men. When using the imputed data, the significant difference between the intervention and control groups disappeared. The summaries of the regression models can be found in Tables S3-S9 in Multimedia Appendix 1

Discussion

Principal Findings

The aim of our study was 2-fold: (1) to measure the effectiveness of a brief BRI provided in the self-test motivating people with severe suicidal thoughts to contact the Dutch suicide prevention helpline and (2) to specifically evaluate the effectiveness of the BRI in increasing service use by high-risk groups for suicide such as men and middle-aged people.

This study was the first occasion in which respondents were actively recruited among the anonymous website visitors of the Dutch suicide prevention helpline. Recruiting the respondents went faster than expected; in just under 3 months, we had reached the targeted number of respondents. Respondents' characteristics were in line with expectations based on previous research on the helpline. They were mainly female, young, and having severe suicidal thoughts [5-7]. Almost half of the respondents (358/772, 46.4%) were not receiving treatment from mental health services at the time. This percentage was higher among the male (148/244, 60.7%) than the female respondents (192/493, 38.9%). At the time of the study, 11.9% (n=92) of the respondents were on a waiting list for mental health services (men: 21/244, 8.6%; women: 68/493, 13.8%). The higher percentage of men not receiving treatment from mental health services for suicidality is in line with previous research [10,24,25]. These gender differences in help-seeking behavior are likely to be a factor behind the higher number of suicides among men compared with women [26,27]. During the pilot study in 2021, 7252 people filled in the self-test. The majority of them (n=5200, 72%) scored higher than the cutoff point for severe suicidal thoughts, and the mean SIDAS score was 27 on a scale ranging from 0 to 50 [21]. In this study, 2124 website visitors completed the SIDAS and the demographic questions in the screening phase. More than three-quarters (78%) scored higher than the cutoff point for severe suicidal thoughts (SIDAS≥21). The mean SIDAS score in the screening phase was 28.4. Among the eventual study sample, the mean SIDAS score was 33, as the BRI was intended for the high-risk group.

The most selected barriers to the helpline in both groups were "I don't think that my problems are serious enough," "I don't think that 113 can help me," and "I'm scared to talk about my feelings." These barriers align with known barriers to seeking professional help in people with suicidal ideation, such as a low perceived need for support, a strong inclination toward self-reliance and stigmatizing attitudes toward suicide, mental health issues, and professional treatment [10]. The results indicated no difference in the use of a direct link to the helpline immediately at the end of the few-minute–long trial. However, the results do show an impact on one's self-reported likelihood of contacting the helpline. In addition, respondents who received the BRI were more positive about the self-test than those in the control condition. When looking at the high-risk groups men and middle-aged people specifically, the results were comparable to the whole group of respondents. Brief eHealth interventions tend to struggle with achieving significant effects on behavioral outcomes [28,29]. In their review of methods for human-centered eHealth development, Kip et al [30] describe many varieties of methods and products that can be used throughout the development process to prioritize user perspectives to enhance the effectiveness and usability of eHealth solutions.

Limitations

Although barriers to help-seeking have been studied in traditional mental health care, persuading reluctant individuals with severe suicidal ideation in a web-based environment toward professional help is still a largely unexplored area. Therefore, this study, with its randomized controlled design, large sample sizes, and hard-to-reach and at-risk study population, brings a valuable contribution to the suicide prevention literature. However, this study has some limitations that should be taken into consideration when interpreting its results. First and most importantly, the BRI in this study was only text-based. For future research, it would be valuable to determine the effects of different types of components (eg, video material) in a similar BRI. Second, due to the anonymity of the suicide prevention helpline, it was not possible to include a follow-up measurement. It is therefore not feasible to determine if those who did not use the link to the helpline after the intervention did contact the helpline at a later moment in time. Because of this, the self-reported likelihood of contacting the helpline has also been measured as well as the use of the direct link. Third, there were relatively high dropout rates during the RCT (36.8%). This is not entirely unexpected for a short, anonymous, web-based survey but still noteworthy. In order to assess the robustness of the results, analyses were carried out on the complete cases (per protocol) as well as on an imputed data set (intention-to-treat). Finally, we recognize the importance of involving people with lived and living experiences of suicidality in research. Although the content and barriers in the RCT were based on a qualitative pilot study among the self-test users, it would be desirable to involve people with lived experience throughout the research process [21]. To give researchers the confidence and willingness to involve the research population, Orygen developed guidelines for safely and effectively involving young people with lived experience in research [31].

Practical Implications and Future Research

This short text-based BRI was not effective in inducing direct behavioral change. However, this low-cost and low-effort method to reach those who are highly at risk and are reluctant to contact helpline services gives more insight into the perceived barriers to service. The 5 barriers in the RCT were chosen based on a pilot study among self-test users. These barriers may differ in different cultural and social contexts. For future research, it would be advantageous to join efforts with other helplines to see how much these barriers vary between countries. Furthermore, it would have added value to monitor helpline use for subgroups after the launch of a BRI. In addition, it would also be useful to seek input from help seekers who have had contact with the helpline, explore their experiences during interactions, and identify effective communication strategies. Our follow-up research is focused on using video materials in the BRI.

Conclusions

The short BRI, built in a self-test for suicidal thoughts, aimed to persuade individuals with severe suicidal ideation to contact the Dutch suicide prevention helpline. Although the BRI could not induce direct behavioral change in our population with severe suicidal thoughts, these few minutes of intervention did manage to increase the self-reported likelihood of contacting the helpline. When tailored to its users, a BRI may have the potential to be a low-cost, highly scalable, and easily implementable method to increase service use for helplines worldwide.

Acknowledgments

We would like to thank all participants of this study for their valuable contribution. We would also like to express our gratitude to Scarlet Hemkes for her insightful content; Salim Salmi and Anouk Romers for their technical assistance; and Siyu Chen, whose thesis served as a great source of inspiration for this study [32]. This research was funded by 113 Suicide Prevention, which is itself mainly funded by the Ministry of Health, Welfare & Sports of the Netherlands. All authors report no financial relationships with commercial interests.

Data Availability

The data generated or analyzed during this study will not be publicly available due to organization policies regarding help-seeker data but are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary materials regarding perceived barriers, use of the direct link to the helpline, and sensitivity analyses.

[DOCX File, 44 KB - mental_v11i1e56396_app1.docx]

Checklist 1

CONSORT-eHEALTH checklist (V 1.6.1).

[PDF File, 1340 KB - mental_v11i1e56396_app2.pdf]

References

1. Jia X, Pang Y, Liu LS. Online health information seeking behavior: a systematic review. Healthcare (Basel) 2021 Dec 16;9(12):1740. [doi: 10.3390/healthcare9121740] [Medline: 34946466]
2. Marton C, Choo CW. A review of theoretical models of health information seeking on the web. J Doc 2012 Apr 20;68(3):330-352. [doi: 10.1108/00220411211225575]
3. Jaarverslag 2021. Annual report 2021. : 113 Suicide Prevention; 2022 URL: https://jaarverslag-2021.113.nl/ [accessed 2024-08-23]
4. Publieksjaarverslag 2022. Annual report 2022. : 113 Suicide Prevention; 2023 URL: https://publieksjaarverslag-2022.113.nl/ [accessed 2023-09-21]
5. Mokkenstorm JK, Eikelenboom M, Huisman A, et al. Evaluation of the 113Online suicide prevention crisis chat service: outcomes, helper behaviors and comparison to telephone hotlines. Suicide Life Threat Behav 2017 Jun;47(3):282-296. [doi: 10.1111/sltb.12286] [Medline: 27539122]
6. Salmi S, Mérelle S, Gilissen R, van der Mei R, Bhulai S. Detecting changes in help seeker conversations on a suicide prevention helpline during the COVID- 19 pandemic: in-depth analysis using encoder representations from transformers. BMC Public Health 2022 Mar 18;22(1):530. [doi: 10.1186/s12889-022-12926-2] [Medline: 35300638]
7. Van der Burgt MCA, Mérelle S, Beekman ATF, Gilissen R. The impact of COVID-19 on the suicide prevention helpline in the Netherlands. Crisis 2023 Jul;44(4):285-291. [doi: 10.1027/0227-5910/a000863] [Medline: 35723060]
8. Van Spijker BAJ, Batterham PJ, Calear AL, et al. The Suicidal Ideation Attributes Scale (SIDAS): community-based validation study of a new scale for the measurement of suicidal ideation. Suicide Life Threat Behav 2014 Aug;44(4):408-419. [doi: 10.1111/sltb.12084] [Medline: 24612048]

9. Janssen W, Van Raak J, Van der Lucht Y, Van Ballegooijen W, Mérelle S. Can outcomes of a chat-based suicide prevention helpline be improved by training counselors in motivational interviewing? a non-randomized controlled trial. Front Digit Health 2022;4:871841. [doi: 10.3389/fdgth.2022.871841] [Medline: 35801118]
10. Han J, Batterham PJ, Calear AL, Randall R. Factors influencing professional help-seeking for suicidality: a systematic review. Crisis 2018 May;39(3):175-196. [doi: 10.1027/0227-5910/a000485] [Medline: 29052431]
11. Bruffaerts R, Demyttenaere K, Hwang I, et al. Treatment of suicidal people around the world. Br J Psychiatry 2011 Jul;199(1):64-70. [doi: 10.1192/bjp.bp.110.084129] [Medline: 21263012]
12. Hom MA, Stanley IH, Joiner TE. Evaluating factors and interventions that influence help-seeking and mental health service utilization among suicidal individuals: a review of the literature. Clin Psychol Rev 2015 Aug;40:28-39. [doi: 10.1016/j.cpr.2015.05.006] [Medline: 26048165]
13. Ali MM, Lackey S, Mutter R, McKeon R. The relationship between perceived unmet mental health care needs and suicidal ideation and attempt. Adm Policy Ment Health 2018 Sep;45(5):709-715. [doi: 10.1007/s10488-018-0856-z] [Medline: 29476291]
14. Roberts T, Miguel Esponda G, Torre C, Pillai P, Cohen A, Burgess RA. Reconceptualising the treatment gap for common mental disorders: a fork in the road for global mental health? Br J Psychiatry 2022 Sep;221(3):553-557. [doi: 10.1192/bjp.2021.221] [Medline: 35137680]
15. Gunnell D. A population health perspective on suicide research and prevention: what we know, what we need to know, and policy priorities. Crisis 2015;36(3):155-160. [doi: 10.1027/0227-5910/a000317] [Medline: 26266821]
16. O'Connor RC, Worthman CM, Abanga M, et al. Gone too soon: priorities for action to prevent premature mortality associated with mental illness and mental distress. Lancet Psychiatry 2023 Jun;10(6):452-464. [doi: 10.1016/S2215-0366(23)00058-5] [Medline: 37182526]
17. Suicide worldwide in 2019: global health estimates. World Health Organization. 2021. URL: https://www.who.int/publications/i/item/9789240026643 [accessed 2024-08-23]
18. Oinas-Kukkonen H, Harjumaa M. Persuasive systems design: key issues, process model, and system features. CAIS 2009;24(1):485-500. [doi: 10.17705/1CAIS.02428]
19. Jaroszewski AC, Morris RR, Nock MK. Randomized controlled trial of an online machine learning-driven risk assessment and intervention platform for increasing the use of crisis services. J Consult Clin Psychol 2019 Apr;87(4):370-379. [doi: 10.1037/ccp0000389] [Medline: 30883164]
20. Berkelmans G, van der Mei R, Bhulai S, Gilissen R. Identifying socio-demographic risk factors for suicide using data on an individual level. BMC Public Health 2021 Sep 18;21(1):1702. [doi: 10.1186/s12889-021-11743-3] [Medline: 34537046]
21. Van der Burgt MCA, Mérelle S, Brinkman WP, Beekman ATF, Gilissen R. Breaking down barriers to a suicide prevention helpline: protocol for a web-based randomized controlled trial. JMIR Res Protoc 2023 Apr 24;12:e41078. [doi: 10.2196/41078] [Medline: 37093641]
22. Attkisson CC, Greenfield TK. The Client Satisfaction Questionnaire (CSQ) Scales Outcome Assessment in Clinical Practice: Williams & Wilkins; 1995.
23. Van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. J Stat Softw Foundat Open Access Statis 2011 Dec 12;45(3):1-67 [FREE Full text] [doi: 10.18637/jss.v045.i03]
24. Tang S, Reily NM, Arena AF, et al. People who die by suicide without receiving mental health services: a systematic review. Front Public Health 2021;9:736948. [doi: 10.3389/fpubh.2021.736948] [Medline: 35118036]
25. Rens E, Portzky G, Morrens M, Dom G, Van den Broeck K, Gijzen M. An exploration of suicidal ideation and attempts, and care use and unmet need among suicide-ideators in a Belgian population study. BMC Public Health 2023 Sep 7;23(1):1741. [doi: 10.1186/s12889-023-16630-7] [Medline: 37679752]
26. Tang S, Reily NM, Batterham PJ, et al. Correlates of non-receipt of formal mental health services among Australian men experiencing thoughts of suicide. J Affect Disord 2023 Jan;11:100455. [doi: 10.1016/j.jadr.2022.100455]
27. Stene-Larsen K, Reneflot A. Contact with primary and mental health care prior to suicide: a systematic review of the literature from 2000 to 2017. Scand J Public Health 2019 Feb;47(1):9-17. [doi: 10.1177/1403494817746274] [Medline: 29207932]
28. Batterham PJ, Calear AL, Sunderland M, et al. A brief intervention to increase uptake and adherence of an internet-based program for depression and anxiety (enhancing engagement with psychosocial interventions): randomized controlled trial. J Med Internet Res 2021 Jul 27;23(7):e23029. [doi: 10.2196/23029] [Medline: 34313595]
29. Van Gemert-Pijnen J, Nijland N, Van Limburg M, et al. A holistic framework to improve the uptake and impact of eHealth technologies. J Med Internet Res 2011 Dec 5;13(4):e111. [doi: 10.2196/jmir.1672] [Medline: 22155738]
30. Kip H, Keizer J, da Silva MC, Beerlage-de Jong N, Köhle N, Kelders SM. Methods for human-centered eHealth development: narrative scoping review. J Med Internet Res 2022 Jan 27;24(1):e31858. [doi: 10.2196/31858] [Medline: 35084359]
31. Webb M, Cooper C, Hemming L, et al. Guidelines for involving young people with lived and living experience of suicide in suicide research. : Orygen; 2023.
32. Chen S. An interactive suicide ideation self-test service for helping people resolve barriers towards contacting a suicide prevention helpline. TU Delft. URL: http://resolver.tudelft.nl/uuid:e1be41a7-313e-4683-98f7-ffefab31219c [accessed 2024-08-23]

Abbreviations

BRI: barrier reduction intervention
CSQ: Client Satisfaction Questionnaire
OR: odds ratio
RCT: randomized controlled trial
SIDAS: Suicidal Ideation Attributes Scale

Edited by J Torous; submitted 15.01.24; peer-reviewed by A AL-Asadi, P Batterham; revised version received 11.05.24; accepted 28.05.24; published 05.09.24.

Please cite as:
Van der Burgt MCA, Mérelle S, Brinkman WP, Beekman ATF, Gilissen R
Breaking Down Barriers to a Suicide Prevention Helpline: Web-Based Randomized Controlled Trial
JMIR Ment Health 2024;11:e56396
URL: https://mental.jmir.org/2024/1/e56396
doi:10.2196/56396

Talk Time Differences Between Interregional and Intraregional Calls to a Crisis Helpline: Statistical Analysis

Robin Turkington[1,*], MSc, PhD; Courtney Potts[1,*], MSc; Maurice Mulvenna[1,*], MPhil, PhD; Raymond Bond[1,*], PhD; Siobhán O'Neill[2,*], MPsychSc, PhD; Edel Ennis[2,*], PhD; Katie Hardcastle[3,*], MSc; Elizabeth Scowcroft[3,*], MSc, PhD; Ciaran Moore[4,*], BCom; Louise Hamra[4,*], BA

1

2

3

4

*all authors contributed equally

Corresponding Author:
Maurice Mulvenna, MPhil, PhD

Abstract

Background: National suicide prevention strategies are general population-based approaches to prevent suicide by promoting help-seeking behaviors and implementing interventions. Crisis helplines are one of the suicide prevention resources available for public use, where individuals experiencing a crisis can talk to a trained volunteer. Samaritans UK operates on a national scale, with a number of branches located within each of the United Kingdom's 4 countries or regions.

Objectives: The aim of this study was to identify any differences in call duration across the helpline service in order to determine whether service varied interregionally and intraregionally and to determine the impact of calls answered in the same region as the caller, compared with calls answered in a different region on the duration of calls made from landlines to Samaritans UK.

Methods: Calls may be routed by Samaritans, wherein the telephony system sends the call to the next available volunteer, irrespective of location; therefore, individuals may be routed to a branch within the same region as the caller's current region (intraregional calls) or routed to a branch that is in a different region from that of the caller's current region (interregional calls). The origin of calls by region was identified using the landline prefix of the anonymized caller identifier, along with the region of the destination branch (as branch details are recorded in the call details record). First, a Levene's test of homogeneity of variance was carried out for each condition, that is, England calls and Scotland calls. Thereafter, for each condition, a one-way ANOVA or one-way analysis of means was carried out to evaluate any significant differences in call duration.

Results: ANOVA results showed that there are significant differences in call durations between intraregional calls and interregional calls (P<.001). Across all conditions within this study, callers stayed on the phone for a shorter period of time when routed to a branch that is within the same region as the call origin than if they were put through to a branch within a different region than the call origin.

Conclusions: Statistical analyses showed that there were significant differences between interregional and intraregional calls. On average, callers to crisis helplines stayed on the phone for a shorter period of time if they were routed to a branch within the same region in which the call originated than if they were routed to a branch in a different region of origin. The findings from this study have practical applications, which may allow crisis helplines to manage their resources more effectively and improve caller satisfaction with the service.

(JMIR Ment Health 2024;11:e58162) doi:10.2196/58162

KEYWORDS

crisis helplines; call duration; mental health; suicide; suicidal; suicide prevention; population-based; help-seeking behavior; Samaritans; UK; telephony; telephone; telephones; one-way analysis; call; calls; talk time; support; talk time differences

Introduction

Crisis helplines are one of the oldest suicide prevention resources available for public use [1,2]. They are based on the premise that suicide may be prevented by supporting callers who are in a crisis situation, which is defined as a transient state of psychological disequilibrium where an individual's coping mechanisms are no longer working [3-5].

In global helpline evidence, characteristics of the call content and call outcomes have been examined [5,6] and helper behavior(s) and intervention styles have been examined through silent monitoring to determine which aspects of helper behavior and intervention styles were significantly related to positive outcomes [7]. In addition, there is evidence of subclasses of callers to crisis helplines, with different call characteristics, such as call duration. These variations may reflect the different needs of callers from each caller group [8-10].

There are many crisis helplines in operation within the United Kingdom; one of the oldest in operation is Samaritans. Samaritans UK is a volunteer-based listening service, where the role of the volunteer is to listen to the caller in a respectful, nonjudgmental, and nondirective manner. The model of crisis helplines such as Samaritans is to offer one-off support; however, many individuals use telephony support more than once and often call the service repeatedly over a period of time [8-10].

Given the role of crisis helplines in suicide prevention, it is important to understand the factors that affect access to these services. National suicide prevention strategies describe the importance of connectedness as a protective factor against suicide [11], which is also a prominent part of theories that explain suicidal behavior [12]; the length of call may reflect the extent to which a caller feels supported and able to discuss his or her personal situation. Depending on predetermined criteria set by the crisis helpline, calls may or may not be routed to the nearest branch to call origin but could be routed to a branch where there is a volunteer available to answer the call. This is common practice in many other telephony-based services, as it allows the client to be connected to a service agent much faster. For example, at Samaritans, callers contacting the service from one region of the United Kingdom may be answered by a branch in a different region of the United Kingdom.

The objective of this study was to assess whether interregional and intraregional calls differ in call duration. Thus, this research seeks to answer the question: Do callers stay on the phone for a longer period of time depending on whether they are put through to a branch within their own region or to a branch within another region?

Methods

Data Collection

Data used for this study were provided by Samaritans UK. Calls made to this service originate from the constituent regions within the United Kingdom and are routed to Samaritans branches throughout the United Kingdom. Samaritans operate with a single helpline with no regional routing. This means that callers can be routed to any Samaritans branch within the United Kingdom regardless of where the call originated. The data analyzed in this study were from January 2015 to August 2018. During that period, Samaritans received 25,177,944 calls; of these, 4,647,567 (18.5%) were made from landline numbers. Calls to the Samaritans Welsh language line were excluded from this study. The data contained the following variables: caller ID, an anonymous identifier for the caller; country-specific region of origin (derived from the first number of digits within the caller ID variable representing the prefix for the phone number region); destination branch name (place where call was routed); and call duration. In addition, the average call duration for each caller to each region within the data was computed.

Ethical Considerations

Ethical approval for this work was provided by Ulster University's Psychology Filter Committee for application "Understanding Samaritans caller behaviour: a machine learning analysis to identify latent sub population and model caller behaviour," dated August 8, 2016 (FCPSY-08082016).

Data Wrangling

This stage consisted of identifying the calls within the data that were made from landline phone numbers. The data contained a "caller ID" variable which is an anonymous identifier for the caller. The first number of digits within the caller ID variable represented the prefix for the phone number region from where the caller made the call. After the end of the prefix in the identifier, the remaining digits are hashed, therefore, anonymizing the caller's identity. From this, calls that were made by landlines were identified and then extracted for the next stage of the wrangling process.

A landline prefix is specific to an area within a country or region. The calls made by each landline number were categorized by region of origin. A total of 608 places of origin were identified within the data, which equated to 4,647,567 phone calls.

Another variable within the call data that was used was the destination branch name, which is the location of the Samaritans branch where the call was routed. There are 187 destination branches and each of these branches was categorized by region. At this stage, it is now possible to filter calls made from each region of origin to each region of destination. Average call duration for each caller to each region within the data was extracted.

Statistical Analysis

R Studio (version 3.4.4; Posit PBC) was used to conduct data wrangling and statistical analysis. The ggplot2 package (Posit PBC) [13] was used to create data visualizations. Base R functions (R Core Team) were used to conduct data summaries.

Results

Overview

Shorter-duration calls are defined as those less than the mean duration of calls overall, where the mean duration is around 1000 seconds (approximately 16 minutes). Therefore shorter-duration calls are less than 16 minutes, while longer-duration calls are greater than 16 minutes.

The average call duration from callers in each region to other regions was subjected to statistical analysis using Levene's test of homogeneity of variance and one-way ANOVA or one-way analysis of means. The following subsections detail the analysis for each of the regions in turn, comparing intraregion and then interregion calls.

Throughout the next 4 subsections, the conditions will be referred to as <region of call origin>_<region of branch answering call>. For example, England_NI pertains to calls that originated from landlines in England and answered by a branch in Northern Ireland (NI).

England to England Versus Other Regions

Levene's test of homogeneity of variance assumption was violated ($F_{3,\ 278036}$=9.366; P<.001). A one-way ANOVA or one-way analysis of means was carried out to uncover any significant differences in average call duration between interregional and intraregional calls. There was a significant difference in average call duration between interregional and intraregional calls for England ($F_{3,\ 51876}$=17.504; P<.001).

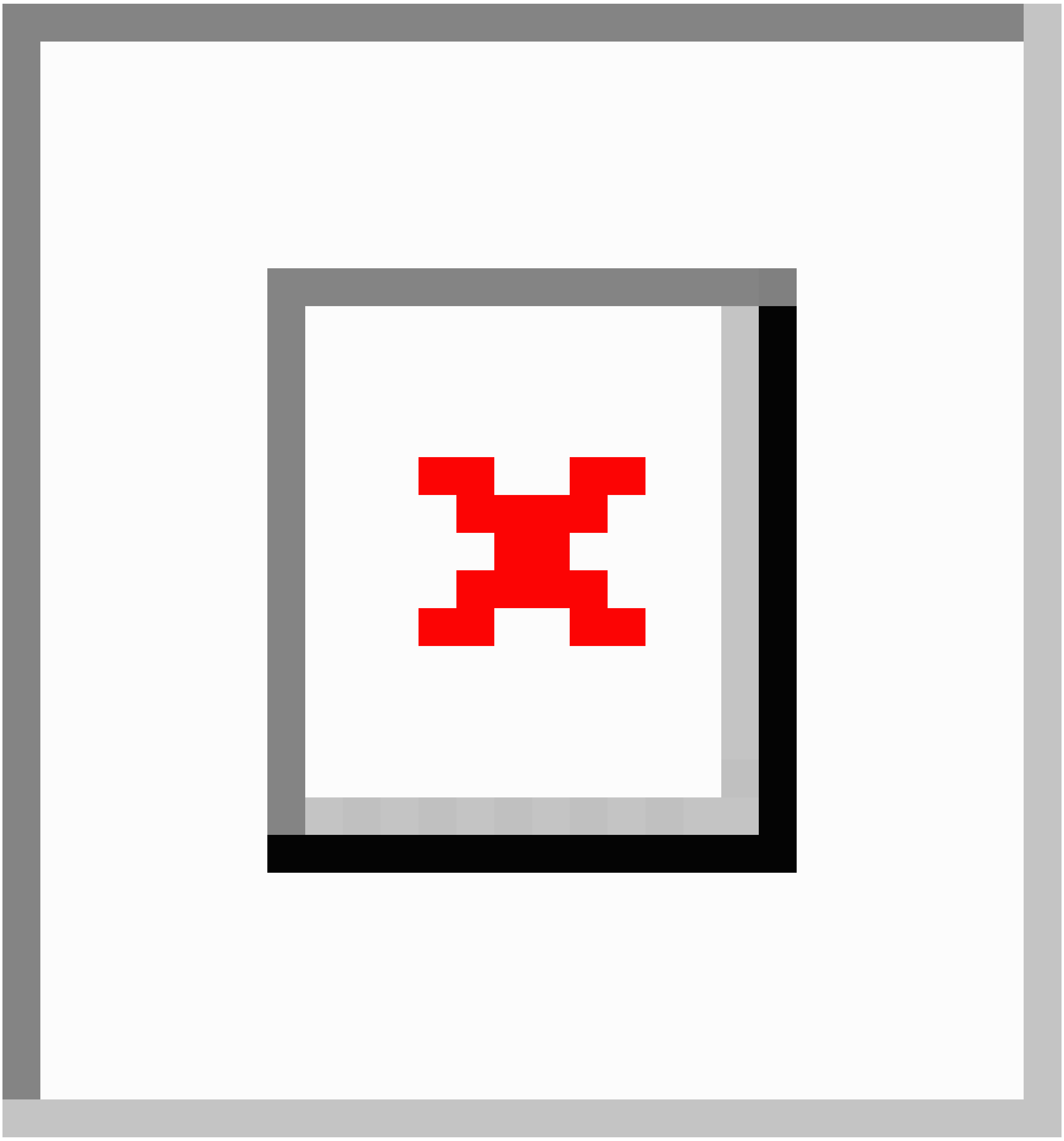
Table 1 displays the pairwise comparisons of the levels within the England condition. Callers from England stayed on the phone for significantly longer if they were put through to a branch in NI than if they were put through to a branch in England by 38 seconds on average (difference between mean durations = 37.61 seconds; P<.001). Callers from England stayed on the phone for longer if they were put through to a branch in Scotland than if they were put through to a branch in England by 31 seconds on average (difference = 31.17 seconds; P<.001). Pairwise comparisons of average call duration from England landlines to England branches versus other regions were not significant for all other comparisons (Table 1).

As shown in Figure 1, there was a higher density of calls with a shorter duration made from England to England than in other conditions. Meanwhile, there was a marginally higher density of calls of longer duration made from England to Scotland than in other conditions. Apart from these 2 observations, there was little variation between the conditions elsewhere. Figure 1 also displays the average call duration between England landlines to England branches and other regional branches: ordered from shortest to longest mean duration: England-England (mean call duration 966 seconds), then England-Wales (mean call duration 977 seconds), England-Scotland (mean call duration 997 seconds), and finally England-NI (mean call duration 1003 seconds).

Table . Pairwise comparisons of average call duration from England landlines to England branches versus other regions.

Group	Difference between means (duration in seconds)	*P* value
England_NI-England_England	37.61	<.001
England_Scotland-England_England	31.17	<.001
England_Wales-England_England	11.52	.13
England-Scotland-England_NI	−6.44	.44
England_Wales-England_NI	−26.09	.01
England_Wales-England-Scotland	−19.65	.04

Figure 1. Density plot of call durations from England to England versus other regions (log scale 10) and average call durations (95% CI).



Scotland to Scotland Versus Other Regions

Levene's test of homogeneity of variance assumption was violated ($F_{3,\ 31477}$=15.042; P<.001). A one-way ANOVA or one-way analysis of means was carried out to uncover any statistically significant differences between average call durations between interregional and intraregional calls. There was a significant difference in average call duration between interregional and intraregional calls for Scotland ($F_{3,\ 6528.5}$=37.341; P<.001).

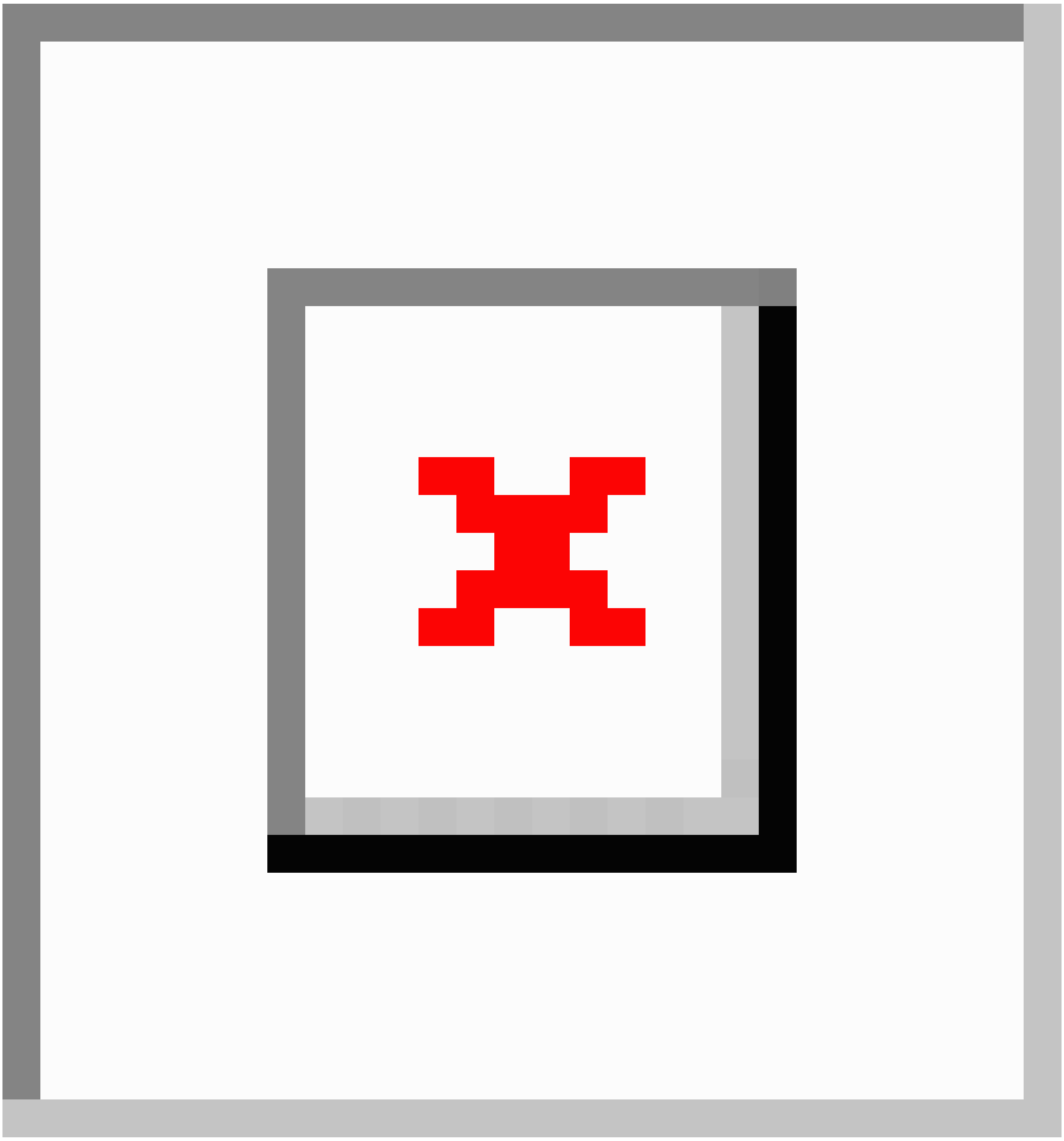
Table 2 displays pairwise comparisons for each level within the Scotland condition. Callers from Scotland stayed on the phone for 120 seconds longer on average if they were put through to a branch in England than if they were put through to a branch in Scotland (difference = −118.43; P<.001). When callers from Scotland were put through to a branch in England, they stayed on the phone for 105 seconds longer than if they were put through to a branch in Wales (difference = −105; P<.001). Callers from Scotland stayed on the phone for 120 seconds longer on average if they were put through to a branch in NI than if they were put through to a branch in Scotland (difference = −116.6; P<.001). Callers from Scotland stayed on the phone for around 103 seconds longer on average if they were put through to a branch in NI than if they put through to a branch in Wales (difference = −103.17; P<.001). Pairwise comparisons of average call duration from Scotland landlines to Scotland branches versus other regions were not significant for all other comparisons (Table 2).

As illustrated in Figure 2, a higher density of calls with a shorter duration was observed from Scotland-Scotland than the other conditions. In terms of calls with a longer duration, Scotland-Scotland had the same density as Scotland-NI. Scotland-England had the highest density of calls with a longer duration, while Scotland-Wales had the lowest density of longer duration calls. Figure 2 also displays the average call duration between Scotland-Scotland and other regional branches. Ordered from shortest to longest mean duration: Scotland-Scotland (mean call duration 916 seconds), Scotland-Wales (mean call duration 930 seconds), Scotland to NI (mean call duration 1033 seconds), and finally Scotland-England (mean call duration 1035 seconds).

Table . Pairwise comparisons of average call duration from Scotland landlines to Scotland branches versus other regions.

Group	Differences between means (duration in seconds)	*P* value
Scotland_NI-Scotland_England	−1.83	.93
Scotland_Scotland-Scotland_England	−118.43	<.001
Scotland_Wales-Scotland_England	−105	<.001
Scotland_Scotland-Scotland-NI	−117	<.001
Scotland_Wales-Scotland_NI	−103.17	<.001
Scotland_Wales-Scotland_Scotland	13.43	.66

Figure 2. Density plot of call durations from Scotland (SCO) to Scotland versus other regions (log scale 10) and average call durations (95% CI).



Wales to Wales Versus Other Regions

Levene's test of homogeneity of variance assumption was violated ($F_{3,\ 16160}$=13.088; P<.001). A one-way ANOVA or one-way analysis of means was carried out to uncover any significant differences in average call duration between interregional and intraregional calls. There was a significant difference between interregional and intraregional calls for Wales ($F_{3,\ 4518}$=31.211; P<.001).

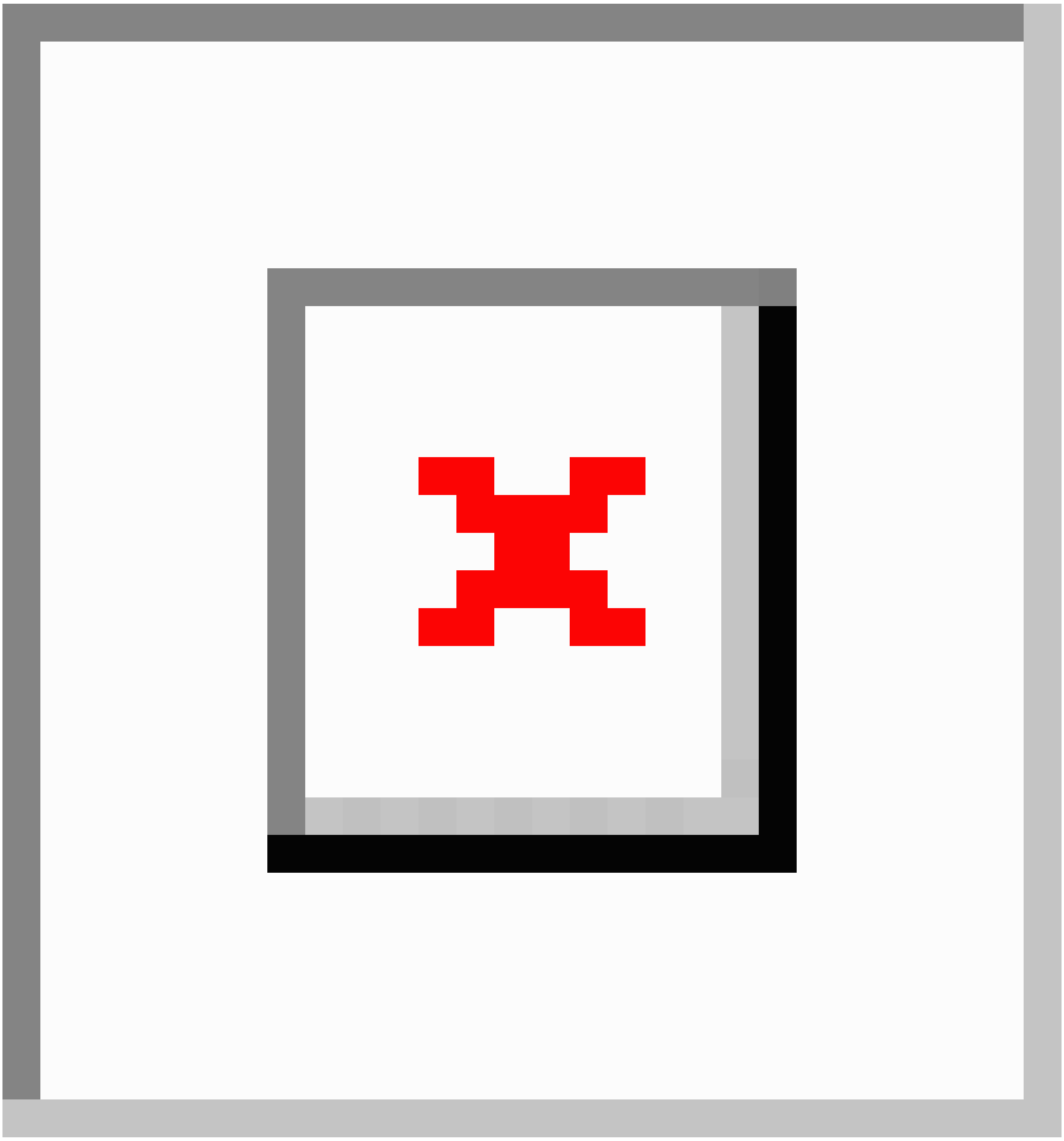
Table 3 displays pairwise comparisons for each level within the Wales condition. Callers stayed on the phone for 174 seconds longer if they were put through to a branch in England than if they were put through to a branch in Wales (difference [seconds] = −175.23; P<.001). Callers stayed on the phone for 168 seconds longer if they were put through to a branch in NI than if they were put through to a branch in Wales (difference [seconds] = −167.93; P<.001). Callers stayed on the phone for 138 seconds longer if they were put through to a branch in Scotland than if they were put through to a branch in Wales (difference [seconds] = −137.04; P<.001). Pairwise comparisons of average call duration from Wales landlines to Wales branches versus other regions were not significant for all other comparisons (Table 3).

As shown in Figure 3, there was an increase in variation across all Wales conditions compared with previous regional conditions. There was a higher density of calls with a shorter duration for the Wales-Wales condition. In terms of the calls with longer duration, there was much more variation between conditions. From lowest to highest density for longer duration calls, Wales-Wales had the lowest density of longer duration calls, followed by Wales-Scotland, Wales-NI, and Wales-England. Figure 3 also displays the average call duration between Wales to Wales and other regions; from shortest to longest, Wales-Wales (mean call duration 835 seconds), then Wales-Scotland (mean call duration 972 seconds), Wales-NI (mean call duration 1003 seconds), and finally Wales-England (mean call duration 1010 seconds).

Table . Pairwise comparisons of average call duration from Wales landlines to Wales branches versus other regions.

Group	Differences between means (duration in seconds)	*P* value
Wales_NI-Wales_England	−7.3	.8
Wales_Scotland-Wales_England	−38.2	.2
Wales_Wales-Wales_England	−175.23	<.001
Wales_Scotland-Wales_NI	−30.89	.46
Wales_Wales-Wales_NI	−167.93	<.001
Wales_Wales-Wales_Scotland	−137.04	<.001

Figure 3. Density plot of call durations from Wales to Wales versus other regions (log scale 10) and average call durations (95% CI).



NI to NI Versus Other Regions

Levene's test of homogeneity of variance assumption was violated ($F_{3,8626}$=8.8469; P<.001). A one-way ANOVA or one-way analysis of means was carried out to uncover any significant differences in average call duration between interregional and intraregional calls. There was a significant difference between interregional and intraregional calls for NI ($F_{3,2078}$=20.40; P<.001).

Table 4 displays pairwise comparisons for each level within the NI condition. Callers stayed on the phone for 143 seconds longer if they were put through to a branch in England than if they were put through to a branch in in NI (difference [seconds] = −142.67; P<.001). Callers stayed on the phone for 191 seconds longer if they were put through to a branch in Scotland than if they were put through to a branch in NI (difference [seconds] = 191.47; P<.001). Callers stayed on the phone for 176 seconds longer if they were put through to a branch in Wales than if they were put through to a branch in NI (difference [seconds] = 175.74; P<.001). Pairwise comparisons of average call durations from Wales landlines to Wales branches versus other regions were not significant for all other comparisons.

Figure 4 shows considerable variation between the NI conditions. There appears to be variation between conditions for shorter-duration calls; NI-NI had a higher density of calls with a shorter duration than the other conditions, followed by NI-England, NI-Wales, and with NI-Scotland having the lowest density of shorter-duration calls. NI-Scotland and NI-Wales had similar density of longer-duration calls, closely followed by NI-England; NI-NI had the lowest density of longer-duration calls. Figure 4 also displays the average talk duration between NI-NI and other regions: from shortest to longest, NI-NI (mean call duration 692 seconds), then NI-England (mean call duration 834 seconds), NI-Wales (mean call duration 867 seconds), and finally NIScotland (mean call duration 883 seconds).

The aim of this study was to explore any potential differences in call duration between interregional calls and intraregional calls within the United Kingdom made to a national crisis helpline. This study analyzed Samaritans UK landline-based calls (N=4,708,205 calls) that were made to the service from January 2015 until August 2018. The origin of the calls was determined based on the landline prefixes that formed the area code digits of an anonymized caller identifier enabling the categorization of calls by region of origin.

The destination regions were determined by categorizing the destination branches, which allowed for comparisons of call duration from region to region (Figure 5).

Table . Pairwise comparisons of average call duration from Northern Ireland (NI) landlines to NI branches versus other regions.

Group	Differences between means (duration in seconds)	*P* value
NI_NI-NI_England	−142.67	<.001
NI_Scotland-NI_England	48.8	.23
NI_Wales-NI_England	33.07	.52
NI_Scotland-NI_NI	191.47	<.001
NI_Wales-NI_NI	175.74	<.001
NI_Wales-NI-Scotland	−15.73	.075

Figure 4. Density plot of call durations from NI to NI versus other regions (log scale 10) and average call durations (95% CI).

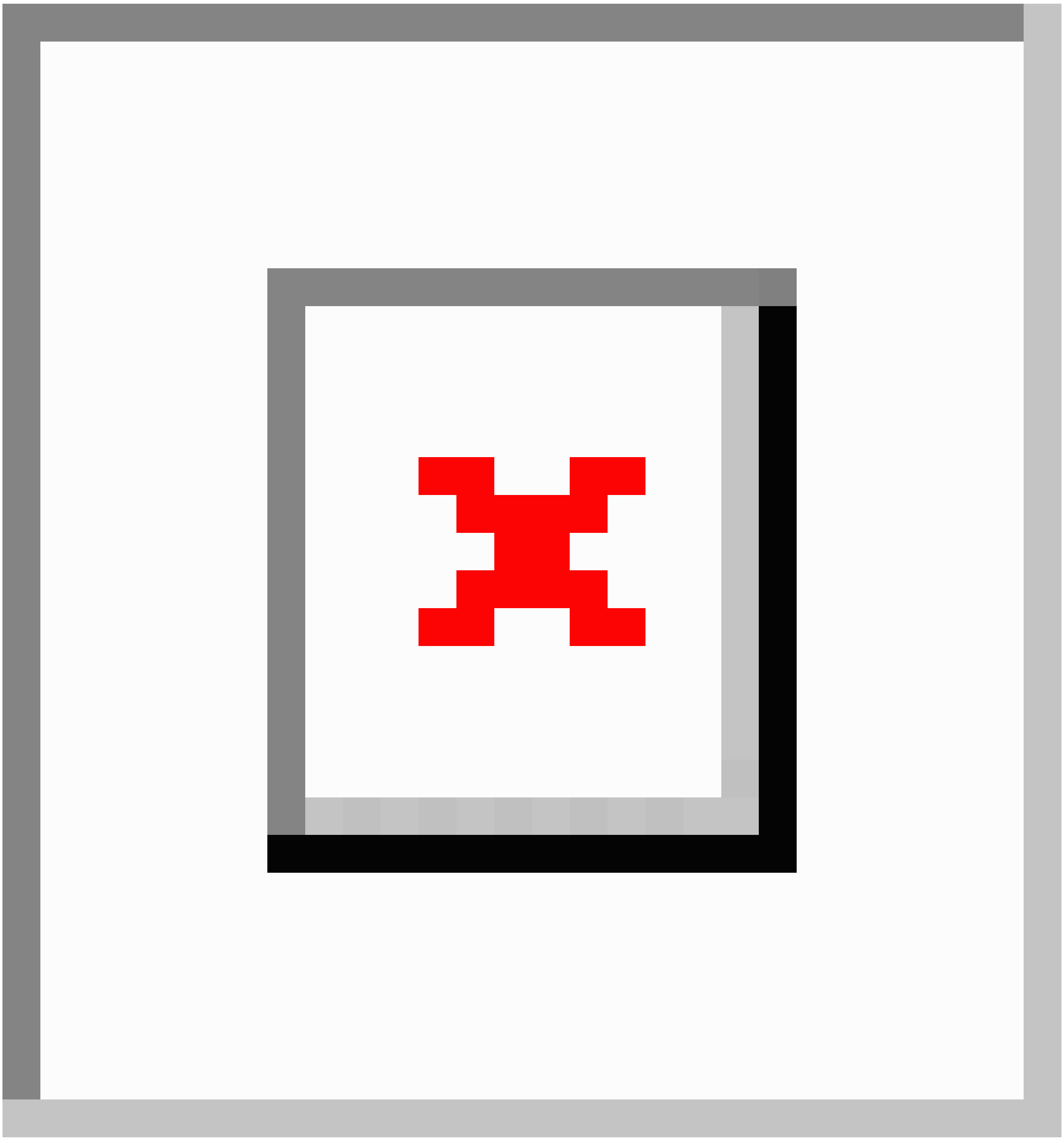
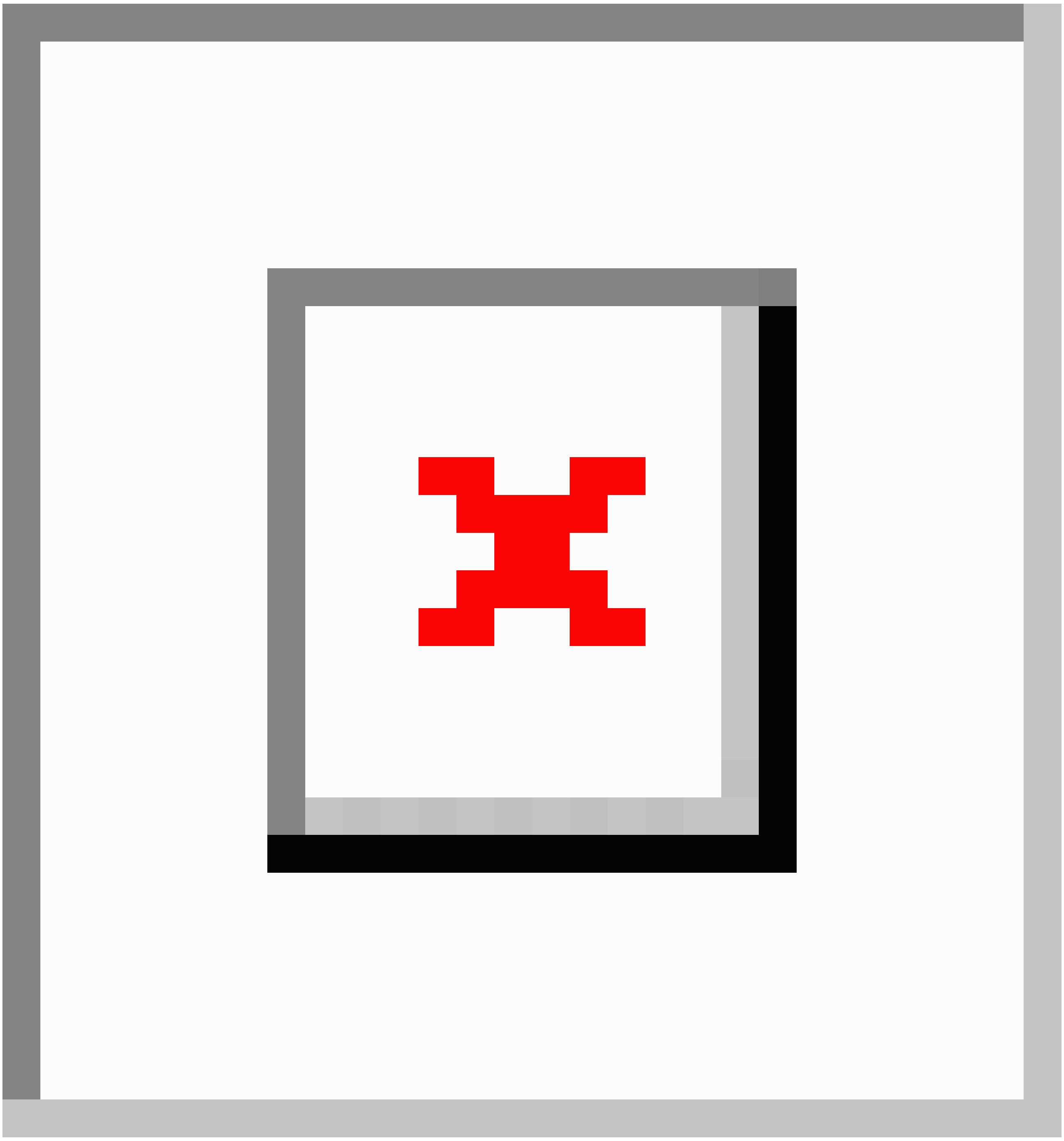


Figure 5. Mean durations of calls between the constituent regions of the United Kingdom.



Discussion

Principal Findings

Until now, this is an aspect of caller behavior that had not been examined in relation to how callers interact with crisis helplines. Statistical analyses showed that there were significant differences between interregional and intraregional calls. On average, callers to crisis helplines stayed on the phone for a shorter amount of time if they were routed to a branch within the same region in which the call originated than if they were routed to a branch in a different region of origin. Across all regions, there was a higher density of shorter-duration calls for intraregional conditions than in interregional conditions. Conversely, intraregional conditions had the lowest density of longer-duration calls than interregional conditions. Both findings were consistent across all regional conditions.

It is important to point out that the reasons behind different call durations for interregional calls are unknown. It may be the case that callers may feel more encouraged to engage in lengthy conversations and disclose more information about their crisis to a volunteer within a different region due to having a heightened sense of anonymity. While it is the role of the volunteer to have a nonjudgmental approach and uphold confidentiality regarding call information, callers may perceive that there is a risk that they know the volunteer, or that they might have mutual contacts.

Another possible explanation for the significant differences in talk time between regions is that there could be variations in accents within interregional calls. All intraregional call conditions had the shortest mean call duration, had a higher density of shorter-duration calls, and had the lowest density of longer-duration calls. Callers and volunteers may be better able to understand each other if they both speak in the same accent. In contrast, callers and volunteers who are both from different regions may take longer to process the accent of one another. Indeed, studies have examined the impairments of language processing due to foreign accents. Clark and Garrett [14] found that perceptual processing speed of listeners was much slower by an average of 100 - 150 milliseconds when sentences produced in the English language were spoken with accented speech (ie, Spanish or Chinese accent) than if spoken in nonaccented speech (ie, native speech). In a series of experiments, Floccia and colleagues [15] concluded that regional accent normalization is exemplified by an initial temporary perturbation in speech processing, resulting in longer reaction times in detecting words spoken in an unfamiliar accent. The authors also state that the differences could also be down to basic differences in stimuli across different accents, meaning that a speaker with a different accent to that of the listener may speak at a various rate. However, this explanation is just one possibility, and this assumes that volunteers have regional accents akin to that area and to date we do not have evidence to confirm that is the case. Another important consideration is the effect of accents more generally, outside of understanding speech or language processing, as it may also relate to how easily the caller identifies with the volunteer and vice versa.

Ultimately, further work should seek to understand the reasons behind differing call durations. For example, future work could include a qualitative analysis with Samaritans volunteers across these regions to explore their views and with callers to explore their thoughts on using the service. Questions for future research include the following: Do callers make a judgment as to a listener's location (eg, based on accent or local knowledge, or other)? Do callers feel they can identify more with listeners from the same location/with the same accent/who share local knowledge? Does this perception (of location and/or identity) influence willingness to disclose/caller openness/satisfaction with service?

Limitations

It is not possible to determine the cause of the differences that have been found with any certainty from these data. Some crisis helplines will record qualitative data regarding the content of the call, such as the main presenting reason that was cited by the caller. Additional information such as this may provide important context and some idea as to call complexity. It may be intuitive to think that if a call is complex in nature, that the call will last longer. However, a limitation to this study is that such qualitative information was not available to complement the analysis of the call log data.

It was possible to conduct only the analysis on calls made from landlines within the call log data. In total, while there were 25,177,944 calls within the data set provided by Samaritans UK, only landline calls (18.7% of all calls) were analyzed for this study. Most calls were made from a mobile telephone and it was not possible to determine the location from where these calls were made. Therefore, findings from this study may not be generalizable to all callers who contact crisis helplines as the analysis was limited to landline callers only. In addition, calls in Wales to the Welsh language line were not explored during this study.

Implications

The findings from this study could have practical applications which may allow crisis helplines to manage their resources more effectively and improve caller satisfaction with the service. In the interests of freeing up service capacity, intraregional routing could potentially allow for more callers to get in contact with a volunteer over time. However, it may be that shorter calls result in less disclosure and thus callers may not prefer this. It may also be the case that callers prefer to speak to a volunteer from the same region, which may promote better understanding by the volunteer of the issues that underpin the caller's crisis. Qualitative research is warranted to determine caller preferences regarding whether they are routed to a branch within or in a different region from themselves. While results achieved in this study were statistically significant, consideration must be taken by the relevant stakeholders into determining whether these results are, in fact, practically significant. While some statistically significant results yielded a difference upward of 180 seconds between regions, there needs to be careful consideration from crisis helplines who operate a multiregional routing service, as to whether intraregional routing of calls is a cost-effective solution to streamlining the service so that more calls from more callers can be answered at any one time; however, this would have to be balanced against service quality and caller satisfaction.

Conclusions

This study presents a temporal analysis of more than 4.6 million landline phone calls made to a UK national crisis helpline. The aim was to determine whether callers stayed on the phone for a longer or shorter time depending on whether they were routed to a branch within the same region of origin of their call or in a different region. The study found that across all conditions, callers stayed on the phone for a longer period on average if routed to a branch that was within a different region from which the call originated. Intraregional call conditions had a higher density of shorter calls than interregional call conditions; this finding was also consistent across all conditions. Potential explanations for this finding may be that callers may feel more anonymous if routed to an interregional branch and may disclose more, which leads to longer call times, or it may be the case that callers and volunteers may experience some difficulty in understanding nonnative-accented speech which also results in longer call duration. Another possible reason is that callers find it harder to connect with or identify with the volunteer and thus the calls take longer. While these explanations are speculative, further research is warranted to determine whether intraregional or interregional branches are preferred by callers to national crisis helplines.

Conflicts of Interest

None declared.

References

1. Coveney CM, Pollock K, Armstrong S, Moore J. Callers' experiences of contacting a national suicide prevention helpline: report of an online survey. Crisis 2012 Jul;33(6):313-324. [doi: 10.1027/0227-5910/a000151] [Medline: 22759662]
2. Bassilios B, Harris M, Middleton A, Gunn J, Pirkis J. Characteristics of people who use telephone counseling: findings from secondary analysis of a population-based study. Adm Policy Ment Health 2015 Sep;42(5):621-632. [doi: 10.1007/s10488-014-0595-8] [Medline: 25234346]
3. Kalafat J, Gould MS, Munfakh JLH, Kleinman M. An evaluation of crisis hotline outcomes. Part 1: nonsuicidal crisis callers. Suicide Life Threat Behav 2007 Jun;37(3):322-337. [doi: 10.1521/suli.2007.37.3.322] [Medline: 17579544]
4. Sawyer JB, Jameton EM. Chronic callers to a suicide prevention center. Suicide Life Threat Behav 1979;9(2):97-104. [Medline: 483356]
5. Spittal MJ, Fedyszyn I, Middleton A, et al. Frequent callers to crisis helplines: who are they and why do they call? Aust N Z J Psychiatry 2015 Jan;49(1):54-64. [doi: 10.1177/0004867414541154] [Medline: 24972602]
6. Turkington R, Mulvenna M, Bond R, et al. Behavior of callers to a crisis helpline before and during the COVID-19 pandemic: quantitative data analysis. JMIR Ment Health 2020 Nov 6;7(11):e22984. [doi: 10.2196/22984] [Medline: 33112759]
7. Mishara BL, Chagnon F, Daigle M, et al. Which helper behaviors and intervention styles are related to better short-term outcomes in telephone crisis intervention? Results from a silent monitoring study of calls to the U.S. 1-800-SUICIDE Network. Suicide Life Threat Behav 2007 Jun;37(3):308-321. [doi: 10.1521/suli.2007.37.3.308] [Medline: 17579543]
8. Grigorash A, O'Neill S, Bond R, Ramsey C, Armour C, Mulvenna MD. Predicting caller type from a mental health and well-being helpline: analysis of call log data. JMIR Ment Health 2018 Jun 11;5(2):e47. [doi: 10.2196/mental.9946] [Medline: 29891472]
9. O'Neill S, Bond RR, Grigorash A, Ramsey C, Armour C, Mulvenna MD. Data analytics of call log data to identify caller behaviour patterns from a mental health and well-being helpline. Health Informatics J 2019 Dec;25(4):1722-1738. [doi: 10.1177/1460458218792668] [Medline: 30222034]
10. Turkington R, Bond R, Mulvenna M, Potts C, O'Neill S, Armour C. User archetype discovery by cluster analysis of caller log data: tenure evolution is stable as time period reduces. Presented at: ECCE 2019: Proceedings of the 31st European Conference on Cognitive Ergonomics; Sep 10, 2019; Belfast, Northern Ireland, United Kingdom. [doi: 10.1145/3335082.3335090]
11. Scottish Government. Creating Hope Together: Suicide Prevention Strategy 2022 to 2032: Scottish Government; 2022. URL: https://www.gov.scot/binaries/content/documents/govscot/publications/strategy-plan/2022/09/creating-hope-together-scotlands-suicide-prevention-strategy-2022-2032/documents/creating-hope-together-scotlands-suicide-prevention-strategy-2022-2032/creating-hope-together-scotlands-suicide-prevention-strategy-2022-2032/govscot%3Adocument/creating-hope-together-scotlands-suicide-prevention-strategy-2022-2032.pdf [accessed 2024-08-19]
12. Zareian B, Klonsky ED. Connectedness and suicide. In: Alternatives to Suicide: Beyond Risk and Toward a Life Worth Living: Elsevier; 2020:135-158. [doi: 10.1016/B978-0-12-814297-4.00007-8]
13. Wickham H. Ggpolt2 Elegant Graphics for Data Analysis Use R! Ser: Springer; 2016. URL: https://link.springer.com/book/10.1007/978-3-319-24277-4 [accessed 2024-08-31]
14. Clarke CM, Garrett MF. Rapid adaptation to foreign-accented English. J Acoust Soc Am 2004 Dec;116(6):3647-3658. [doi: 10.1121/1.1815131] [Medline: 15658715]
15. Floccia C, Goslin J, Girard F, Konopczynski G. Does a regional accent perturb speech processing? J Exp Psychol Hum Percept Perform 2006 Oct;32(5):1276-1293. [doi: 10.1037/0096-1523.32.5.1276] [Medline: 17002537]

Abbreviations

UK: United Kingdom
NI: Northern Ireland

Edited by J Torous; submitted 07.03.24; peer-reviewed by A Hosny, CY Hsu, WY Hsu; revised version received 03.06.24; accepted 12.06.24; published 19.09.24.

Please cite as:
Turkington R, Potts C, Mulvenna M, Bond R, O'Neill S, Ennis E, Hardcastle K, Scowcroft E, Moore C, Hamra L
Talk Time Differences Between Interregional and Intraregional Calls to a Crisis Helpline: Statistical Analysis
JMIR Ment Health 2024;11:e58162
URL: https://mental.jmir.org/2024/1/e58162
doi:10.2196/58162

Using Large Language Models to Understand Suicidality in a Social Media–Based Taxonomy of Mental Health Disorders: Linguistic Analysis of Reddit Posts

Brian Bauer[1,*], PhD; Raquel Norel[2,*], PhD; Alex Leow[3,4], MD, PhD; Zad Abi Rached[5]; Bo Wen[2], PhD; Guillermo Cecchi[2], PhD

1

2

3

4

5

*these authors contributed equally

Corresponding Author:
Brian Bauer, PhD

Abstract

Background: Rates of suicide have increased by over 35% since 1999. Despite concerted efforts, our ability to predict, explain, or treat suicide risk has not significantly improved over the past 50 years.

Objective: The aim of this study was to use large language models to understand natural language use during public web-based discussions (on Reddit) around topics related to suicidality.

Methods: We used large language model–based sentence embedding to extract the latent linguistic dimensions of user postings derived from several mental health–related subreddits, with a focus on suicidality. We then applied dimensionality reduction to these sentence embeddings, allowing them to be summarized and visualized in a lower-dimensional Euclidean space for further downstream analyses. We analyzed 2.9 million posts extracted from 30 subreddits, including r/SuicideWatch, between October 1 and December 31, 2022, and the same period in 2010.

Results: Our results showed that, in line with existing theories of suicide, posters in the suicidality community (r/SuicideWatch) predominantly wrote about feelings of disconnection, burdensomeness, hopeless, desperation, resignation, and trauma. Further, we identified distinct latent linguistic dimensions (well-being, seeking support, and severity of distress) among all mental health subreddits, and many of the resulting subreddit clusters were in line with a statistically driven diagnostic classification system—namely, the Hierarchical Taxonomy of Psychopathology (HiTOP)—by mapping onto the proposed superspectra.

Conclusions: Overall, our findings provide data-driven support for several language-based theories of suicide, as well as dimensional classification systems for mental health disorders. Ultimately, this novel combination of natural language processing techniques can assist researchers in gaining deeper insights about emotions and experiences shared on the web and may aid in the validation and refutation of different mental health theories.

(JMIR Ment Health 2024;11:e57234) doi:10.2196/57234

KEYWORDS

natural language processing; explainable AI; suicide; mental health disorders; mental health disorder; mental health; social media; online discussions; online; large language model; LLM; downstream analyses; trauma; stress; depression; anxiety; AI; artificial intelligence; explainable artificial intelligence; web-based discussions

Introduction

Suicide rates have increased by 35% since 1999, and suicide remains a leading cause of death in the United States [1]. Despite concerted efforts, our ability to predict, explain, or treat suicide risk has not significantly improved over the past 50 years [2,3]. Thus, a top public health priority is understanding factors that contribute to suicide risk. Recent meta-analytic work, comprising suicide risk factor research over the past 50 years (using 365 studies), found that no single set of risk factors (eg, mood disorders or impulsivity) accurately predict future suicidal thoughts and behaviors [2]. Past risk factor studies have been limited by (1) potential sampling biases (eg, overrepresentation of clinical populations), (2) structured clinical interviews and surveys, and (3) laboratory-based (rather than naturalistic) settings. With the rapid increase in the use of web-based platforms, such as Reddit, people experiencing mental health symptoms have new outlets for sharing experiences, seeking support, and engaging in discussion regarding their mental health. Platforms such as Reddit provide unique opportunities for studying the experiences and perspectives of individuals at risk of suicide in the context of other mental pathologies and stressors [4-6]. To overcome previous limitations in suicide risk factor research, this study aims to analyze posts from a web-based community dedicated to providing support for individuals in crisis (ie, the r/SuicideWatch subreddit), to involve individuals who may not present for mental health studies or disclose their suicide risk and to obtain more nuanced insights into suicidality from the naturalistic and open-ended nature of anonymous web-based forums.

Understanding the factors contributing to suicidality is crucial for developing effective prevention strategies and interventions. Prominent theories of suicide—such as the Interpersonal Theory of Suicide (ITS) [7], Three-Step Theory (3ST) [8], and Integrated Motivational–Volitional (IMV) model [9]—are referred to as "ideation-to-action" frameworks. These theories attempt to explain how people develop suicidal ideation and transition to suicidal behaviors. Several common variables among ideation-to-action frameworks include feeling disconnected and burdensome to others, feelings of entrapment and hopelessness, and factors that may increase peoples' capability to die by suicide (eg, traumatic experiences). Similar to many other psychological theories, they have been developed through researchers observing data patterns and testing their hypotheses (mainly) through self-reported survey data. Analyzing linguistic patterns in Reddit posts may provide an avenue for suicide theory exploration, confirmation, and refutation for these ideation-to-action frameworks, which could significantly impact future and existing intervention targets and assessment practices.

The use of natural language processing (NLP) with machine learning (ML) to gain new insights into mental health topics has increased dramatically over the last decade [10]. In mental health, this approach has mainly been used to confirm existing hypotheses through extracting meaning from texts (NLP) and then classifying these extractions (ML); however, this combination approach can be equally useful for exploration and discovery [10]. Specifically in suicide research, NLP and ML have primarily been used to help improve the accuracy of suicide risk identification [11]. However, NLP combined with ML is less frequently used in both mental health and suicide research to derive theoretical perspectives. Newer large language models (LLMs) that use Bidirectional Encoder Representations from Transformers (BERT) allow researchers to capture more complexities in human language than previous approaches, which are ideal for discovery as well as for testing directional hypotheses. Furthermore, the recent advances in explainable artificial intelligence (XAI) can be applied to NLP to help improve transparency, trustworthiness, and understanding of results in the context of mental health [12]. Using these modern techniques in tandem may help provide critical insights into suicide risk. The primary aim of this study is to analyze the content of posts from the r/SuicideWatch subreddit as well as mental health–related and non–mental health–related subreddits, with the goal of contributing to our understanding of suicide risk to ultimately improve prevention strategies and interventions. Although there are several other anonymous platforms available for individuals to discuss suicide, we chose Reddit due to the size of its userbase (over 400,000 members), which may aid the generalizability of our findings, and for practical reasons—namely, data availability. For this, we used LLMs to produce numerical representations of these posts—called *embeddings* [13]—which may reveal *unique* suicidality linguistic patterns; we also used recent developments in generative LLMs [14] and XAI [15], which turn abstract embeddings into natural language text, to identify connections with theories of suicidal behavior.

Our study was primarily data driven, as we used *generic* LLM embeddings of the posts and only applied theoretical constructs for post hoc interpretation. This means that the numerical representation of the posts was based on how the sentences in them are related to sentences in very large text corpora used for training the LLM that cover vast swaths of topics.

Methods

Ethical Considerations

Reddit users are made aware that their posts are publicly accessible through Reddit's Terms and Conditions. No personal identifying information (eg, names, locations, or IP addresses) were collected. Further, none of the authors participated in any discussions; thus, it was not necessary to inform users that their posts may be used for research. Because the collected data set is publicly available and already deidentified, the University of Georgia Human Subjects Office reviewed the submission and assigned a determination of "Not Human Research."

Data Procurement, Selection, and Preprocessing

We downloaded posts from a list of subreddits from October 1 to December 31, 2022, and from October 1 to December 31, 2010, using The-Eye.eu [16], which contains an archive of Reddit's full submission until December 2022. We used Python (Python Software Foundation) to process the data. Posts that were removed or deleted were not used. Empty entries or entries containing just a "?" were not considered. We analyzed 16 subreddits related to mental health and 14 subreddits not related to mental health to serve as controls. In Table 1, we compared the number of posts and words per post by 2 subreddit groups (mental health and controls). We saw that the number of posts was similar in both groups, but the range and variability (ie, SDs for the number of words and words per post) were much higher in the mental health subreddits. We noted that for the 2010 data, in general, the posts were much longer than those from 2022 data; in the 2022 data, some of the posts were just a single word (eg, "Pls," "Yuh," and "Help"), whereas for older data (2010), the posts were much longer.

Table . Statistical descriptors of the 16 subreddits related to mental health and the 14 subreddits used as controls.

Subreddits and statistics		2010 data		2022 data	
		Mean (SD)	Range	Mean (SD)	Range
Mental health					
	Posts per subreddit	246 (225)	14-785	182,000 (240,000)	15,900-928,000
	Words per subreddit	409,000 (415,000)	42,000-1,600,000	34,700,000 (48,800,000)	2,700,000-189,000,000
	Words per post	1700 (446)	1000-3000	183 (31)	112-225
Controls					
	Posts per subreddit	—[a]	—	195,000 (243,000)	9700-831,000
	Words per subreddit	—	—	21,900,000 (28,300,000)	1,500,000-95,000,000
	Words per post	—	—	114 (27)	72-163

[a]Not applicable.

Linguistic Analysis and Interpretation

Overall Approach

After obtaining the posts, we followed these steps: (1) represented the posts in the latent space of LLM embeddings; (2) computed a representation for each subreddit, averaging all its posts; (3) computed different metrics of the structure of the subreddits in the embedding space; and (4) applied interpretation techniques to these structural metrics to obtain insights about the relationship between suicidality and other self-identified groups.

LLM Embeddings

Semantic text embedding is an NLP technique used to represent the meaning of text in numerical form. It accounts for the context of words or phrases rather than just their individual representations. By using advancements in text embedding, we used a more precise method in NLP using BERT [17]. Text embedding assigns a numerical vector to each text, enabling texts with similar contexts to be closer in the vector space. This allows us, using mathematical tools, to better understand and analyze the semantic similarities and differences between different texts. To represent a subreddit, we compute the centroid (average) of the embeddings of all the posts assigned to it. As mentioned above, these embeddings are *unbiased*: we did not use the metadata related to the subreddit provenance nor applied any theory-driven construct [18].

Subreddits Structure

To obtain a measure of similarity between subreddits, for each post, we computed a "linguistic label"—the label of the nearest centroid in the embedding space. Then, for each subreddit, we computed the percentage of posts that were assigned to any "linguistic label," including the original label; the proportion of posts "linguistically assigned" to a subreddit other than the original one is a measure of the similarity between them.

Hierarchical Clustering

Hierarchical clustering operates on the principle of iteratively merging the closest pair of clusters, where the definition of "closeness" varies according to different linkage criteria. We used the Ward linkage method, where the distance between 2 clusters is the increase in the summed square distance from each point to the centroid of its cluster after merging the clusters. A dendrogram (a tree-like diagram) visually represents the process and results of hierarchical clustering. Each leaf corresponds to a data point, and branches represent the successive merging of clusters, with the height of each merge proportional to the distance between the combined clusters. By examining a dendrogram, one can intuitively grasp the data's structure and decide on an appropriate number of clusters by cutting the tree at a specific height. This visual tool aids in interpreting the complex relationships and nested structures within the data, offering insights into the underlying patterns and groupings. We performed hierarchical clustering on the embedding representative of the subreddit, that is, the average of all the embedding vectors of the subreddit.

Dimensionality Reduction

We factorized the subreddits centroids in the embedding space using singular value decomposition (SVD). We analyzed the relative location of the subreddits in the reduced representation of the first 3 SVD components, understood as latent semantic dimensions.

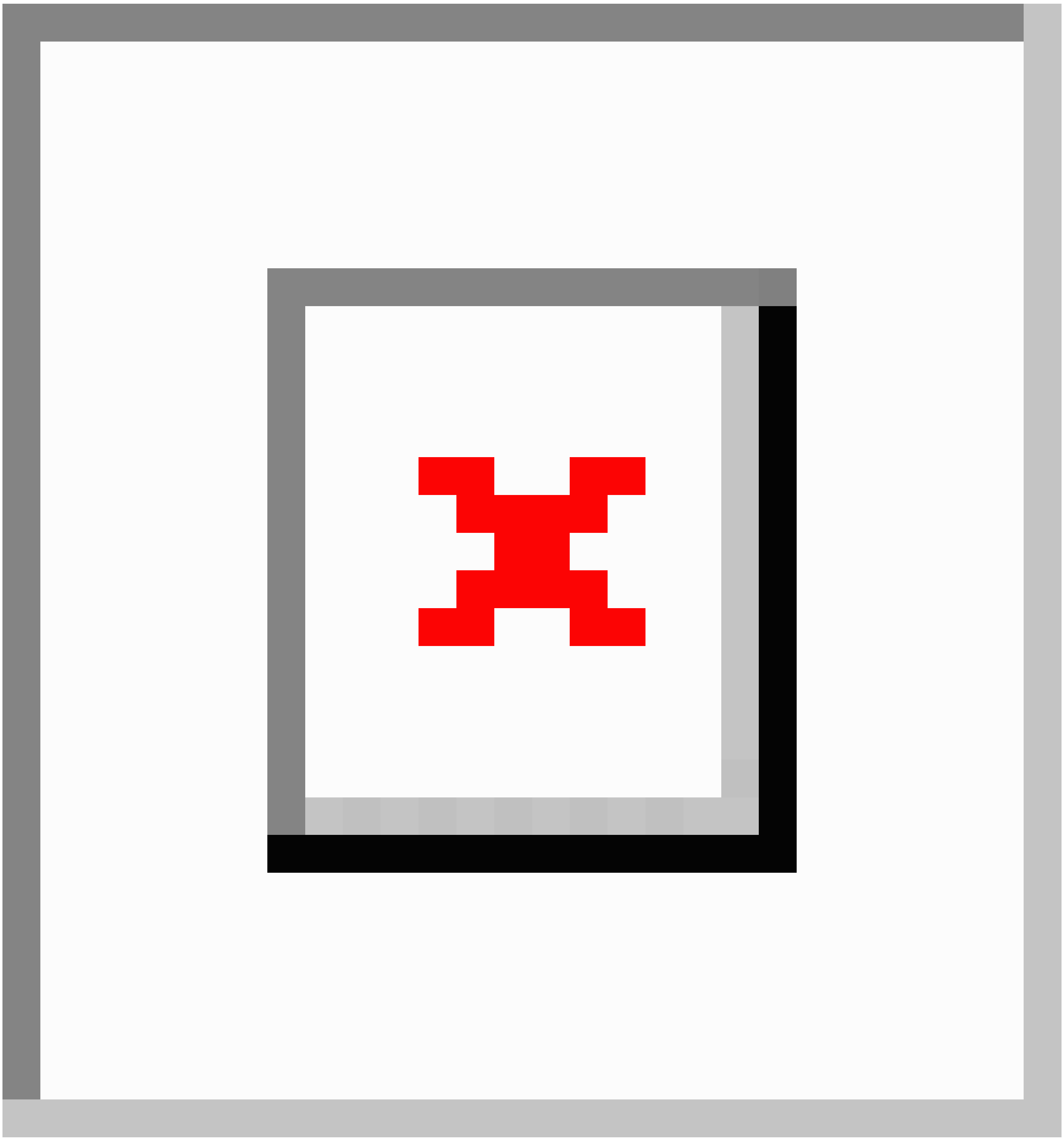
Generative LLM

We used a generative LLM to obtain insights into the latent patterns of the SVD components. In a Jupyter Notebook environment, we configured the *Langchain* library to use the GPT-4-0613 model (OpenAI) [14] with a zero-temperature setting, ensuring deterministic outputs for consistency in interpretation. We used the *ConversationChain* module coupled with *ConversationBufferMemory* to facilitate an interactive and memory-aware dialogue with GPT-4. This setup enabled us to iteratively query the model with our prompts and data, ensuring a contextually rich and coherent analysis of the subreddit postings. The use of a verbose mode in the conversation setup provided detailed logging of the model's responses, further aiding in the transparency and traceability of our analytical process. We identified posts with the top and bottom 5 embedding positions in the first 3 dimensions of SVD projections, totaling 30 extreme postings.

XAI Techniques

XAI techniques are designed to provide insights into the factors or features that contribute to an artificial intelligence system's outputs, allowing users to understand and validate the reasoning behind those decisions. With this in mind, we use ProtoDash [15], a technique used to choose representative examples that effectively represent the overall distribution of a data set, to help with the interpretation of the model. We fed the LLM embeddings and the associated posts from the r/SuicideWatch subreddit to ProtoDash and asked for the 5 most representative posts in the data set. Those selected prototypes of the data set were then fed to ChatGPT, to help obtain insight into the data. Figure 1 shows a visual overview of the study methodology.

Figure 1. Schematic description of the processing workflow in this paper. All Reddit postings were first fed into BERT to yield posting-level embeddings in a high-dimensional space, followed by dimensionality reduction (into 3D) using SVD. This procedure allows us to extract prototypical postings (using ProtoDash) as well as extreme postings along each of the 3 SVD axes, which were then fed into ChatGPT for semantic interpretations. Last, we also performed hierarchical clustering to recover the relational structure between different subreddits. BERT: Bidirectional Encoder Representations from Transformers; GPT: Generative Pre-trained Transformer; SVD: singular value decomposition.



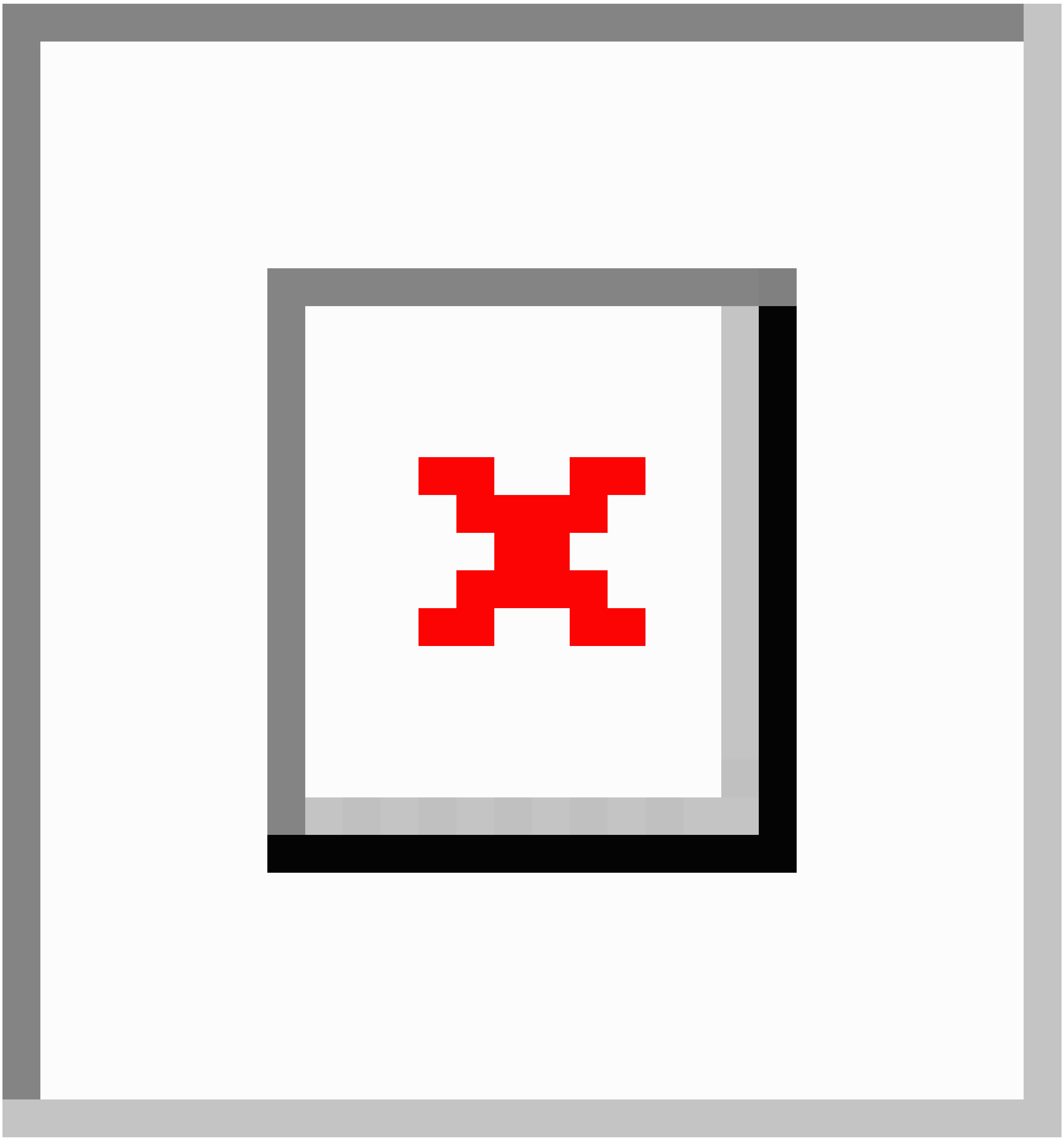
Results

Unless explicitly indicated, we presented results on the 2022 data; the 2010 data were used to ascertain the stability of the structural features we determined with the most recent and larger data set.

Verbosity in Mental Health Subreddits

The total number of posts and the verbosity (the average number of words per post in a subreddit) are shown in Figure 2. R/Depression is the most active of the mental health subreddits, and only behind the r/DnD (referring to Dungeons and Dragons) and r/gaming subreddits overall. The subreddits were sorted by verbosity (green line), which shows higher values for mental health posts as opposed to non–mental health posts. Moreover, within mental health subreddits, there are differences of more than 30% between the least verbose subreddits (r/schizophrenia and r/EDAnonymous [eating disorders anonymous]) and the high-verbosity subreddits (r/PTSD [posttraumatic stress disorder] and r/Depression).

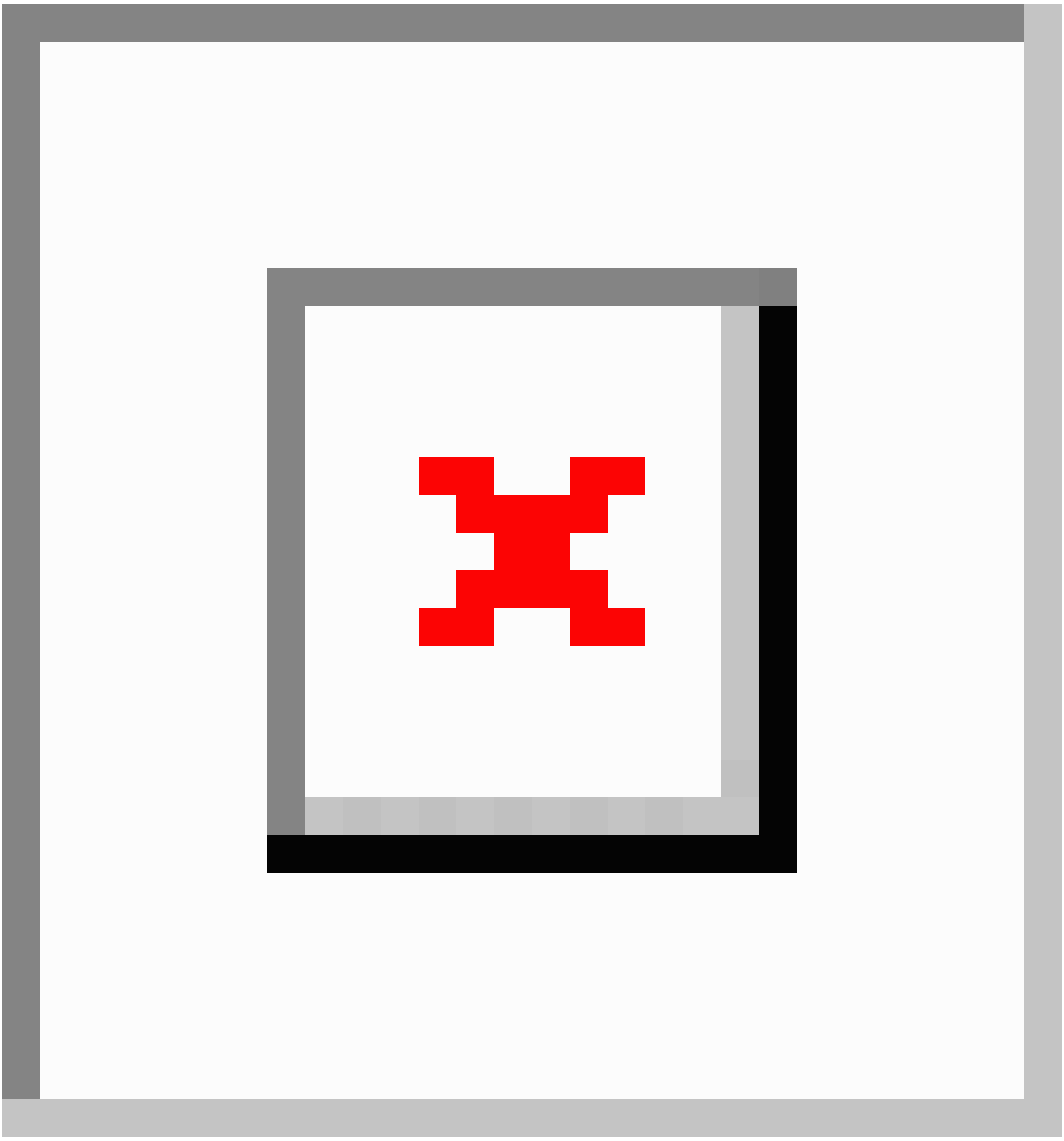
Figure 2. The total number of posts (bar heights) and verbosity (number of words per post; green line) of subreddits. The verbosity of MH subreddits (blue) is significantly higher than non-MH ones (orange). BPD: borderline personality disorder; DnD: Dungeons and Dragons; EDAnonymous: eating disorders anonymous; MH: mental health; PTSD: posttraumatic stress disorder.



Structure of the Linguistic Embedding Space

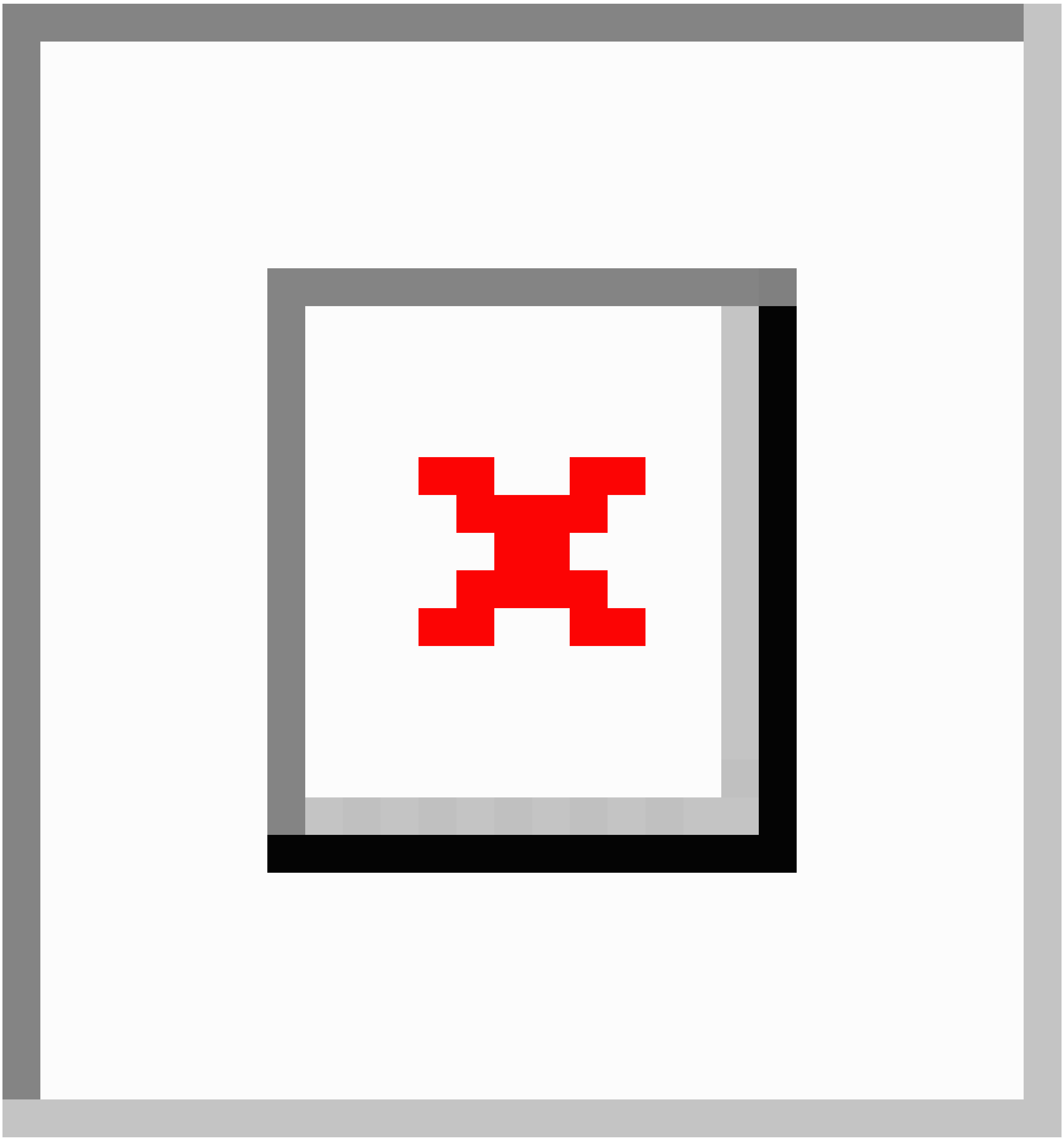
The plot in Multimedia Appendix 1 shows the measure of similarity to r/SuicideWatch for all subreddits. Almost half of the posts (234,406/479,321; 48.9%) from r/SuicideWatch were closer to their centroid than to any other centroid, with an additional 12.7% (60,854/479,321) being the closest to r/Depression, followed by r/BPD (borderline personality disorder) and r/SocialAnxiety. The extension of this approach to all subreddits is shown in Figure 3, which represents inter-subreddit similarities in the width of the links, thresholded at 7%. Besides the strong associations already present in Figure 2 (of r/SuicideWatch with r/Depression and r/BPD), there are associations of r/BPD with r/Depression; r/SocialAnxiety and r/PTSD; and several associations between the different anxiety-related subreddits, r/Psychosis and r/schizophrenia, the 2 bipolar subreddits, and r/addiction with r/alcoholism.

Figure 3. Proportion of posts from a given subreddit that are the "closest" to a different subreddit centroid. The r/SuicideWatch centroid is colored red, nodes connected to it are colored purple, and the rest of nodes are colored blue. The width of the edge is proportional to the number of posts that are the closest to each centroid. addition: r/addiction; alcoholism: r/alcoholism; anxiety: r/SocialAnxiety; bipolar: r/Bipolar; bipolarR: r/BipolarReddit; BPD: r/BPD (borderline personality disorder); depression: r/Depression; EDA: r/EDAnonymous (eating disorders anonymous); healthanx: r/HealthAnxiety; mhealth: r/MentalHealth; millness: r/mentalillness; psychosis: r/Psychosis; ptsd: r/PTSD (posttraumatic stress disorder); schiz: r/schizophrenia; socialanx: r/SocialAnxiety; suicide: r/SuicideWatch.



The dendrogram in Figure 4 represents the result of the hierarchical clustering of the centroid coordinates for all the mental health subreddits in 2022 and 2010. The horizontal axis representing the linkage distance illustrates and supports the sequential merging of clusters. Clusters that merge at lower distances are more akin; as the distance increases, the clusters amalgamate into broader categories. The cluster with the shortest distance is grouping the subreddits r/BipolarReddit and r/Bipolar. The following cluster consists of the subreddits that discuss mental illness (r/mentalillness) and mental health (r/MentalHealth). The third most similar cluster is formed by r/Psychosis and r/schizophrenia. R/Anxiety and r/HealthAnxiety cluster together (green), as do the subreddits regarding addiction (r/alcoholism and r/addiction; red). Importantly, the clustering of the 2022 and 2010 subreddits were highly consistent, supporting the notion that the posts by the different subreddits revolve around the same topics related to the mental health condition they identify with.

Figure 4. Dendrogram of mental health vectors, using Ward linkage of subreddit's embeddings. The horizontal axis represents the linkage distance; clusters that combine at shorter distances are more similar to each other; as the distance grows, these clusters join together to form larger, more general categories. BPD: borderline personality disorder; EDAnonymous: eating disorders anonymous; PTSD: posttraumatic stress disorder.



Interpretation of the Linguistic Embedding Space

To obtain insights into the meaning of these topics, we performed a SVD factorization of the embedding space and used interpretation techniques on the resulting factors. Figure 5 shows the relative location of all the subreddits in the space determined by the first 2 SVD components. With the exception of r/Mindfulness, which is close to the mental health subreddits, there is a clear separation of classes along the SVD1 dimension. We also observed that r/SuicideWatch was ranked the second highest in SVD2, suggesting that this dimension may contain patterns relevant to suicidality.

The result of this procedure is presented in Table 2, which can be summarized by the following labels and directionalities: SVD1, *Emotional Well-Being: Despair to Resilience*; SVD2, *Seeking Understanding and Support: Closing In to Reaching Out*; and SVD3, *Severity of Distress: Low to High*. Using these axis interpretations, we mapped the results from the second and third SVD projection only for the mental health subreddits, for better visualization and interpretation; Figure 6 shows only mental health cases, which fall within the left portion of Figure 5. A prominent feature of the 2 plots is that r/SuicideWatch mapped onto the high end of the *Understanding and Support (Reaching Out)* dimension, as well as on the high end of the *Severity of Distress (High Distress)* dimension in a completely data-driven way.

- Axis label prompt: “Identify the name of the axis spanned by these two groups of extreme postings given that if one extreme is hot, the other extreme is cold, then the axis is temperature.”
- High values prompt: “What are the similarities among the top (highest values on the axis) five postings.”
- Low values prompt: “What are the similarities among the bottom (lowest values on the axis) five postings.”

Figure 5. SVD factorization of the embedding space. With the exception of r/Mindfulness, all non-mental health subreddits have positive value on the first SVD component, whereas all mental health subreddits have negative value on the first SVD component. SVD: singular value decomposition.

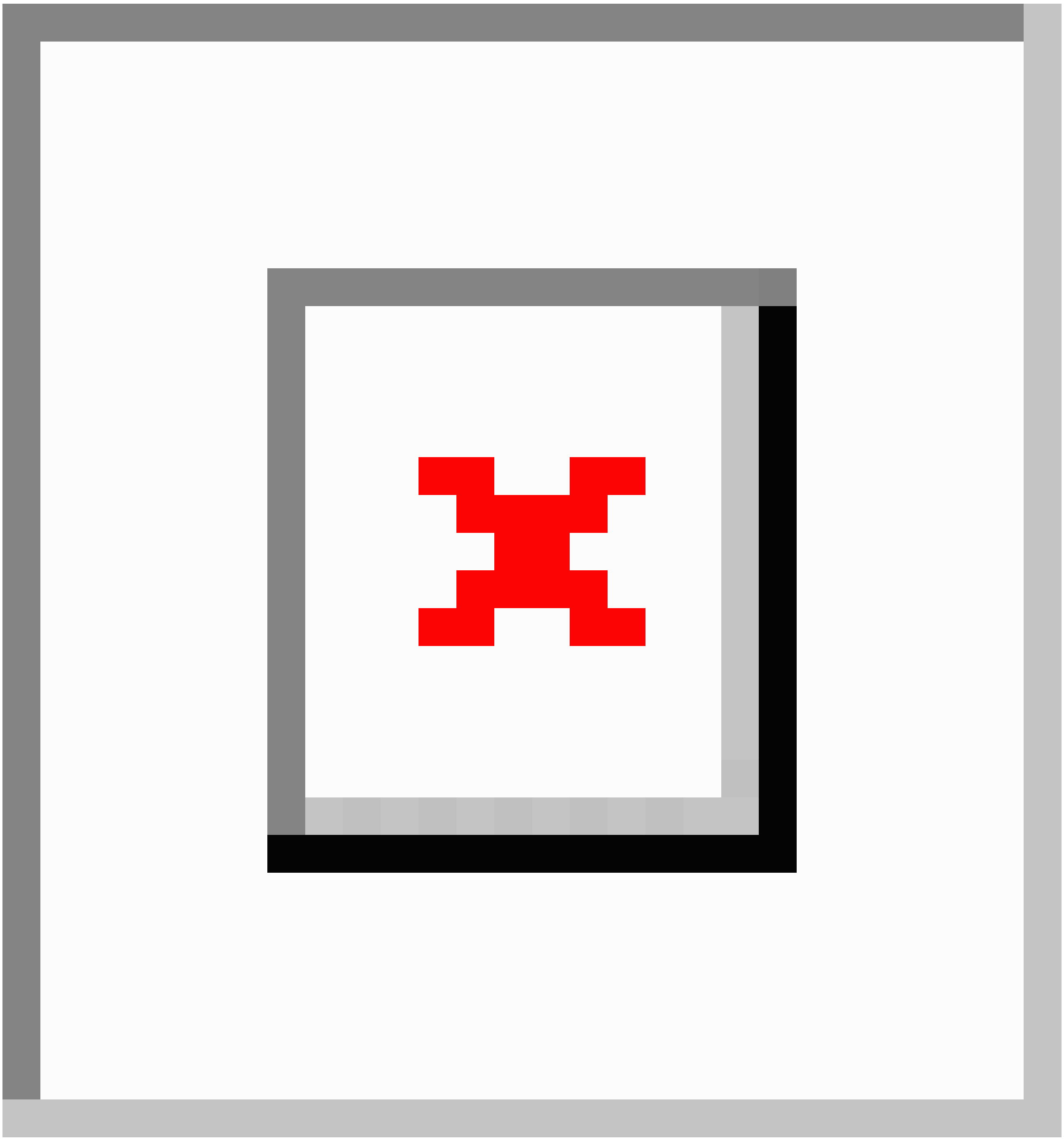
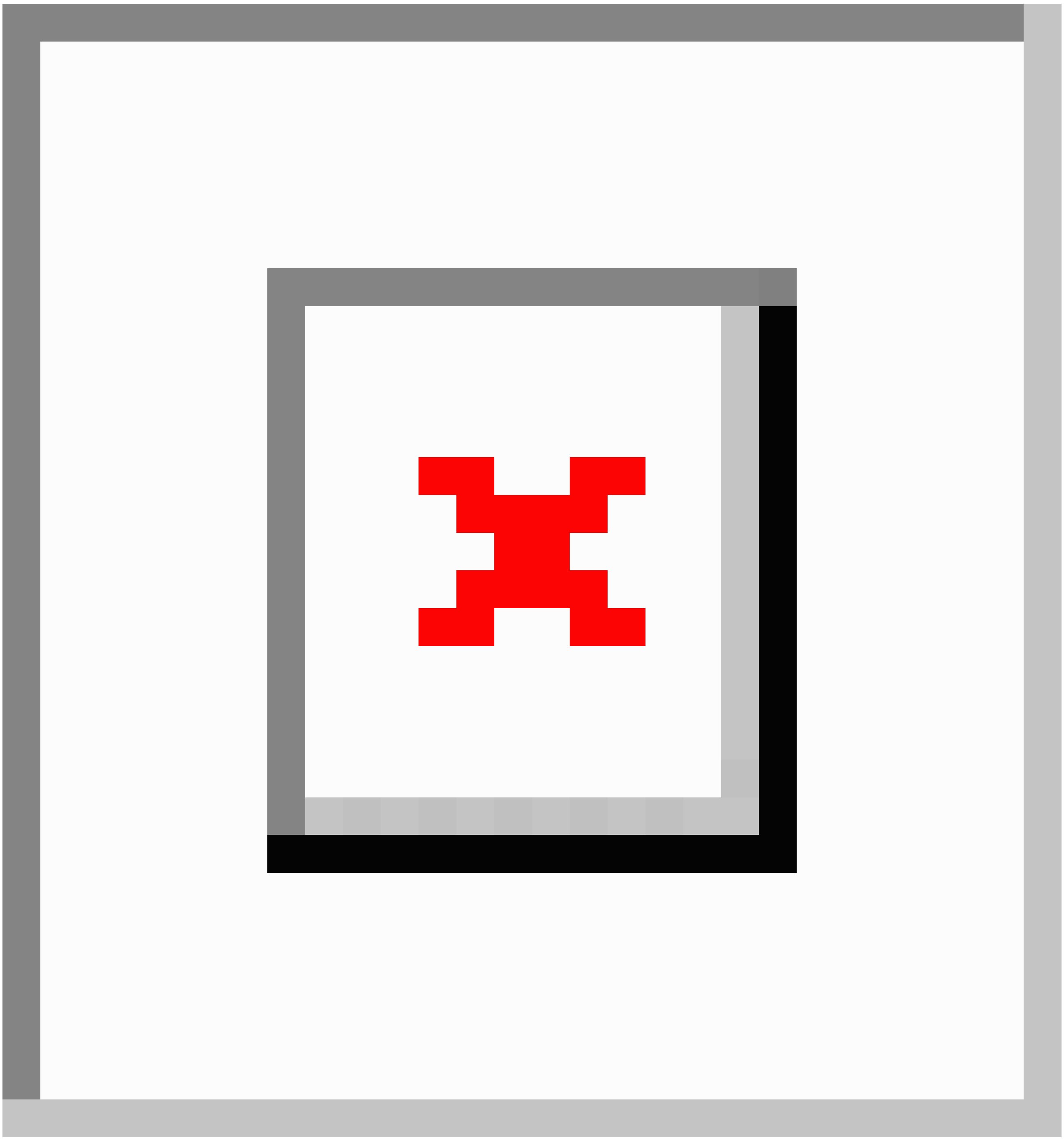


Table . Interpretation based on GPT-4 for prompts for the main 3 singular value decomposition (SVD) axes. Text from GPT-4 has been edited for grammar and clarity.

Values	Interpretation and axis label		
	SVD1, Emotional Well-Being: Despair to Resilience	SVD2, Seeking Understanding and Support: Closing In to Reaching Out	SVD3, Severity of Distress: Low to High
High	• Struggle and resilience • Experiencing significant anxiety, depression, and life changes • Learning to trust themselves and their abilities to handle their situations • Actively seeking help and trying to find ways to manage their mental health (eg, therapy, self-care, and relaxation techniques) • Making progress toward self improvement and happiness	• Seeking guidance or advice • Desire for information, advice, or validation • Reaching out for insights, recommendations, or shared experiences • This emotion is intertwined with feelings of uncertainty, curiosity, and a desire for understanding or improvement	• More intense and distressing content • Severe mental health struggles, including suicidal thoughts and feelings of extreme despair • Darker tone and more desperate
Low	• Narrative of struggle and despair • Feelings of sadness and emptiness • Express a sense of hopelessness about their situation • Previously sought help (eg, therapy and medication) but feel that these methods have been ineffective	• Feelings of internal conflict, frustration, and being overwhelmed by personal challenges • Deep sense of pain stemming from mental health struggles, physical health issues, or personal insecurities • There’s a recurring theme of individuals grappling with their emotions and seeking understanding, validation, or support	• Moderate distress and struggles with mental health • Discusses personal experiences with mental health struggles, seeking help, and dealing with anxiety and social situations

Figure 6. SVD factorization of the embedding space for mental health subreddits. Centroids (representing each subreddit) are sorted from lowest to highest value on the second SVD factor in the top panel and on the third SVD factor in the bottom panel. On both panels, that is, for the second and third SVD components, the 4 extreme values correspond to r/SuicideWatch, r/SocialAnxiety, r/Depression, and r/BPD. BPD: borderline personality disorder; EDAnonymous: eating disorders anonymous; PTSD: posttraumatic stress disorder; SVD: singular value decomposition.



To complement our previous analyses that relied on extreme postings, here, we instead leveraged prototypical or representative postings within the r/SuicideWatch subreddit. To this end, we used ProtoDash to select the top 3 most representative postings, which were then fed into GPT-4 with the prompt of “Are these postings in line with current theories about suicide and suicidal ideation?” The following summarizes GPT-4’s response:

> *The first post describes feelings of hopelessness, despair, and anxiety, with a clear intent to commit suicide. The second post, despite stating a lack of suicidal intent, expresses chronic emotional distress, a lack of enjoyment in life, and recurring thoughts of wanting to die, which may actually be signs of suicidal ideation. The third represents a state of severe emotional distress, feelings of hopelessness, despair, loneliness, and a sense of being misunderstood and neglected; it also mentions previous suicide attempts, a plan for a final attempt, and mentions hearing voices, which could suggest a psychotic disorder.*

These posts are consistent with current theories of suicidality. The ITS, for example, posits that individuals are more likely to die by suicide when they have both the desire to die, often stemming from feelings of burden and social isolation, and the capability to do so, often developed through previous exposure to painful or fear-inducing experiences.

Discussion

Principal Findings

We combined analytic and interpretability techniques to study linguistic contents in large numbers of postings derived from the r/SuicideWatch subreddit in relation to other mental health subreddits, as well as select non–mental health subreddits. This allows us to better understand, both qualitatively and quantitatively, how suicidal ideation linguistically presents in popular social media sites (Reddit). Our results offer new insights into the emotional and thematic content shared by individuals at risk of suicide on Reddit.

First, by applying GPT-4 to postings spanning the top 3 axes in our dimensionally reduced embeddings across all subreddit postings, we were able to determine the linguistic meanings of these axes, which describe generally what people tend to discuss on anonymous web-based mental health forums, including suicide-related discourse. These axes were (1) resilience versus despair, (2) validation versus advice, and (3) high versus low distress. R/SuicideWatch was characterized by narratives of struggles and despair (eg, hopelessness and previous treatment being ineffective) and showed the highest values of nearly all mental health subreddits for seeking advice (more so than validation) and high levels of distress. R/SuicideWatch posts also had the highest values for seeking guidance and advice—just above r/BPD and r/Depression and just below r/SocialAnxiety. These specific communities and those with associated disorders or phenomena may seek advice or guidance (ie, practical solutions) more than others because of their chronicity, recurrence, and historical difficulty to treat. For example, spanning across 50 years of intervention research for suicidal thoughts and behaviors, recent meta-analytic evidence shows that overall intervention effects are small regardless of the intervention or suicide-related outcome studied [3]. Further, although effective treatments are available, when left untreated, social anxiety and BPD are often chronic conditions [19,20], and depressive disorders are highly recurrent [21]. Individuals on Reddit may therefore be searching for additional solutions, potentially because past treatment engagement was not adequately effective in reducing their symptoms. Last, regarding distress, our study is in line with previous studies examining suicide-related social media posts that characterizes these postings as indicating high levels of distress [22].

More generally, these 3 axes exemplify broad common themes during clinical appointments, with people sharing messages of *hope and resilience* or *despair and hopelessness* (referred to as "Emotional Well-Being" by GPT-4). Similarly, individuals on mental health subreddits mention a desire for *validation or support* and *problem-solving or solutions*, both of which are core to several effective psychosocial treatments. Last, subjective distress represents a core component of what helps define a mental health disorder [23] and varies widely from disorder to disorder as well as individually, as indicated by the values along this axis. Together, these axes may indicate that a therapy-like process naturally occurs in Reddit communities, where posters provide messages of despair or resilience with different degrees of distress and are searching for validation and solutions to their experiences.

Using ProtoDash in conjunction with GPT-4, we extracted and summarized common thematic and emotional contents from r/SuicideWatch postings. Results support the central variables within contemporary theories of suicide for explaining why the desire for suicide develops. Here, the 3 most prototypical posters predominantly wrote about feelings of disconnection, burden, hopelessness, desperation, resignation, and trauma. The ITS, 3ST, and IMV theories of suicide all state that disconnection (eg, thwarted belongingness), perceived burdensomeness, and feeling that their issues are intractable (eg, entrapment, hopelessness, and resignation) are necessary elements for developing a desire to die by suicide. Further, desperation—having a deep sense of despair, feeling overwhelmed, and lacking the ability to improve current conditions—is consistent with psychache [24] and the recently proposed diagnostic criteria for acute suicide conditions (eg, suicide crisis syndrome [25]), each of which cite despair as a core criterion for the development of suicidal ideation. Last, experiencing trauma (broadly defined) was frequently discussed in postings. Experiencing traumatic events is not posited as a necessary and sufficient condition for the development of suicidal ideation, but it has been put forth as a contributor for why people become capable of dying by suicide [7-9]. However, previous meta-analytic work has correlated traumatic experiences such as abuse with future suicidal ideation [2], and if the term is taken colloquially (ie, negative life stressors), it may contribute to theoretical constructs such as defeat or humiliation [9,25].

Next, we identified linguistically defined natural groupings among mental health subreddits. Our results seemed to suggest three different clusters: (1) r/SuicideWatch, r/Depression, r/MentalHealth, r/BPD, and r/Social Anxiety; (2) r/Psychosis, r/schizophrenia, and r/Bipolar; and (3) r/alcoholism, r/EDAnonymous, and r/addiction (see Figures 4 and 5). However, the 3 clusters mentioned above appear to (largely) support the superspectra put forth by the Hierarchical Taxonomy of Psychopathology (HiTOP)—a recent statistically driven diagnostic classification system for mental disorders [26]—compared to traditional clusters found in the *Diagnostic Statistical Manual of Mental Disorders, Fifth Edition* [23]. The three clusters mentioned above generally correspond to the following HiTOP spectra: (1) Internalizing Disorders; (2) Psychotic Disorders; and (3) Disinhibited Externalization Disorders. However, other mental health subreddits such as r/HealthAnxiety, r/Anxiety, r/mentalillness, and r/PTSD did not map as neatly onto any 1 dimension. In contrast with HiTOP, r/BPD was more aligned with internalizing disorders than externalizing disorders in this study. Similarly, while eating disorders (r/EDAnonymous) and substance or alcohol use disorders (r/addiction) are classified under different HiTOP spectra (Internalizing Disorders and Disinhibited Externalizing Disorders, respectively), their strong phenotypic associations and comorbidities are well documented in the literature (see a recent study examining their shared genetic risks [27]). Taken as a whole, our results also support a more nuanced and dimensional view of HiTOP spectra (eg, internalizing vs externalizing disorders), where posters in r/BPD are more often discussing unsatisfying relationships, feelings of emptiness, desires for self-harm, anger, and other cognitive criteria rather than solely discussing externalizing behaviors (eg, physical fights and risky behaviors) that are central to the BPD HiTOP spectra (ie, externalizing disorders).

As indicated by the hierarchical structure of the dendrogram in Figure 4, the linguistic features in r/SuicideWatch have substantial overlap with postings in r/Depression, r/BPD, and r/SocialAnxiety. This is likely due to suicidal ideation being a symptom of depression and suicidal (and parasuicidal) behaviors, which are common features of BPD [23]. Regarding the (somewhat unexpected) linguistic similarities between postings from r/SuicideWatch and r/SocialAnxiety, we posit that they are likely driven by mentioning social and interpersonal issues (eg, loneliness), which are common among both. These similarity findings may help serve as additional validity for our results and may have implications for recently proposed diagnostic criteria for suicide-related thoughts and behaviors, such as acute suicide affective disorder [28] and suicide crisis syndrome [25], that highlight abrupt or accentuated feelings of social disconnection or social withdrawal as an indicator of suicide crisis risk.

From Figure 6, we note in the top panel that r/BPD had the third highest value on the axis for “Understanding and Support,” with r/SuicideWatch being the second highest. In the bottom panel, r/BPD has the fourth highest value on the axis for “Severity of Distress,” next to r/Depression, whereas r/SuicideWatch had the highest value. Individuals with BPD are at a notably higher risk of suicide, with commonalities between the 2 including impulsivity, intense emotional dysregulation, and chronic feelings of emptiness. These individuals often struggle with unstable interpersonal relationships and heightening feelings of loneliness and rejection, which can trigger suicidal thoughts and behaviors. Additionally, a significant proportion of those with BPD have a history of trauma and may have co-occurring mental health disorders, such as depression or anxiety, further exacerbating the risk. The prevalence of self-harm behaviors in individuals with BPD, although not always indicative of suicidal intent, is also a critical risk factor [29-32].

Last, the representative posts selected by ProtoDash and interpreted by ChatGPT showed that r/SuicideWatch discussions emphasized feelings of empathy, support, understanding, and gratitude. In addition, ChatGPT found that typical replies offered messages of hope that things will improve and encouraged seeking help, traveling to beautiful places or to find solace, and connecting with nature to find “self love.” Many of these actions (eg, empathy, support, understanding, and encouraging help-seeking behavior) are what professional organizations (eg, National Suicide Prevention Lifeline and American Foundation for Suicide Prevention) advocate friends, family, and communities provide to individuals in crisis. Although some of the proposed suggestions do not have an evidence base for being effective in reducing suicidal desire (eg, traveling to beautiful places), survivors of suicide or individuals with lived experience (eg, other members of r/SuicideWatch) may provide additional perspectives that can be helpful when used alongside evidence-based therapies and interventions [33]. Overall, these findings indicate that some web-based communities, such as r/SuicideWatch, could be a source of support for many individuals experiencing suicidal thoughts and can act in accordance with the suggestions put forth by several professional suicide prevention organizations.

Limitations and Future Directions

We note a few limitations of our study. First, the data were limited to a 3-month period, which may not be sufficient to fully capture the range of experiences and emotions expressed in the r/SuicideWatch subreddit. Additionally, instead of more broadly looking into other social media platforms, the study focused solely on Reddit, and thus, the findings may not generalize to other web-based platforms. Future research could expand the time frame of data collection; explore other web-based platforms; and integrate additional data sources, such as user comments, to provide a more comprehensive understanding of web-based expressions of suicide risk. Last, the results from this study could not be validated against external criteria such as established measures of suicide risk or clinician judgment, potentially limiting the credibility of our findings. Future studies could incorporate multiple perspectives to help understand the accuracy and reliability of the extracted thematic interpretations.

Conclusion

In conclusion, we used a novel combination of NLP techniques to detect and interpret linguistic patterns of mental health subreddits to better understand how suicidal ideation presents in web-based communities. LLM embeddings allowed for a nuanced analysis of subreddit content that revealed unique patterns and shared themes that are specific to suicide-related content. Further, dimensional reduction revealed latent dimensions of mental health discussions and helped identify relationships between various subreddits. Last, we used generative LLM for XAI to gain deeper insights into the emotions and experiences of individuals posting about suicidal thoughts. Our results supported contemporary theories of suicide. Our study highlights the potential use of web-based linguistic patterns as valuable data sources to better understand mental health disorders and suicidality.

Acknowledgments

We thank the Reddit community for their openness in sharing their experiences and emotions, which made this research possible. We also acknowledge the contributions of ChatGPT (OpenAI) [14] for its assistance in data interpretation.

Data Availability

The data sets used and analyzed during this study were downloaded from The-Eye.eu [16].

Authors' Contributions

BB, RN, AL, and GC created the study concept. BB, AL, RN and GC were major contributors to writing the manuscript. ZAR provided the code to access Reddit data from The-Eye.eu [16]. BW provided the code to access ChatGPT programmatically. RN collected and analyzed the data. All authors read and approved the final manuscript.

Conflicts of Interest

AL is a cofounder of Keywise AI and has served as an adviser or consultant for Otsuka US and Buoy Health. BB, RN, ZAR, BW, and GC have no conflicts of interest.

Multimedia Appendix 1
Proportion of posts from a given subreddit “closest” to the r/SuicideWatch centroid.
[PDF File, 13 KB - mental_v11i1e57234_app1.pdf]

References

1. Garnett MF, Curtin SC. Suicide mortality in the United States, 2001-2021. NCHS Data Brief 2023 Apr;464:1-8. [Medline: 37093258]
2. Franklin JC, Ribeiro JD, Fox KR, et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol Bull 2017 Feb;143(2):187-232. [doi: 10.1037/bul0000084] [Medline: 27841450]
3. Fox KR, Huang X, Guzmán EM, et al. Interventions for suicide and self-injury: a meta-analysis of randomized controlled trials across nearly 50 years of research. Psychol Bull 2020 Dec;146(12):1117-1145. [doi: 10.1037/bul0000305] [Medline: 33119344]
4. Low DM, Rumker L, Talkar T, Torous J, Cecchi G, Ghosh SS. Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on Reddit during COVID-19: observational study. J Med Internet Res 2020 Oct 12;22(10):e22635. [doi: 10.2196/22635] [Medline: 32936777]
5. Fineberg SK, Leavitt J, Deutsch-Link S, et al. Self-reference in psychosis and depression: a language marker of illness. Psychol Med 2016 Sep;46(12):2605-2615. [doi: 10.1017/S0033291716001215] [Medline: 27353541]
6. Birnbaum ML, Norel R, van Meter A, et al. Identifying signals associated with psychiatric illness utilizing language and images posted to Facebook. NPJ Schizophr 2020 Dec 3;6(1):38. [doi: 10.1038/s41537-020-00125-0] [Medline: 33273468]
7. van Orden KA, Witte TK, Gordon KH, Bender TW, Joiner TEJ. Suicidal desire and the capability for suicide: tests of the interpersonal-psychological theory of suicidal behavior among adults. J Consult Clin Psychol 2008 Feb;76(1):72-83. [doi: 10.1037/0022-006X.76.1.72] [Medline: 18229985]
8. Klonsky ED, May AM. The Three-Step Theory (3ST): a new theory of suicide rooted in the “ideation-to-action” framework. Int J Cogn Ther 2015 Jun;8(2):114-129. [doi: 10.1521/ijct.2015.8.2.114]
9. O’Connor RC. The integrated motivational-volitional model of suicidal behavior. Crisis 2011;32(6):295-298. [doi: 10.1027/0227-5910/a000120] [Medline: 21945841]
10. Le Glaz A, Haralambous Y, Kim-Dufor DH, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res 2021 May 4;23(5):e15708. [doi: 10.2196/15708] [Medline: 33944788]
11. Arowosegbe A, Oyelade T. Application of natural language processing (NLP) in detecting and preventing suicide Ideation: a systematic review. Int J Environ Res Public Health 2023 Jan 13;20(2):1514. [doi: 10.3390/ijerph20021514] [Medline: 36674270]
12. Arnaud E, Elbattah M, Moreno-Sánchez PA, Dequen G, Ghazali DA. Explainable NLP model for predicting patient admissions at emergency department using triage notes. In: 2023 IEEE International Conference on Big Data (BigData): IEEE; 2023:4843-4847. [doi: 10.1109/BigData59044.2023.10386753]
13. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on May 24, 2019. [doi: 10.48550/arXiv.1810.04805]
14. ChatGPT. OpenAI. URL: https://chat.openai.com/ [accessed 2024-05-01]
15. Gurumoorthy KS, Dhurandhar A, Cecchi G, Aggarwal C. Efficient data representation by selecting prototypes with importance weights. In: 2019 IEEE International Conference on Data Mining (ICDM): IEEE; 2019:260-269. [doi: 10.1109/ICDM.2019.00036]
16. Reddit/Pushshift archives. The-Eye.eu. URL: https://the-eye.eu/redarcs/ [accessed 2024-05-01]
17. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. arXiv. Preprint posted online on Aug 27, 2019. [doi: 10.48550/arXiv.1908.10084]

18. Srivastava A, Selloni A, Bilgrami ZR, et al. Differential expression of anomalous self-experiences in spontaneous speech in clinical high-risk and early-course psychosis quantified by natural language processing. Biol Psychiatry Cogn Neurosci Neuroimaging 2023 Oct;8(10):1005-1012. [doi: 10.1016/j.bpsc.2023.06.007] [Medline: 37414359]
19. Crowell SE, Beauchaine TP, Linehan MM. A biosocial developmental model of borderline personality: elaborating and extending Linehan's theory. Psychol Bull 2009 May;135(3):495-510. [doi: 10.1037/a0015616] [Medline: 19379027]
20. Keller MB. Social anxiety disorder clinical course and outcome: review of Harvard/Brown Anxiety Research Project (HARP) findings. J Clin Psychiatry 2006;67(Suppl 12):14-19. [Medline: 17092191]
21. Burcusa SL, Iacono WG. Risk for recurrence in depression. Clin Psychol Rev 2007 Dec;27(8):959-985. [doi: 10.1016/j.cpr.2007.02.005] [Medline: 17448579]
22. O'Dea B, Wan S, Batterham PJ, Calear AL, Paris C, Christensen H. Detecting suicidality on Twitter. Internet Interv 2015 May;2(2):183-188. [doi: 10.1016/j.invent.2015.03.005]
23. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition: American Psychiatric Association; 2013. [doi: 10.1176/appi.books.9780890425596]
24. Shneidman ES. Commentary: suicide as psychache. J Nerv Ment Dis 1993 Mar;181(3):145-147. [doi: 10.1097/00005053-199303000-00001]
25. Schuck A, Calati R, Barzilay S, Bloch-Elkouby S, Galynker I. Suicide crisis syndrome: a review of supporting evidence for a new suicide-specific diagnosis. Behav Sci Law 2019 May;37(3):223-239. [doi: 10.1002/bsl.2397] [Medline: 30900347]
26. Kotov R, Krueger RF, Watson D, et al. The Hierarchical Taxonomy of Psychopathology (HiTOP): a quantitative nosology based on consensus of evidence. Annu Rev Clin Psychol 2021 May 7;17:83-108. [doi: 10.1146/annurev-clinpsy-081219-093304] [Medline: 33577350]
27. Munn-Chernoff MA, Johnson EC, Chou YL, et al. Shared genetic risk between eating disorder- and substance-use-related phenotypes: evidence from genome-wide association studies. Addict Biol 2021 Jan;26(1):e12880. [doi: 10.1111/adb.12880] [Medline: 32064741]
28. Tucker RP, Michaels MS, Rogers ML, Wingate LR, Joiner TEJ. Construct validity of a proposed new diagnostic entity: acute suicidal affective disturbance (ASAD). J Affect Disord 2016 Jan 1;189:365-378. [doi: 10.1016/j.jad.2015.07.049] [Medline: 26476421]
29. Paris J. Suicidality in borderline personality disorder. Medicina 2019 May 28;55(6):223. [doi: 10.3390/medicina55060223]
30. Xu YE, Barron DA, Sudol K, Zisook S, Oquendo MA. Suicidal behavior across a broad range of psychiatric disorders. Mol Psychiatry 2023 Jul;28(7):2764-2810. [doi: 10.1038/s41380-022-01935-7] [Medline: 36653675]
31. Scheunemann J, Kühn S, Biedermann SV, et al. Implicit cognitions on self-injurious and suicidal behavior in borderline personality disorder. J Behav Ther Exp Psychiatry 2023 Jun;79:101836. [doi: 10.1016/j.jbtep.2023.101836] [Medline: 36709601]
32. Toma S, Sinyor M, Mitchell RHB, Schaffer A. Transdiagnostic suicidality in depression: more similar than different. Acta Psychiatr Scand 2023 Sep;148(3):219-221. [doi: 10.1111/acps.13600] [Medline: 37605854]
33. Watling D, Preece M, Hawgood J, Bloomfield S, Kõlves K. Developing an intervention for suicide prevention: a rapid review of lived experience involvement. Arch Suicide Res 2022;26(2):465-480. [doi: 10.1080/13811118.2020.1833799] [Medline: 33073734]

Abbreviations

3ST: Three-Step Theory
BERT: Bidirectional Encoder Representations from Transformers
BPD: borderline personality disorder
EDAnonymous: eating disorders anonymous
HiTOP: Hierarchical Taxonomy of Psychopathology
IMV: Integrated Motivational–Volitional
ITS: Interpersonal Theory of Suicide
LLM: large language model
ML: machine learning
NLP: natural language processing
PTSD: posttraumatic stress disorder
SVD: singular value decomposition
XAI: explainable artificial intelligence

Edited by J Torous; submitted 08.02.24; peer-reviewed by A Hassan, M Elbattah, T Adebambo; revised version received 28.03.24; accepted 29.03.24; published 16.05.24.

Please cite as:
Bauer B, Norel R, Leow A, Rached ZA, Wen B, Cecchi G
Using Large Language Models to Understand Suicidality in a Social Media–Based Taxonomy of Mental Health Disorders: Linguistic Analysis of Reddit Posts
JMIR Ment Health 2024;11:e57234
URL: https://mental.jmir.org/2024/1/e57234
doi:10.2196/57234

Emerging Trends of Self-Harm Using Sodium Nitrite in an Online Suicide Community: Observational Study Using Natural Language Processing Analysis

Sudeshna Das[1], PhD; Drew Walker[2], MS; Swati Rajwal[3], MTech; Sahithi Lakamana[1], MS; Steven A Sumner[4], MD; Karin A Mack[4], PhD; Wojciech Kaczkowski[4], PhD; Abeed Sarker[1,5], PhD

1
2
3
4
5

Corresponding Author:
Sudeshna Das, PhD

Abstract

Background: There is growing concern around the use of sodium nitrite (SN) as an emerging means of suicide, particularly among younger people. Given the limited information on the topic from traditional public health surveillance sources, we studied posts made to an online suicide discussion forum, "Sanctioned Suicide," which is a primary source of information on the use and procurement of SN.

Objective: This study aims to determine the trends in SN purchase and use, as obtained via data mining from subscriber posts on the forum. We also aim to determine the substances and topics commonly co-occurring with SN, as well as the geographical distribution of users and sources of SN.

Methods: We collected all publicly available from the site's inception in March 2018 to October 2022. Using data-driven methods, including natural language processing and machine learning, we analyzed the trends in SN mentions over time, including the locations of SN consumers and the sources from which SN is procured. We developed a transformer-based source and location classifier to determine the geographical distribution of the sources of SN.

Results: Posts pertaining to SN show a rise in popularity, and there were statistically significant correlations between real-life use of SN and suicidal intent when compared to data from the Centers for Disease Control and Prevention (CDC) Wide-Ranging Online Data for Epidemiologic Research (=0.727; P<.001) and the National Poison Data System (=0.866; P=.001). We observed frequent co-mentions of antiemetics, benzodiazepines, and acid regulators with SN. Our proposed machine learning–based source and location classifier can detect potential sources of SN with an accuracy of 72.92% and showed consumption in the United States and elsewhere.

Conclusions: Vital information about SN and other emerging mechanisms of suicide can be obtained from online forums.

(JMIR Ment Health 2024;11:e53730) doi:10.2196/53730

KEYWORDS

online suicide community; suicide; sodium nitrite; sodium nitrite sources; mental health; adolescent; juvenile; self harm; Sanctioned Suicide; online forum; US; public health; surveillance; data mining; natural language processing; machine learning; usage; suicidal; accuracy; consumption; information; United States

Introduction

Background

Suicide rates in the United States continue to rise and remain near their highest levels in more than 2 decades [1,2]. There were 48,183 suicides in the United States in 2021, which is 5% higher than the reported number in 2020 [3]. Poisoning is the most common mechanism of suicide attempts in the United States [4] and the third-leading mechanism involved in suicides [5]. One factor complicating suicide prevention efforts is the continual emergence and promotion of new means by which one can attempt suicide, such as novel substances.

Since 2019, a growing trend of using sodium nitrite (SN), a common food additive, for self-harm has been reported [6,7]. SN has traditionally been used as a food preservative and coloring agent, in addition to use as a corrosion inhibitor. As such, it is widely available for purchase. An alarming development has been the sale of "suicide kits" in online marketplaces, which comprise SN in addition to instructional material on attempting suicide [8]. Instances of the use of such suicide kits have been reported in the literature [9]. Furthermore, media reports of celebrity suicides from SN ingestion have raised public awareness of this means of self-harm. As a widely available, water-soluble salt with reported lethal dosages of as low as 0.7 g [10], the potentially growing popularity of SN as a suicide mechanism is concerning. Between 2018 and 2020, the annual suicide rate involving SN increased from 0.01 to 0.09 per 100,000 person-years in the United States [11]. Although there are concerns about increased youth suicides due to media contagion [12], studies show that adhering to suicide reporting guidelines [13] can raise awareness and have a protective effect through the coverage of positive coping mechanisms [14].

The consumption of SN induces methemoglobinemia, a condition resulting in hypoxia, and if not treated promptly, it can result in death. Although the unintentional consumption of SN due to misleading or dubious storage practices is of concern [15], the consumption of SN with the intent of self-harm has also been reported in countries such as Australia, Portugal, and South Korea [6]. It is believed that global popularization of SN, instruction on its use in suicide, and sharing of information about procuring SN has been facilitated by online forums such as “Sanctioned Suicide” [15], about which little is known.

Online Suicide Forums

Online forums provide a platform for users with similar interests to share their views on common topics of interest. Internet support forums exist for a wide variety of health-related issues, including mental health and suicide-related behaviors. Sanctioned Suicide is the successor of the eponymous subreddit (a topic-specific forum), which was banned in March 2018 for violating Reddit’s policies on content promoting self-harm and specific suicide methods [15]. The purpose of the website, as mentioned in their frequently asked questions [16], is to allow individuals to discuss suicide—including suicide methods—without the content screening that occurs on more prominent social media platforms. Thus, this forum encapsulates a large amount of suicide-related information that can be of high utility for planning and enacting public health measures to prevent suicides. The large volume of data, however, also makes it impractical to manually review the content continuously to generate timely and evolving insights. Automated methods are thus required to optimally leverage this resource of publicly available information.

Objective

Although the use of SN for suicide has elevated to the level of congressional interest in the United States [17], little is known about epidemiologic trends from Sanctioned Suicide that could inform prevention efforts. The key strategies highlighted in the *Suicide Prevention Resource for Action* [18] by the US Centers for Disease Control and Prevention (CDC) call for “data-driven strategic planning with engagement from multi-sectoral partners” for making decisions associated with the prevention of suicides in the United States [17]. In line with the objectives and guidance set out in this resource, in this paper, we adopt a data-driven approach using natural language processing (NLP) and other techniques to study large-scale public posts from Sanctioned Suicide to answer critical questions relevant to suicide prevention efforts:

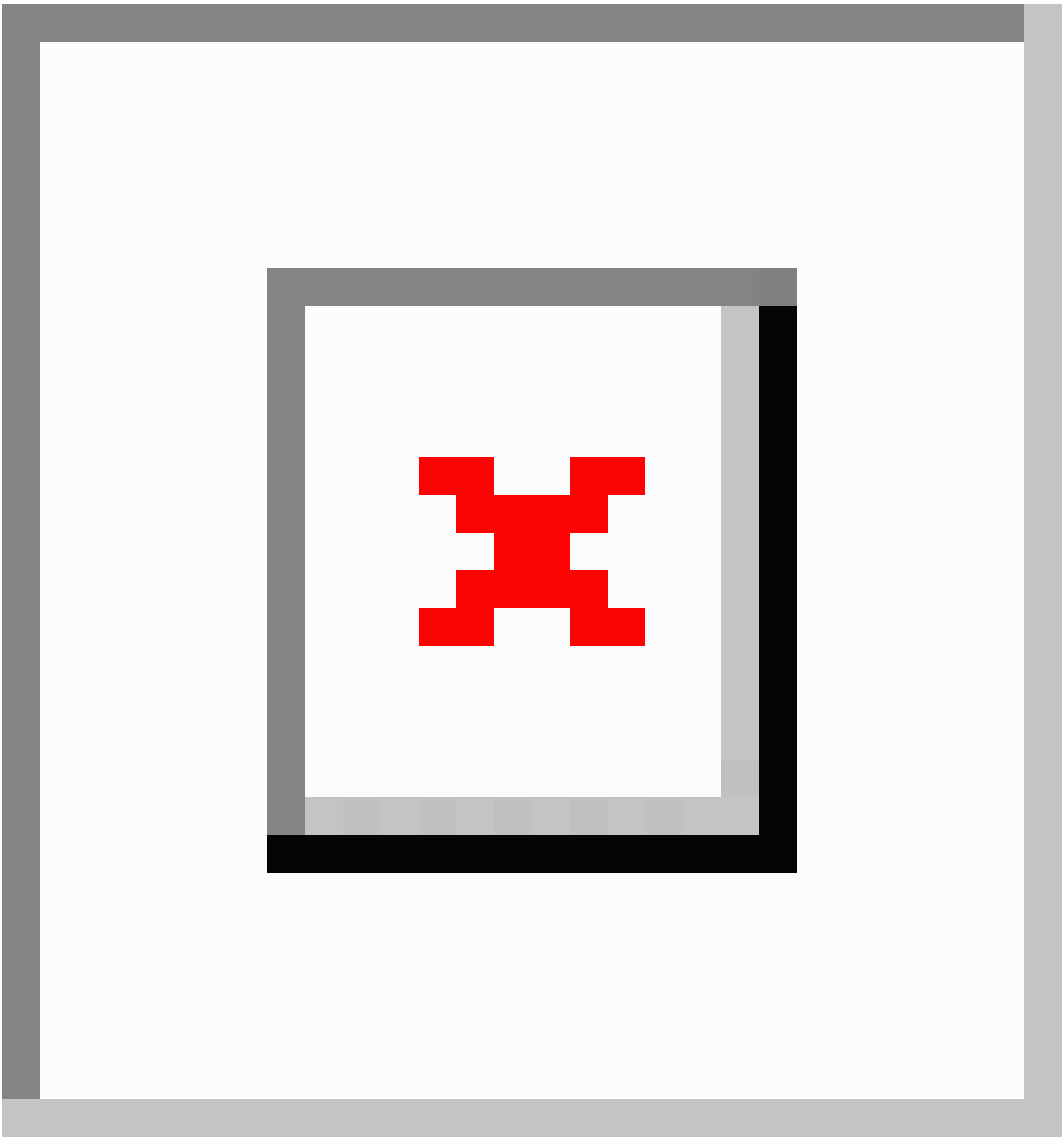
1. Does interest in SN appear to be increasing over time on the forum, and how does this interest compare to other mechanisms of suicide?
2. Are there other co-occurring substances or topics of interest relevant to SN?
3. What are the leading countries and vendors of SN that are being promoted?

Methods

Data Collection and Preprocessing

We collected data from the website “Sanctioned Suicide,” an online community dedicated to discussing “the topic of suicide without the censorship of other places” [16], which has received substantial attention because of its rising popularity in the recent past [15]. Figure 1 presents the structure of the Sanctioned Suicide network. As the figure illustrates, the social network is broadly divided into 3 types of discussion based on topic (*suicide*, *recovery*, and *off-topic*). We collected all posts available on the website from March 22, 2018 (the date on which the website went live), to October 7, 2022 (the date of data collection). We removed duplicate posts and applied preprocessing steps that are standard in NLP, namely tokenization, lowercasing, punctuation removal, stop-word removal, and lemmatization.

Figure 1. Overall structure of the Sanctioned Suicide website. Threads are organized under 3 broad categories: suicide discussion, recovery, and off-topic.



Data-Driven Analysis of Suicide Mechanisms

Suicide Mechanism Detection and Trend Analysis

Our first objective was to detect mentions of SN as well as other specific substances and suicide methods. As is common over social media, many lexical variants are used to discuss SN and other substances or methods. Not including commonly used lexical variants leads to low-sensitivity data collection [19]. We took a data-driven approach to identify all the relevant lexical variants for SN. We trained a Word2Vec model using n-grams (n=1, 2, and 3) from the entire data set, which enabled us to automatically identify lexical expressions that are the most semantically similar to SN. We manually reviewed these terms and identified, in addition to alternative names for SN, keywords associated with other suicide mechanisms. In consultation with our subject matter experts, we manually grouped these keywords into 14 categories: "sodium nitrite," "cyanides," "firearms," "hanging," "acid regulators," "ricin," "plant-based poisons," "antiemetics," "other preservatives," "nitric oxide," "household chemicals," "barbiturates," "benzodiazepines," and "opioids." The complete list of suicide mechanism–related terms identified is given in Table S1 in Multimedia Appendix 1.

We performed automatic searches over the whole data set to compute the frequencies of posts mentioning each method of suicide over time. A post that mentioned any number of lexical variants associated with a suicide method counted toward that method. For example, a post mentioning only "$NaNO_2$" is assigned the group "sodium nitrite," whereas a post mentioning "gun" and "full suspension" is assigned the group labels "firearms" and "hanging." We computed the monthly and normalized frequencies of posts mentioning each of these categories to analyze their temporal trends.

Comparison of Trends With Traditional Data

We compared the temporal trends of SN mentions we discovered from the above analysis with relevant metrics reported in two more traditional sources: (1) intentional exposures to SN in the US National Poison Data System (NPDS) and (2) death counts in the CDC Wide-Ranging Online Data for Epidemiologic Research (WONDER) database. For the first comparison, we compared our data against the quarterly intentional exposures to SN from the NPDS reported by McCann et al [10]. For the second comparison, we compared against the reported deaths under the underlying cause of death codes U03, X60–X84, and Y87.0 and multiple causes of death code T50.6 from the CDC WONDER database. For the latter, to make the comparison better aligned with the traditional data source, we combined our keyword mention counts pertaining to "sodium nitrite" and "other preservatives." We performed a Spearman rank correlation test to assess possible associations between the pairs of statistics.

Co-Occurrence Analysis and Topic Modeling

We performed a co-occurrence analysis to compute the number of times different suicide methods we already identified were mentioned together. The intuition behind this analysis was that suicide methods that are considered together or substances that are taken together (eg, substances taken alongside SN) are likely to be mentioned more frequently in the same posts.

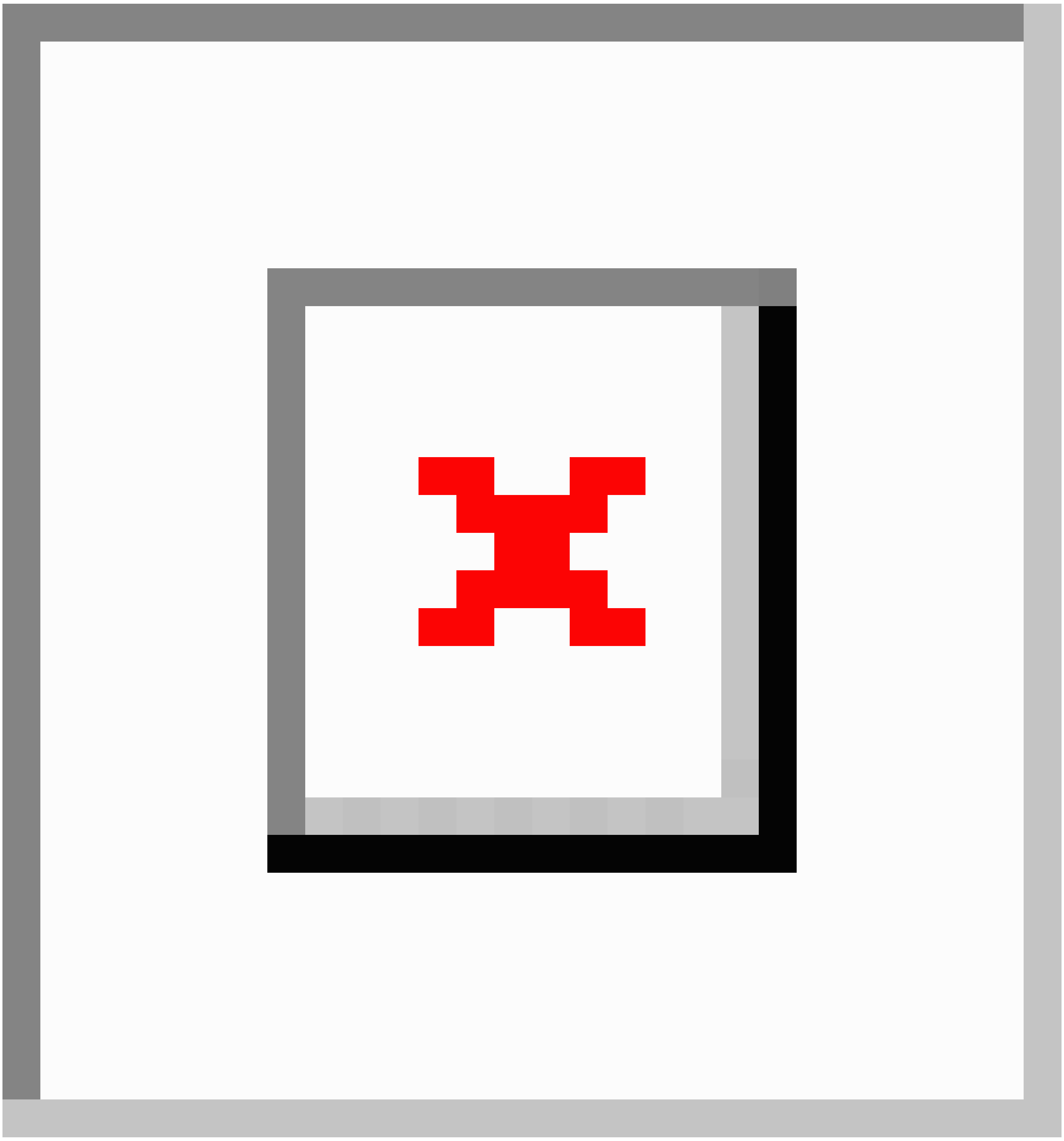
To obtain further insights about the topics associated with SN chatter, we conducted a topic modeling experiment. Topic modeling is a class of unsupervised algorithms that attempt to identify clusters of lexical elements that belong to latent topics from large sets of texts. Since the process is purely unsupervised, the topic clusters are not known a priori. We applied the BERTopic model, an unsupervised topic modeling approach that automatically clusters content from posts into a mathematically optimized number of topics [20].

Identification of SN Sources and Consumer Locations

Named Entity Recognition

Purchasing, sourcing, and procurement of SN were identified as common topics of discussion during the analyses mentioned above (see the *Results* section), so we used a 2-fold approach to identify information about both the retail venues and the geographic locations where SN was being sought for purchase. Specifically, we leveraged named entity recognition (NER) models trained on large web-based data sets to identify potential sources of SN: locations, organizations, and keywords used to look for SN on online marketplaces. NER is an information extraction technique used to discover named entities (such as organizations, locations, etc) in a textual corpus (Figure 2).

Figure 2. Examples of named entity recognition from Sanctioned Suicide posts from March 2018 to September 2022. Detected entities, their spans, and their inferred types are shown. GPE: geopolitical entity; ORG: organization; SN: sodium nitrite.



We used an NER algorithm available in the Python programming language (spaCy NER [21]) to detect possible locations or sources from which people seek or obtain SN. Apart from using spaCy's location entity label, we created a custom entity called "Suicide_Method" that would identify the substance in the text and highlight it. We used the rule-based pattern recognizer in spaCy to detect mentions of SN. In particular, we used the pattern "sn" to detect mentions of SN (the code snippet is shown in Multimedia Appendix 1). Posts where SN was mentioned in conjunction with location names were used for further analysis. The location entity recognizer of spaCy has a reported F_1-score of 0.916 [22].

Location Mention–Intent Classifier

The locations identified by the NER methods represented both sources of SN and consumer locations, and thus, the process required further disambiguation. We modeled this disambiguation as a supervised classification task and trained a transformer-based location classifier to classify the locations obtained from NER as "consumer location" and "purchase location." Transformer models leverage large amounts of "pre-trained" language data, which can then be fine-tuned on a specific task—such as location type identification, in the case of this study [23]. Sentence-level annotation of location mentions was carried out by 2 of the authors to create a gold-standard training data set for fine-tuning the model. The annotation process was carried out iteratively, with the annotation guidelines being refined after each round of annotation. Disagreements were resolved after a detailed discussion between 3 of the authors to reach a consensus after each round. Two rounds of annotation were performed, and interannotator agreement was computed based on Cohen κ, revealing good agreement for both source location (κ=0.80) and consumer location (κ=0.84) [24]. A total of 722 samples were annotated, of which 577 (80%) were used for training and the remaining 145 (20%) were used for testing.

From the many transformer-based models that are publicly available, we chose the RoBERTa model [25], which is based on Bidirectional Encoder Representations From Transformers [26] and has been shown to achieve state-of-the-art results on several language processing tasks similar to those by the base model, including for health-related, social media–based text classification tasks [27]. Our use of "state-of-the-art" refers to the best-performing machine or deep learning models currently available for each task, in the rest of the paper. We used the embeddings obtained from RoBERTa to fine-tune our model using the training data. We computed the distributions of the sources and locations identified automatically for analysis. Classification results and all outputs are provided in the *Results* section.

Ethical Considerations

This study was deemed to be exempt from review (publicly available data) by the Emory University Institutional Review Board. Our analyses use publicly available, user-generated content from an online forum where users remain anonymous by default. We do not use any personally identifiable information and only report aggregated data.

Results

Data Collection and Frequency Analysis

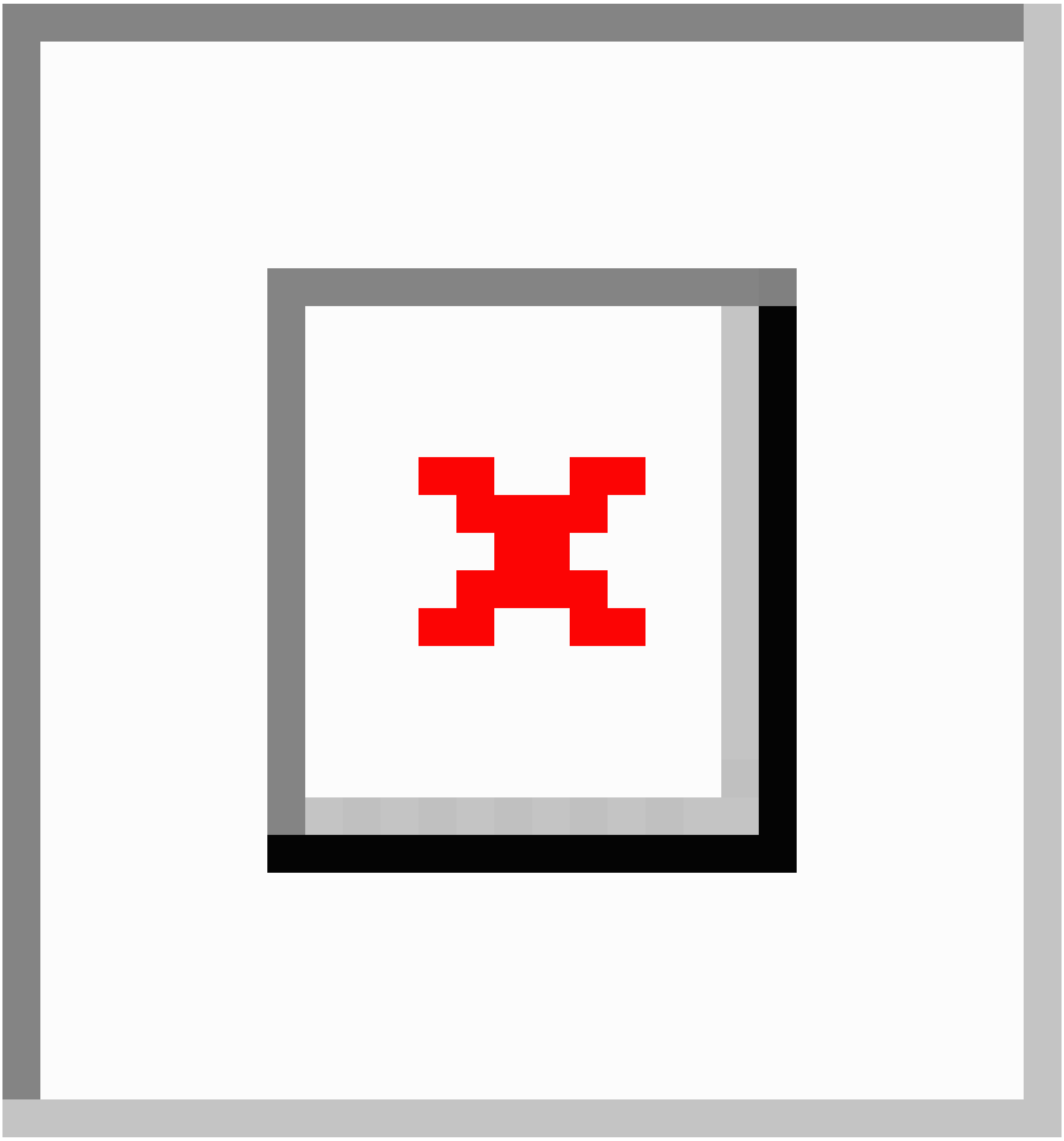
Overview

A total of 1,337,982 posts were collected. Of these, 1,302,620 were posted under the "Discussion" threads; 28,666 under the "Poll" thread; and 6696 under the "Question" threads. Preprocessing and removal of duplicate posts resulted in 1,329,042 total posts. There was a steady increase in the total number of posts from 2018 to 2020, followed by a slight decrease in 2021. The highest number of posts on the website was made in 2020. In the months leading up to September 2022, the final full month of data collection, the total number of monthly posts was generally higher than in the corresponding months in 2021 (Figure S1 in Multimedia Appendix 1).

Temporal SN Trends

Figure 3 shows the relative monthly frequencies for posts mentioning SN, as well as other substances and methods associated with suicides. From the figure, it can be observed that SN was the most popular means of suicide discussed in this online community, and the frequency of mentions of SN increased over time. A sharp rise in SN mentions can be seen towards the end of 2019, with the frequency of mentions remaining elevated thereafter.

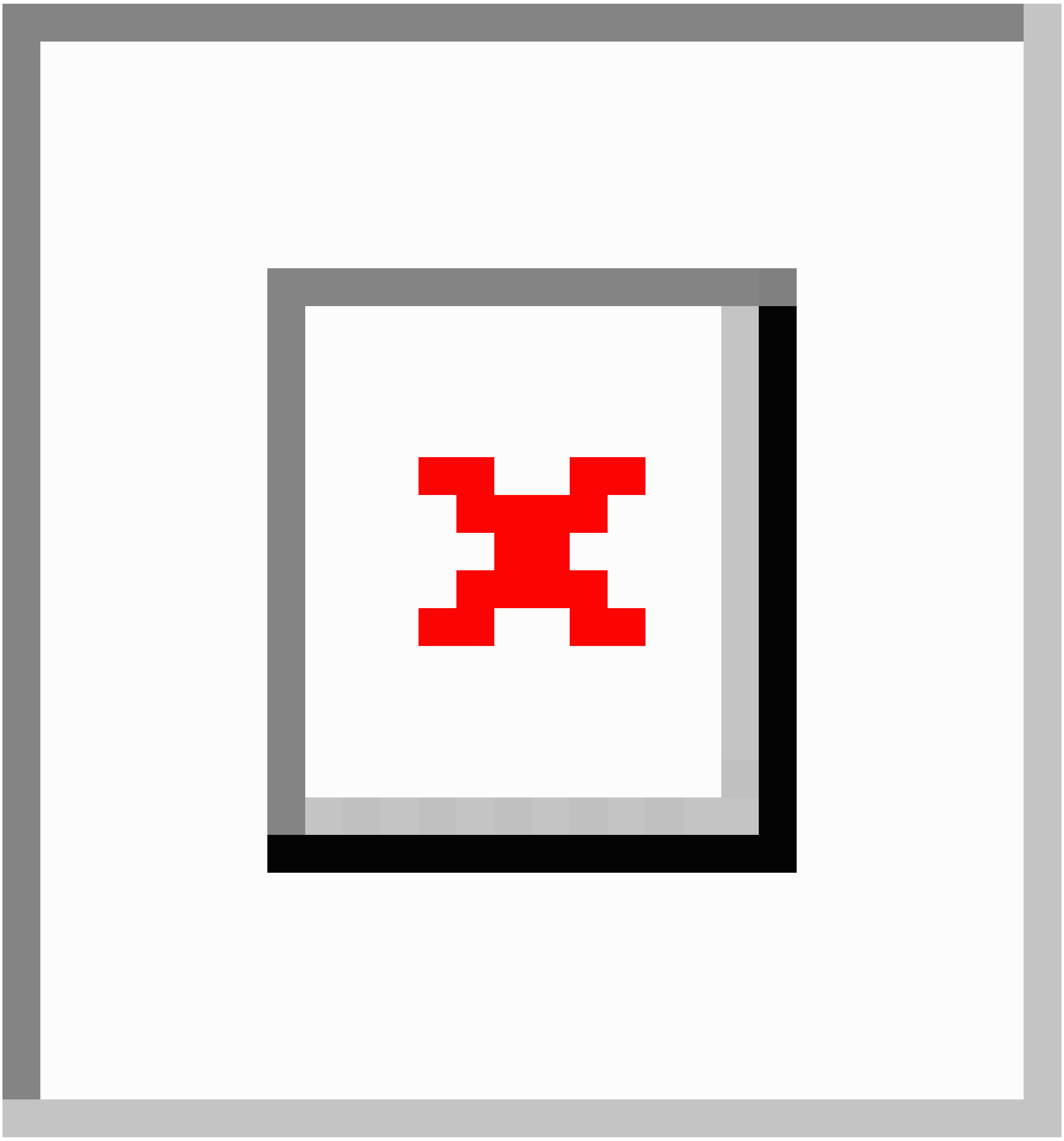
Figure 3. Normalized frequencies of posts on Sanctioned Suicide mentioning potential suicide means and related substances per month from March 2018 to September 2022.



Comparison of Trends With Traditional Data

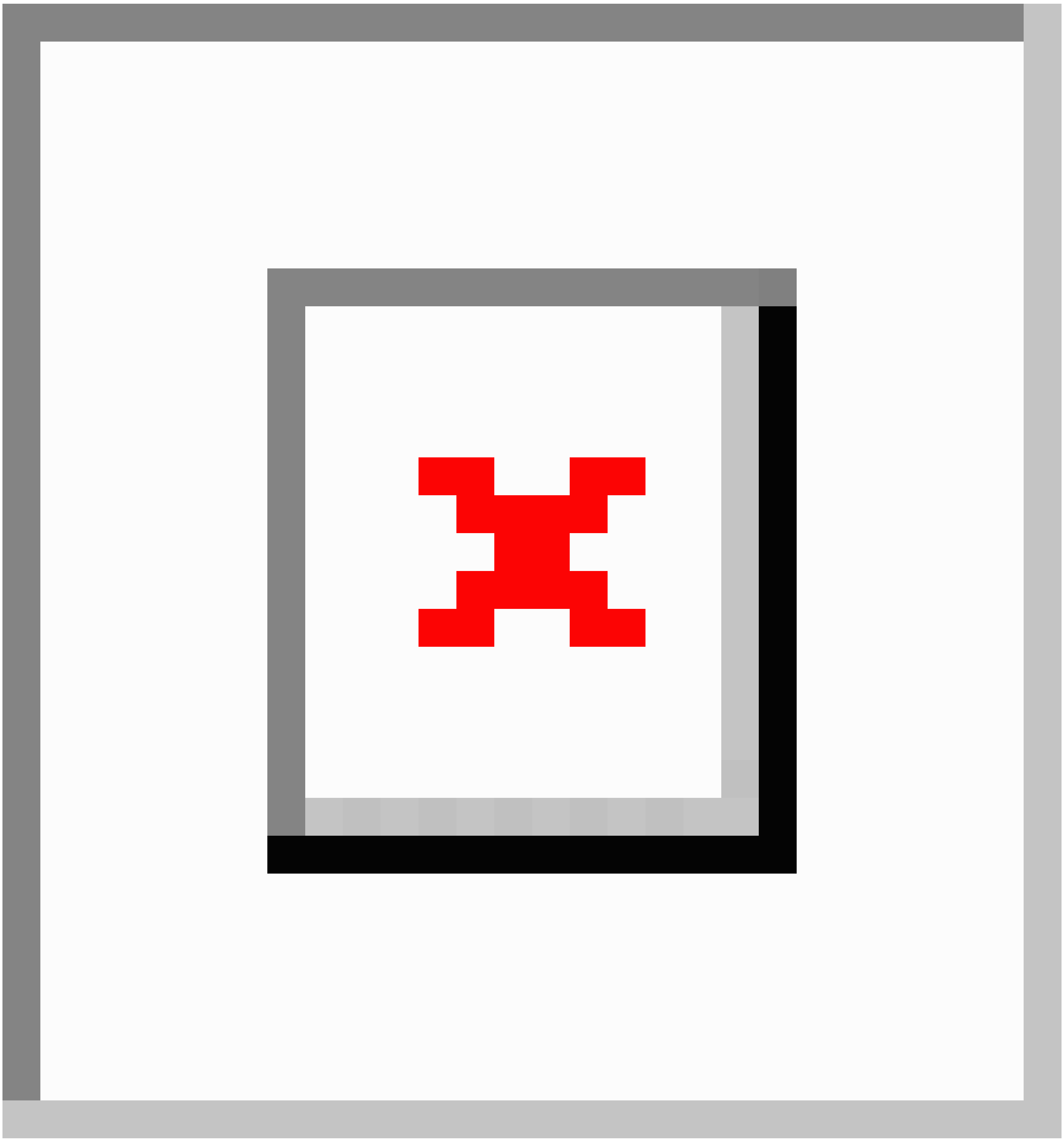
SN-related deaths due to intentional consumption have increased since 2019 [28]. Our analyses align with this uptick of SN ingestions and suicides: a sharp increase in the mentions of purchases of SN was seen toward the end of 2019 (Figure 3). In our comparison against the NPDS data, we obtained a Spearman of 0.866 (*P*=.001), revealing a statistically significant association between the 2 data sources. On visualizing the normalized frequencies of intentional exposures in the NPDS and purchases made as obtained from our data set (see the *Sources and Consumer Locations* section), we found that both the noncumulative and cumulative frequencies showed similar trends (Figure 4) for the 10 quarters spanning from 2018 to 2020 for which data from both sources were available (quarters 1 to 4 for 2018 and 2019, and quarters 1 and 2 for 2020). Although the mentions of purchases on Sanctioned Suicide are not exclusive to the United States, the similar trends are a strong indicator of online suicide community content reflecting real-life suicide incidences.

Figure 4. Comparison of normalized frequencies of purchase mentions on Sanctioned Suicide (SS) versus National Poison Data System (NPDS) exposures as reported in McCann et al [10], from 2018 to 2020. (A) Noncumulative frequencies; (B) cumulative frequencies. Q: quarter.



In the comparison against metrics from the CDC WONDER database, we obtained a Spearman of 0.727 (*P*<.001) for month-by-month frequencies during the time period from March 2018 to December 2021 for SN purchases versus actual deaths and a Spearman of 0.775 (*P*<.001) for keyword mentions versus actual deaths, revealing statistically significant correlation in both cases (Figure 5).

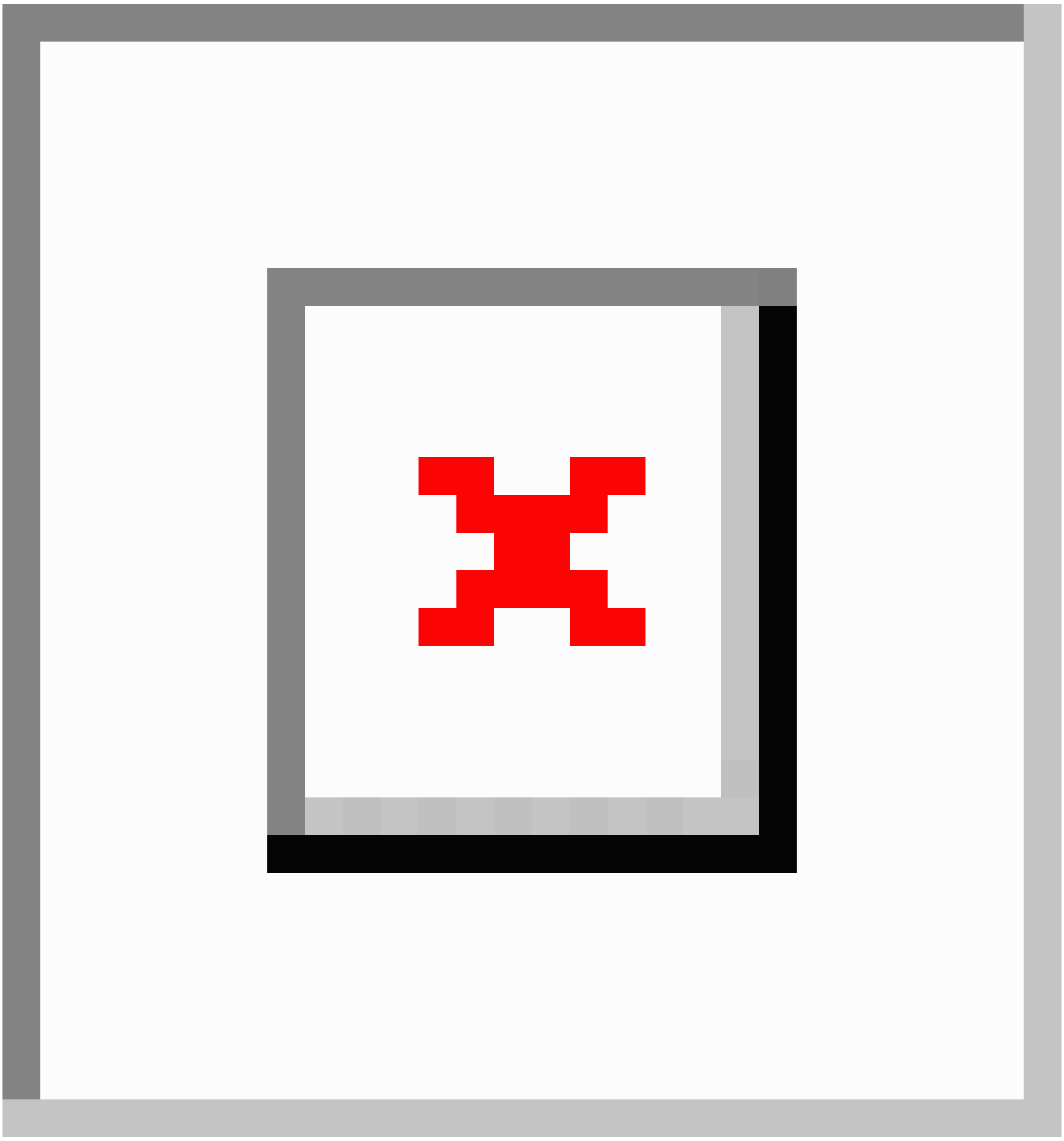
Figure 5. Comparison of normalized frequencies. (A) Purchase mentions of “sodium nitrite“ on Sanctioned Suicide versus actual deaths reported in the CDC WONDER database from 2018 to 2021; (B) keyword mentions of “sodium nitrite” and “other preservatives” on Sanctioned Suicide versus actual deaths reported in the CDC WONDER database. Suicide deaths involving chelating agents were identified by using the *International Classification of Diseases, Tenth Revision* underlying cause of death codes U03, X60–X84, and Y87.0 and multiple causes of death code T50.6. CDC: Centers for Disease Control and Prevention; WONDER: Wide-Ranging Online Data for Epidemiologic Research.



Co-mentioned Suicide Mechanisms

The heatmap in Figure 6 presents the co-occurrence frequencies between different substances and methods. As illustrated in the figure, co-occurring mentions of SN and antiemetics are common. Manual qualitative inspection of these posts revealed that antiemetics are often mentioned when discussing SN, as forum users recommended these substances to ensure that individuals do not feel nauseous after consuming SN.

Figure 6. Heatmap illustrating the most commonly co-occurring suicide mechanism mentions in Sanctioned Suicide posts from March 2018 to September 2022.



Topic Modeling

Topic modeling revealed further insights and some key differences in the content of posts. First, the topics discovered reinforced some of the insights revealed in Figure 3. For the SN group, unigram and bigram topic clusters represented the following:

- Substances that are potentially coingested (benzodiazepines, antiemetics, and metoclopramide);
- Dosage amounts (“tablespoon” and “grams”);
- Sourcing-related questions and information (“I’m looking,” “source,” “ordered,” and “package”);
- Comparison with other mechanisms of suicide (“hanging” and “shotgun”);
- Mechanism of action and symptoms (“hypoxia” and “peaceful way”); and
- Descriptions of experiences, thoughts, and suicide notes (“failed attempt,” “feel like,” and “I’m sorry”).

Figures S2 and S3 in Multimedia Appendix 1 present all the topics.

Sources and Consumer Locations

Overview

Based on the aforementioned topic modeling experiment results and supplemented with manual qualitative inspection of SN-related posts, we curated a list of phrases related to sourcing (eg, “seller,” “bought,” and “purchase”). The complete list is given in Table S2 in Multimedia Appendix 1. Since we collected data from Sanctioned Suicide on October 7, 2022, we extrapolated the frequency of posts from 279 days before (January 1, 2022, to October 7, 2022) by multiplying the per-day frequency with the total number of days in the year. Frequency analysis of sourcing-related posts pertaining to SN shows a sharp rise in “purchase” toward the end of 2019, with the highest raw annual “purchase” frequency observed in 2020 (Figure 7). Figure 8 shows the sourcing frequency (purchase frequency) of SN normalized by the posting frequency of SN-related topics. We found that discussions about the sourcing of SN gradually increased over the years.

Figure 7. Raw yearly purchase frequency of sodium nitrite from Sanctioned Suicide posts from March 2018 to December 2022 (extrapolated to the period from October to December 2022).

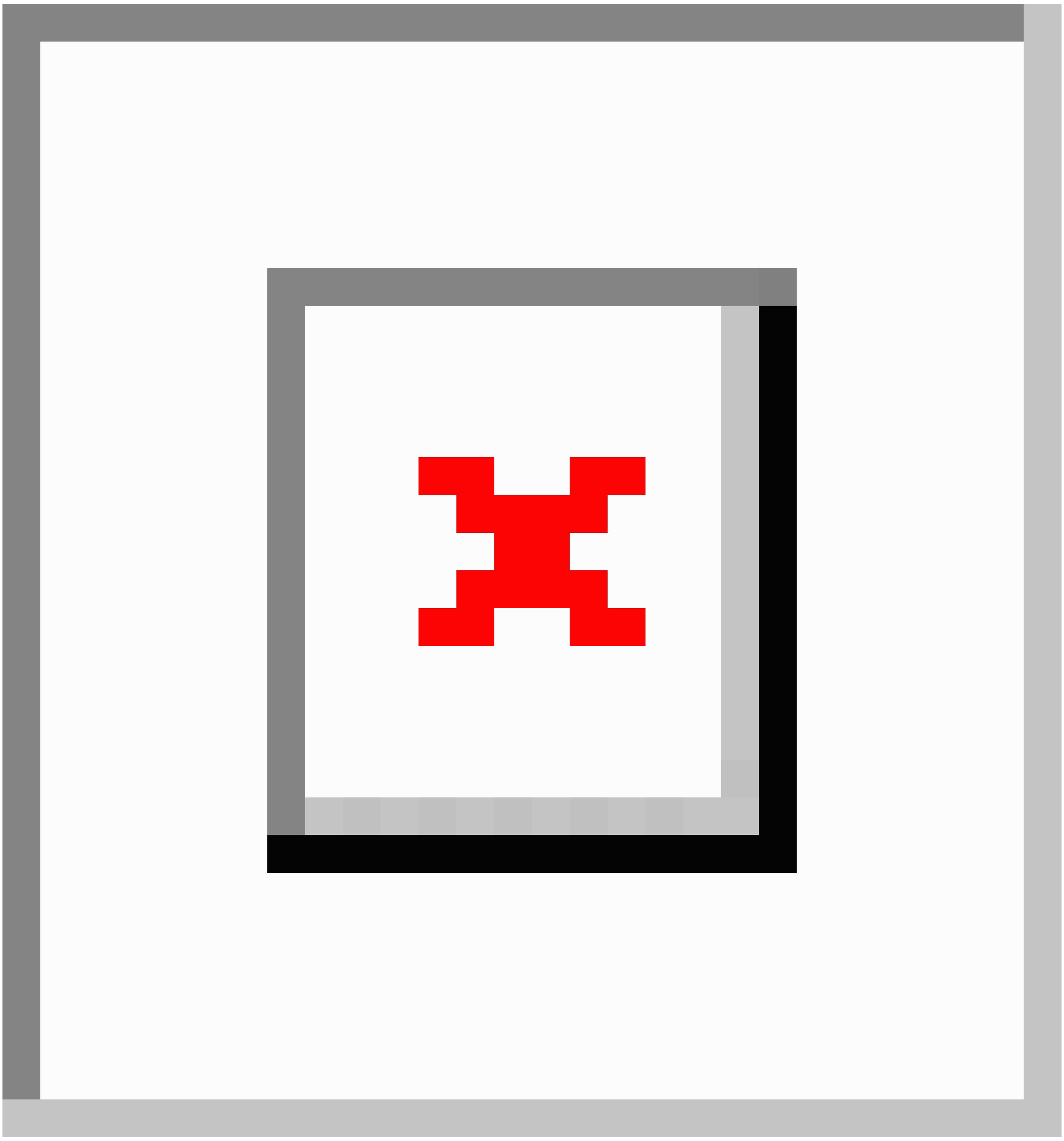
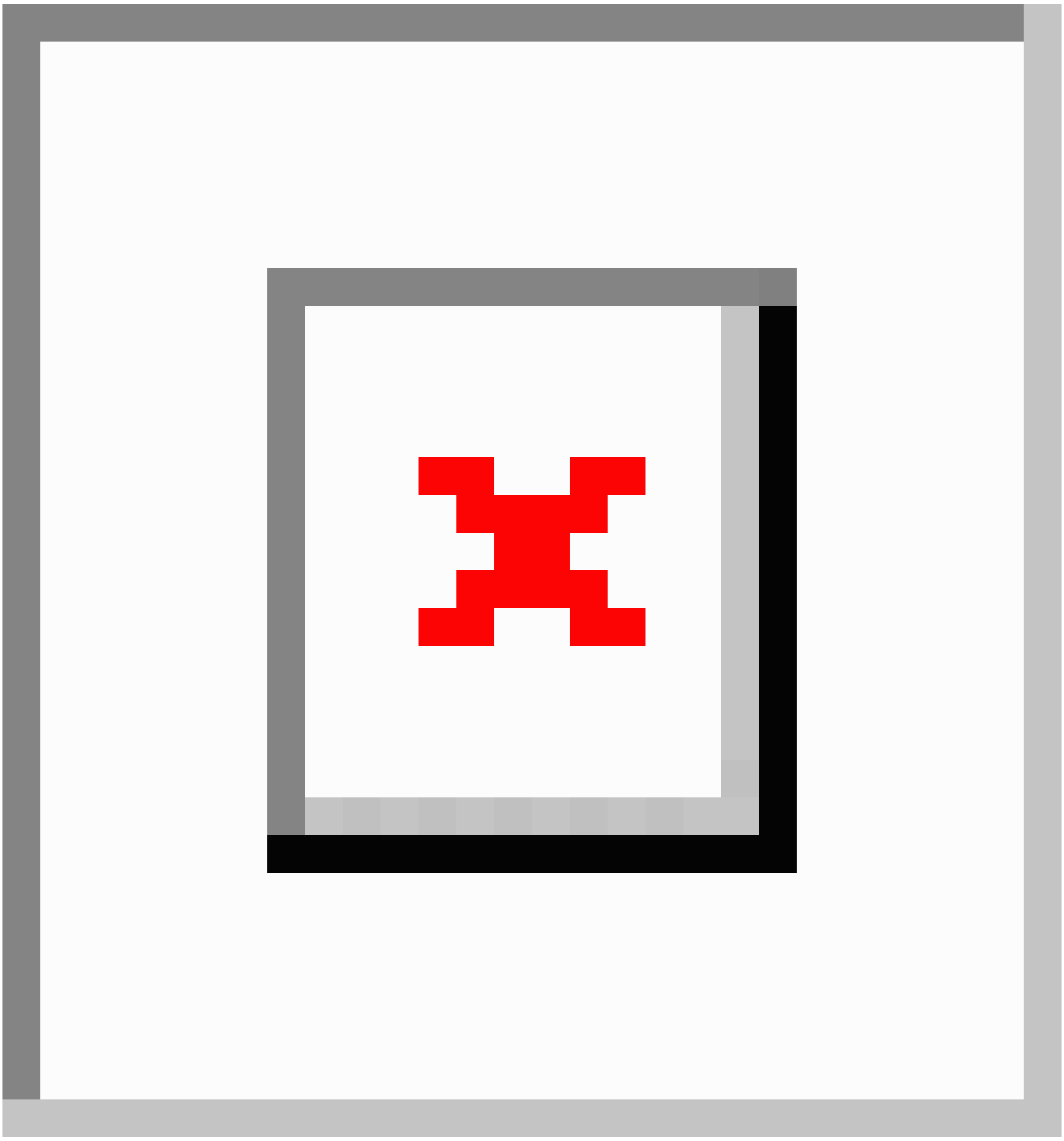


Figure 8. Normalized yearly purchase frequency of sodium nitrate from Sanctioned Suicide posts from March 2018 to December 2022 (extrapolated to the period from October to December 2022).



SN Sources and Consumer Locations

The NER approach detected locations that included countries, states, cities, counties, etc. We manually analyzed the detected locations to create a location-mapping dictionary, which was used to map cities to countries. The United States, the United Kingdom, Canada, China, and Germany were found to be the most popular potential locations for obtaining or using SN (Table 1). Since the locations detected were primarily from the United States, we also mapped locations detected from within the United States to the state level. The states of California, New York, Florida, Texas, and Oregon were found to be mentioned the most often in association with SN.

Table . Identified geographical locations from Sanctioned Suicide posts, from March 2018 to September 2022, associated with the sourcing or use of sodium nitrite.

Rank	Country-level locations		US state–level locations	
	Country	Frequency, n	US state	Frequency, n
1	United States	1043	California	149
2	United Kingdom	774	New York	104
3	Canada	358	Florida	67
4	China	202	Texas	63
5	Germany	185	Oregon	50
6	Mexico	170	Washington	41
7	Australia	164	Pennsylvania	30
8	India	163	North Carolina and Virginia[a]	23
9	Netherlands	146	Illinois	20
10	Switzerland	144	Arizona	18
11	Russia	140	Massachusetts	17
12	France	128	Utah	16
13	Japan and Spain[a]	96	Michigan and Nevada[a]	15
14	Poland	90	Colorado	13

[a]Locations with equal frequencies, presented in alphabetical order.

Location Mention–Intent Classification

Although NER-based geographical sources are informative, we found that consumers often tend to circumvent naming the geographical and online sources from where they obtained SN. Consider the following post: "I ordered it off of the big river in Brazil website." NER identifies "Brazil" as a potential geographical source of SN in this example. However, the user is referring to the online source "Amazon" rather than the geographical source "Brazil." Furthermore, locations mentioned in posts may also refer to the consumer's location rather than the location where SN was sourced. This necessitated building a classifier to identify the intent of mentioning the location. Our proposed location mention–intent classifier achieved an accuracy of 72.92% on the unseen test data set, outperforming traditional machine learning–based baselines (Table 2).

Table . Performance of the transformer-based, location mention–intent classifier on Sanctioned Suicide posts from March 2018 to September 2022.

Model	Random classifier	Support vector machine classifier	Our model
Accuracy	0.25	0.34	0.72
Precision	0.25	0.37	0.73
Recall	0.25	0.32	0.76
F_1-score	0.25	0.35	0.74

Sources and Consumer Locations

Based on the NER and subsequent classification process, the United States, the United Kingdom, Canada, Australia, and China were found to be the most popular source locations for obtaining SN (Table 3). The states of California, New Mexico, Florida, Rhode Island, and Oklahoma were reported to be the most popular sourcing locations for SN within the United States.

Table . Locations where sodium nitrite was potentially sourced from, as obtained from Sanctioned Suicide posts from March 2018 to September 2022.

Rank	Country-level locations		US state–level locations	
	Country	Frequency, n	State	Frequency, n
1	United States	586	California	55
2	United Kingdom	303	New Mexico	44
3	Canada	129	Florida	24
4	Australia	122	Rhode Island	21
5	China	83	Oklahoma	13
6	Germany and Mexico[a]	80	Virginia	12
7	India	67	Minnesota, New York, North Dakota, and Vermont[a]	8
8	Netherlands	53	Arizona, Georgia, Maryland, and Oregon[a]	7
9	Russia	45	Alabama, Illinois, Massachusetts, and Texas[a]	6

[a]Locations with equal frequencies, presented in alphabetical order.

The consumers who were interested in or attempted to obtain SN primarily were from the United States, the United Kingdom, Canada, Mexico, and Australia (Table 4). Within the United States, the states of California, New York, Texas, Florida, and Pennsylvania were found to be the most common locations of consumers attempting to obtain SN. For user-level post frequency, the mean was 72.63 (median 13, IQR 4-46; range 1-14,795).

Table . Locations where consumers attempting to procure sodium nitrite were from, as obtained from the Sanctioned Suicide posts from March 2018 to September 2022.

Rank	Country-level		US State-level	
	Country	Frequency, n	State	Frequency, n
1	United States	150	California	22
2	United Kingdom	107	New York and Texas[a]	10
3	Canada	41	Florida	6
4	Mexico	23	Pennsylvania	5
5	Australia	22	Oregon	4
6	China	21	Colorado, Maryland, Massachusetts, Oklahoma, and Washington[a]	3
7	India	15	Alaska, Illinois, New Jersey, Ohio, and Utah[a]	2
8	Russia and Switzerland[a]	12	Arizona, Connecticut, Georgia, Indiana, Kentucky, Louisiana, Michigan, Mississippi, South Carolina, and Wisconsin[a]	1
9	Brazil	10	N/A[b]	N/A

[a]Locations with equal frequencies, presented in alphabetical order.

[b]N/A: not applicable.

Potential Online Sources

Our NER-based approach also revealed potential online sources of obtaining SN or information about obtaining SN. Online marketplaces were the most commonly mentioned potential sources of SN (Table 5).

Table . Possible internet-based sources of sodium nitrite mentioned from Sanctioned Suicide posts from March 2018 to September 2022.

Rank	Online source	Frequency, n
1	Online	15,601
2	Google	6314
3	YouTube	5173
4	Amazon	2909
5	Facebook	2837
6	eBay	2373
7	Pharmacy	800
8	Walmart	621
9	Online pharmacy	593
10	Craigslist	148
11	Alibaba	84
12	Etsy	81
13	CVS	69
14	Tesco	51
15	Walgreens	45
16	AliExpress	26
17	Taobao	4

Discussion

Principal Findings

Our study is the first to conduct a comprehensive, NLP-based assessment of a large, popular, and public suicide forum. Our findings show that SN is the most popular method of suicide discussed on the forum, perhaps indicating the rising popularity of SN in real life. The trends we discovered suggest that the popularity of SN might still be increasing. Our study also revealed topics associated with SN discussions, among which sourcing was a common one. The application of automated NLP methods such as NER and classification enabled us to rapidly aggregate the locations and sources and compute their frequencies. Our findings and the data mining resources we are releasing with this paper will aid much-needed future research on this topic.

Online Sources of SN

Our analyses show the distinct role of online marketplaces as a source of SN (Table 5). As a common food additive approved for use [29] in several countries, such as the United States [30], the United Kingdom [31], New Zealand, and Australia [32], SN is widely available for procurement. However, SN is now listed as a poison in the United Kingdom and, thus, is considered to be a reportable substance whose sale is regulated and requires an Explosive Precursors and Poisons license [31]. Some online marketplaces, such as Etsy and eBay [33], have now implemented restrictions prohibiting the sale of SN through their website [28]. Overall, however, there remains a wide availability of products containing SN on internet marketplaces.

Trends in SN Use for Self-Harm

SN as a suicide mechanism has been reported in the medical literature as early as 1979 [34]. Since then, there have been a limited number of case reports in the literature through 2019, including one paper presenting 10 cases of SN consumption with the intent for self-harm [6]. Since 2019, case reports of SN-related deaths due to intentional consumption have increased [28], which is reflected in both the online community data as well as reported exposures (Figure 4) and deaths from official databases (Figure 5). This trend of rising instances of SN use for self-harm is concerning and may benefit from broader scientific interest. The literature on the topic is still sparse though, particularly at the intersection of SN and social media or internet-based data.

Utility of Internet-Based Data

Internet-based data hold substantial potential for the surveillance of suicide methods, particularly emerging topics such as the one we studied in this paper. Recent advances in data-centric artificial intelligence methods, particularly NLP, have opened up opportunities for rapidly analyzing such data, as we did in this study. While our paper is the first to take such a data-driven approach to fully describe the contents of this forum from an epidemiologic perspective, other recent papers have attempted to analyze data from it. Sartori et al [35], for example, investigated the impact of COVID-19 by studying posts from this forum, and they found that COVID-19 appeared to be indirectly connected to causes of distress for the users, such as anxiety for the economy, but not directly to the growth of users on the forum. Dilkes [36] took a more linguistic investigation approach and analyzed changes in language to evaluate the social and psychological effect of participation in the forum. In a more recent commentary, Dinis-Oliveira and Durão [37] further highlighted the importance of studying the forum, how it plays a role in providing guidance on how to use SN as a means of suicide, and the rapid increase in its popularity. Although our study takes a necessary next step in the use of this data source for public health work, this information can also be leveraged to address this emerging public health problem in the United States and globally in the form of locally targeted interventions, such as notices for emergency personnel about signs, symptoms, and treatment in locations prone to the issue [38,39].

Conclusion

In this study, we adopted a data-driven approach to analyze the trends in SN mentions on an online suicide forum using NLP and machine learning–based techniques. Our findings show that online forums can be an important source of information about emerging trends in suicide mechanisms. We also show that it is possible to obtain geographical trends of use and sourcing with our proposed location mention–intent classifier, with high accuracy. Since suicide is a rising concern in the United States and worldwide, we believe our study can be key to understanding temporal trends in suicide mechanisms and provide insights to the public health community by leveraging large amounts of online data sources.

Limitations and Future Work

Since we used Sanctioned Suicide as our primary data source, a limitation of our study is that the distribution of users appears to be largely from the United States. This is reflected in both the location-based analyses and the close association of Sanctioned Suicide trends with the CDC WONDER and the NPDS data. Differences in access to the internet and cyber literacy across the world further contribute to this distribution. The posts used in this study were in English, which may also have caused users from non–English-speaking countries to be underrepresented. Our NLP and machine learning methods also impose some limitations—the classification abilities of our models are not perfect, and errors can affect downstream tasks. In our study, however, the presence of a large volume of data helps to offset the influence of individual errors; furthermore, the use of machine learning is necessary in analyzing data that are too large to qualitatively assess by manual processes.

In the future, we aim to attempt to improve our NER and classification approaches so that more accurate information about sources and locations can be obtained in close to real time. We will also attempt to develop NLP tools that can automatically discover novel substances and mechanisms gaining popularity in that community. The resources we are sharing with this publication (lexicons, lists of phrases and keywords, and language models) are intended to support researchers to conduct their own data-driven studies on the topic.

Intentional SN ingestion remains an ongoing concern for suicide prevention work. The ease of accessibility and lethality of the substance present a unique challenge for public health efforts. As nations consider both policy and programmatic options for enhanced prevention opportunities, the use of online data will remain critical to understand emerging trends in a timely fashion.

Acknowledgments

The findings and conclusions in this publication are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Authors' Contributions

SD, DW, and SR conducted the data analyses and investigations. SD, DW, SAS, and AS contributed to the methodology. SL performed data collection and curation. All authors contributed to the original draft. KAM, WK, SAS, and AS contributed to validation and editing. AS supervised the project.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full list of suicide mechanism keywords, posting frequency on Sanctioned Suicide, full list of sourcing-related keywords, sodium nitrate–related topics: unigrams and bigrams, and code snippet.
[DOCX File, 323 KB - mental_v11i1e53730_app1.docx]

References

1. Curtin SC, Hedegaard H, Ahmad FB. Provisional numbers and rates of suicide by month and demographic characteristics: United States, 2020. : National Center for Health Statistics; 2021 Nov NVSS-Vital Statistics Rapid Release no. 16 URL: https://archive.hshsl.umaryland.edu/handle/10713/17347 [accessed 2023-05-30]
2. Curtin SC, Garnett MF, Ahmad FB. Provisional numbers and rates of suicide by month and demographic characteristics: United States, 2021. : National Center for Health Statistics; 2022 Sep NVDD Vital Stat Rapid Release no. 24 URL: https://stacks.cdc.gov/view/cdc/120830 [accessed 2024-04-09]

3. CDC WONDER. Centers for Disease Control and Prevention. URL: https://wonder.cdc.gov/ [accessed 2023-07-07]
4. Wang J, Sumner SA, Simon TR, et al. Trends in the incidence and lethality of suicidal acts in the United States, 2006 to 2015. JAMA Psychiatry 2020 Jul 1;77(7):684-693. [doi: 10.1001/jamapsychiatry.2020.0596] [Medline: 32320023]
5. WISQARS fatal and nonfatal injury reports. Centers for Disease Control and Prevention. URL: https://wisqars.cdc.gov/reports/ [accessed 2023-05-30]
6. Stephenson L, Wills S, van den Heuvel C, Humphries M, Byard RW. Increasing use of sodium nitrite in suicides-an emerging trend. Forensic Sci Med Pathol 2022 Sep;18(3):311-318. [doi: 10.1007/s12024-022-00471-8] [Medline: 35334075]
7. Duong D. Troubling rise in suicides linked with common food preservative. CMAJ 2022 Aug 8;194(30):E1070-E1071. [doi: 10.1503/cmaj.1096009] [Medline: 35940621]
8. Durão C, Pedrosa F, Dinis-Oliveira RJ. A fatal case by a suicide kit containing sodium nitrite ordered on the internet. J Forensic Leg Med 2020 Jul;73:101989. [doi: 10.1016/j.jflm.2020.101989] [Medline: 32658747]
9. Sedhai YR, Atreya A, Basnyat S, Phuyal P, Pokhrel S. The use of sodium nitrite for deliberate self-harm, and the online suicide market: should we care? Med Leg J 2022 Jun;90(2):79-80. [doi: 10.1177/0025817221998119] [Medline: 33906496]
10. McCann SD, Tweet MS, Wahl MS. Rising incidence and high mortality in intentional sodium nitrite exposures reported to US poison centers. Clin Toxicol (Phila) 2021 Dec;59(12):1264-1269. [doi: 10.1080/15563650.2021.1905162] [Medline: 33787434]
11. Khan H, Barber C, Azrael D. Suicide by sodium nitrite poisoning: findings from the National Violent Death Reporting System, 2018-2020. Suicide Life Threat Behav 2024 Apr;54(2):310-316. [doi: 10.1111/sltb.13043] [Medline: 38251179]
12. Niederkrotenthaler T, Stack S, Till B, et al. Association of increased youth suicides in the United States with the release of 13 Reasons Why. JAMA Psychiatry 2019 Sep 1;76(9):933-940. [doi: 10.1001/jamapsychiatry.2019.0922] [Medline: 31141094]
13. Sumner SA, Burke M, Kooti F. Adherence to suicide reporting guidelines by news shared on a social networking platform. Proc Natl Acad Sci U S A 2020 Jul 14;117(28):16267-16272. [doi: 10.1073/pnas.2001230117] [Medline: 32631982]
14. Niederkrotenthaler T, Voracek M, Herberth A, et al. Role of media reports in completed and prevented suicide: Werther v. Papageno effects. Br J Psychiatry 2010 Sep;197(3):234-243. [doi: 10.1192/bjp.bp.109.074633] [Medline: 20807970]
15. Twohey M, Dance GJX. Where the despairing log on, and learn ways to die. The New York Times. 2021 Dec 9. URL: https://www.nytimes.com/interactive/2021/12/09/us/where-the-despairing-log-on.html [accessed 2023-05-30]
16. Rules and FAQ. Sanctioned Suicide. URL: https://sanctioned-suicide.net/threads/rules-and-faq.4/ [accessed 2023-05-30]
17. H. Rept. 117-403 - Departments of Labor, Health and Human Services, and Education, and Related Agencies Appropriations Bill, 2023. Congress.gov. URL: https://www.congress.gov/congressional-report/117th-congress/house-report/403/1?s=1&r=54 [accessed 2024-04-18]
18. Suicide prevention resource for action: a compilation of the best available evidence. Centers for Disease Control and Prevention. 2022. URL: https://www.cdc.gov/suicide/resources/prevention.html [accessed 2024-04-18]
19. Sarker A, Gonzalez-Hernandez G. An unsupervised and customizable misspelling generator for mining noisy health-related text sources. J Biomed Inform 2018 Dec;88:98-107. [doi: 10.1016/j.jbi.2018.11.007] [Medline: 30445220]
20. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv. Preprint posted online on Mar 11, 2022. [doi: 10.48550/arXiv.2203.05794]
21. Industrial-strength natural language processing in Python. spaCy. URL: https://spacy.io/ [accessed 2023-05-30]
22. Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD. Stanza: a Python natural language processing toolkit for many human languages. In: Celikyilmaz A, Wen TH, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations: Association for Computational Linguistics; 2020:101-108. [doi: 10.18653/v1/2020.acl-demos.14]
23. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, et al, editors. Advances in Neural Information Processing Systems 30 (NIPS 2017): Curran Associates, Inc; 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html [accessed 2023-05-30]
24. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr;20(1):37-46. [doi: 10.1177/001316446002000104]
25. Liu Y, Ott M, Goyal N, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. Preprint posted online on Jul 26, 2019. [doi: 10.48550/arXiv.1907.11692]
26. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on May 24, 2019. [doi: 10.48550/arXiv.1810.04805]
27. Guo Y, Ge Y, Yang YC, Al-Garadi MA, Sarker A. Comparison of pretraining models and strategies for health-related social media text classification. Healthcare (Basel) 2022 Aug 5;10(8):1478. [doi: 10.3390/healthcare10081478] [Medline: 36011135]
28. Mudan A, Repplinger D, Lebin J, Lewis J, Vohra R, Smollin C. Severe methemoglobinemia and death from intentional sodium nitrite ingestions. J Emerg Med 2020 Sep;59(3):e85-e88. [doi: 10.1016/j.jemermed.2020.06.031] [Medline: 32713620]
29. GSFA online food additive group details for nitrites. Food and Agriculture Organization of the United Nations. URL: https://www.fao.org/gsfaonline/groups/details.html?id=151 [accessed 2023-05-30]

30. Food additive status list. US Food & Drug Administration. 2022 Aug. URL: https://www.fda.gov/food/food-additives-petitions/food-additive-status-list [accessed 2023-05-30]
31. Supplying explosives precursors and poisons. GOV.UK. URL: https://www.gov.uk/government/publications/supplying-explosives-precursors/supplying-explosives-precursors-and-poison [accessed 2023-05-30]
32. Australia New Zealand Food Standards Code - Standard 1.2.4 - labelling of ingredients. Australian Government, Federal Register of Legislation. URL: https://www.legislation.gov.au/F2008B00604/2013-02-21/text [accessed 2023-05-30]
33. Twohey M, Dance GJX. Lawmakers press Amazon on sales of chemical used in suicides. The New York Times. 2022 Feb 4. URL: https://www.nytimes.com/2022/02/04/technology/amazon-suicide-poison-preservative.html [accessed 2023-05-30]
34. Standefer JC, Jones AM, Street E, Inserra R. Death associated with nitrite ingestion: report of a case. J Forensic Sci 1979 Oct;24(4):768-771. [Medline: 541641]
35. Sartori E, Pajola L, Da San Martino G, Conti M. The impact of COVID-19 on online discussions: the case study of the Sanctioned Suicide forum. In: Ding Y, Tang J, Sequeda J, et al, editors. WWW '23: Proceedings of the ACM Web Conference 2023: Association for Computing Machinery; 2023:4060-4064. [doi: 10.1145/3543507.3583879]
36. Dilkes J. Quantifying changes in language to evaluate the social and psychological effect of participation in a pro-choice suicide forum. In: Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media: Association for the Advancement of Artificial Intelligence; 2022. [doi: 10.36190/2022.69]
37. Dinis-Oliveira RJ, Durão C. When the antidote for cyanide poisonings becomes a nightmare: an alarming outbreak of suicides using kits containing sodium nitrite. Forensic Sci Res 2023 Jun 4;8(2):170-171. [doi: 10.1093/fsr/owad015] [Medline: 37621451]
38. University of Maryland School of Pharmacy. Sodium nitrite poisoning. ToxTidbits. 2022 Dec. URL: https://www.mdpoison.com/media/SOP/mdpoisoncom/ToxTidbits/2022/SodiumNitrite-Dec-2022.pdf [accessed 2024-04-09]
39. Pritzker JB, Ezike N. Health alert. Illinois Department of Public Health. 2019. URL: https://www.dhs.state.il.us/OneNetLibrary/27896/documents/By_Division/MentalHealth/2020/CommunityAlerts/12132019-CommAlert-HealthAlert_sodiumNitrite.pdf [accessed 2024-04-09]

Abbreviations

CDC: Centers for Disease Control and Prevention
NER: named entity recognition
NLP: natural language processing
NPDS: National Poison Data System
SN: sodium nitrite
WONDER: Wide-Ranging Online Data for Epidemiologic Research

Edited by J Torous; submitted 17.10.23; peer-reviewed by MG Kim, N Wang, P Batterham; revised version received 11.03.24; accepted 12.03.24; published 02.05.24.

Please cite as:
Das S, Walker D, Rajwal S, Lakamana S, Sumner SA, Mack KA, Kaczkowski W, Sarker A
Emerging Trends of Self-Harm Using Sodium Nitrite in an Online Suicide Community: Observational Study Using Natural Language Processing Analysis
JMIR Ment Health 2024;11:e53730
URL: https://mental.jmir.org/2024/1/e53730
doi:10.2196/53730

Asynchronous Versus Synchronous Screening for Depression and Suicidality in a Primary Health Care System: Quality Improvement Study

Amelia Sattler[1], MD; Julia Dunn[1,2], MSc; Marleni Albarran[1], BS; Charlotte Berger[3], RPh; Ana Calugar[4], MHSA; John Carper[5], MD; Lalitha Chirravuri[5], BS; Nadine Jawad[6], MPhil; Mira Zein[1,2], MD, MPH; Mark McGovern[1,2], PhD

1

2

3

4

5

6

Corresponding Author:
Amelia Sattler, MD

Abstract

Background: Despite being a debilitating, costly, and potentially life-threatening condition, depression is often underdiagnosed and undertreated. Previsit Patient Health Questionnaire-9 (PHQ-9) may help primary care health systems identify symptoms of severe depression and prevent suicide through early intervention. Little is known about the impact of previsit web-based PHQ-9 on patient care and safety.

Objective: We aimed to investigate differences among patient characteristics and provider clinical responses for patients who complete a web-based (asynchronous) versus in-clinic (synchronous) PHQ-9.

Methods: This quality improvement study was conducted at 33 clinic sites across 2 health systems in Northern California from November 1, 2020, to May 31, 2021, and evaluated 1683 (0.9% of total PHQs completed) records of patients endorsing thoughts that they would be better off dead or of self-harm (question 9 in the PHQ-9) following the implementation of a depression screening program that included automated electronic previsit PHQ-9 distribution. Patient demographics and providers' clinical response (suicide risk assessment, triage nurse connection, medication management, electronic consultation with psychiatrist, and referral to social worker or psychiatrist) were compared for patients with asynchronous versus synchronous PHQ-9 completion.

Results: Of the 1683 patients (female: n=1071, 63.7%; non-Hispanic: n=1293, 76.8%; White: n=831, 49.4%), Hispanic and Latino patients were 40% less likely to complete a PHQ-9 asynchronously (odds ratio [OR] 0.6, 95% CI 0.45-0.8; *P*<.001). Patients with Medicare insurance were 36% (OR 0.64, 95% CI 0.51-0.79) less likely to complete a PHQ-9 asynchronously than patients with private insurance. Those with moderate to severe depression were 1.61 times more likely (95% CI 1.21-2.15; *P*=.001) to complete a PHQ-9 asynchronously than those with no or mild symptoms. Patients who completed a PHQ-9 asynchronously were twice as likely to complete a Columbia-Suicide Severity Rating Scale (OR 2.41, 95% CI 1.89-3.06; *P*<.001) and 77% less likely to receive a referral to psychiatry (OR 0.23, 95% CI 0.16-0.34; *P*<.001). Those who endorsed question 9 "more than half the days" (OR 1.62, 95% CI 1.06-2.48) and "nearly every day" (OR 2.38, 95% CI 1.38-4.12) were more likely to receive a referral to psychiatry than those who endorsed question 9 "several days" (*P*=.002).

Conclusions: Shifting depression screening from in-clinic to previsit led to a dramatic increase in PHQ-9 completion without sacrificing patient safety. Asynchronous PHQ-9 can decrease workload on frontline clinical team members, increase patient self-reporting, and elicit more intentional clinical responses from providers. Observed disparities will inform future improvement efforts.

(JMIR Ment Health 2024;11:e50192) doi:10.2196/50192

KEYWORDS

depression diagnosis; primary health care methods; electronic health records utilization; quality improvement; web-based universal screening methods; suicide prevention and control; screening; depression; asynchronous; synchronous; primary care; suicide; intervention; prevention

Introduction

In 2020, approximately 21 million adults in the United States experienced at least 1 major depressive episode, with only 66% receiving treatment, and nearly 46,000 people dying from suicide [1,2]. Primary care health systems can help prevent suicide by effectively screening and connecting patients with early intervention and treatment for depression [3,4].

The relationship between depression and the risk of suicide is well established and highlights the urgency of proactive intervention [5]. Depression is the most common psychiatric disorder in people who die by suicide [6]. A majority of individuals who die by suicide visited their primary care provider in the year preceding their death [7-9]. Thus, primary care is a pivotal setting for identifying suicide risk and initiating mental health care, starting with the implementation of effective screening processes [3,4].

Despite guidelines and recommendations for preventive annual depression screening that includes suicide risk assessment, health systems face a dilemma in integrating these measures into routine care [10-12]. Little evidence exists on the predictors and outcomes associated with asynchronous depression screening programs, such as which patient groups are more likely to engage in asynchronous depression screening and whether asynchronous screening affects a provider's clinical response [13-15]. Asynchronous previsit patient questionnaires can ease the administrative burden on frontline clinical staff and may help to efficiently identify problems ahead of primary care visits. However, concerns exist regarding the need for an immediate clinical response and the potential liability associated with patient responses indicating a high risk of severe depression that may result in self-harm or suicide. This concern results in health systems compromising by administering incomplete asynchronous depression screening that excludes questions overtly asking patients to report thoughts of self-harm or wishing they were dead. One resolution to this dilemma was the development of the Patient Health Questionnaire-8 (the 8-item version). This measure is an adaptation of the more widely adopted and validated Patient Health Questionnaire-9 (PHQ-9; the 9-item version) but does not include the ninth question, which is about suicidal thoughts [16].

Most depression screening and suicide risk assessments currently occur in an in-person clinical encounter, allowing for immediate evaluation, triage, and management of high-risk patients [17-19]. However, the dramatic shift from direct face-to-face patient care to remote care in the context of the COVID-19 pandemic rendered synchronous clinic-based screening programs insufficient. Meanwhile, during the COVID-19 pandemic, the prevalence of symptoms of depressive disorders nearly quadrupled [20]. This dramatic increase in depressive symptoms during the pandemic further heightened the urgency to adapt and optimize screening processes [21].

Screening for suicide risk remotely and asynchronously raises concerns about the risk of a patient harming themselves when support is not immediately available [22,23], particularly because no standards exist for immediate response to electronic screening for self-harm and suicidal intent [13]. However, there is robust evidence that patients are unlikely to attempt or die by suicide within a week after positively endorsing question 9 on a PHQ-9 [24]. Stated plainly, while question 9 is an important marker of disease severity, it has little predictive utility for acute risk of suicide [16,24,25].

During the pandemic, our primary care system redesigned our depression screening program to be effective for both remote and in-clinic visits by including both asynchronous screening ahead of visits and synchronous screening during visits for those patients who did not complete the asynchronous PHQ-9. Our overarching goal was to increase depression screening during the pandemic while decreasing the administrative burden on overwhelmed frontline clinical teams. We launched an automated electronic previsit depression screening workflow with the support of a multidisciplinary integrated behavioral health (IBH) team. These efforts align with the National Action Alliance for Suicide Prevention's COVID-19 "guidance: screening for suicide risk during telehealth visits [22]. Our program has broad clinical implications, as it increased rates of identifying depression and suicide risk in our patients.

In this quality improvement study, we sought to describe and evaluate the differences in patient characteristics and provider responses based on asynchronous and synchronous completion of a depression screening questionnaire. Little evidence exists on the predictors and outcomes associated with remote depression screening programs [13-15]. Through this study, we aim to contribute valuable insights that can inform future strategies in suicide prevention within primary care settings.

Methods

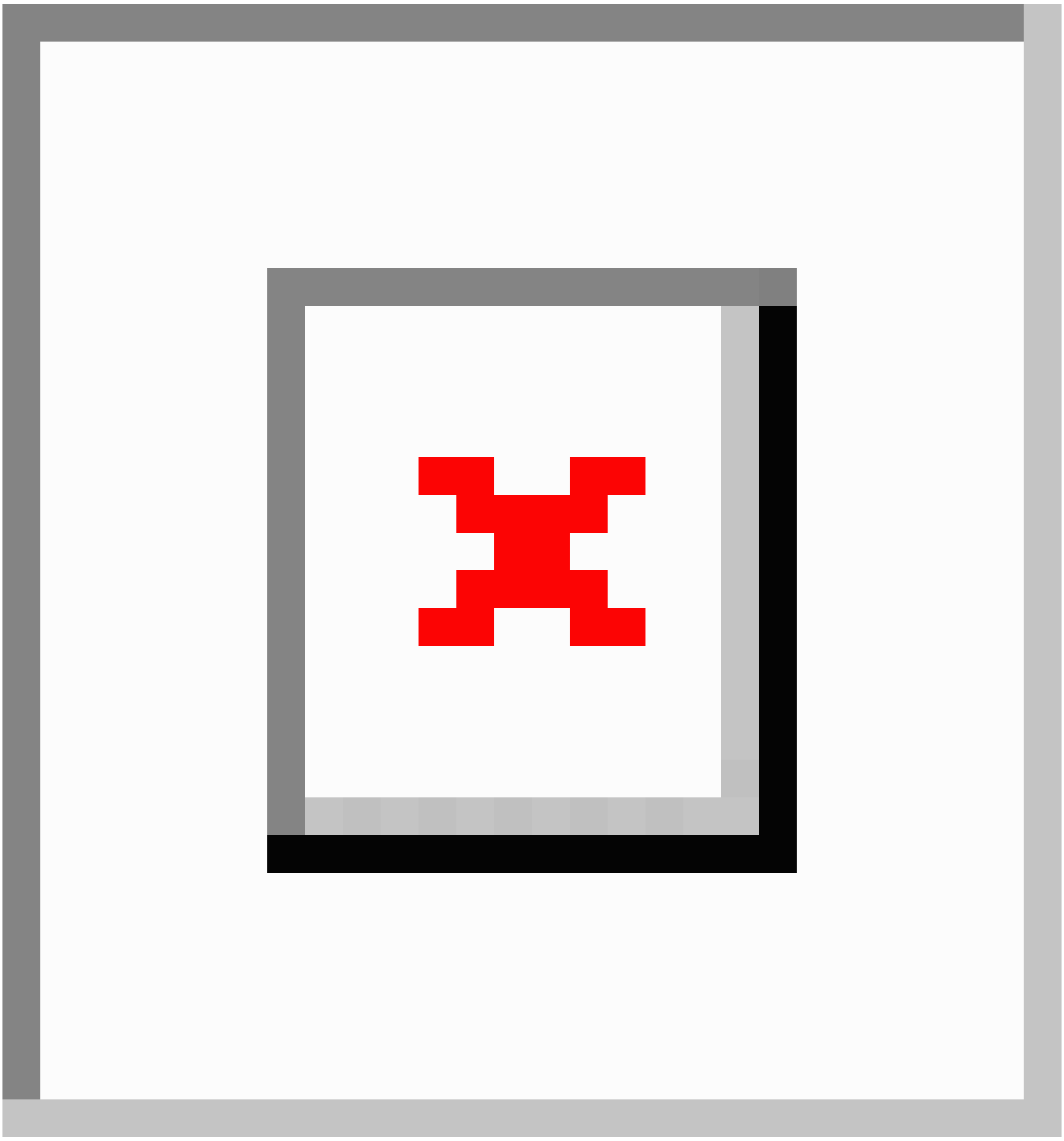
Study Setting

This quality improvement study was conducted in 2 primary care health systems in Northern California from November 1, 2020, to May 31, 2021. Stanford Health Care (SHC) includes 11 clinics and cares for 60,000 patients annually. University Healthcare Alliance (UHA) comprises 32 clinics and cares for 120,000 patients annually.

Depression Screening Program Description

Our universal depression screening program was launched in 33 clinics (28 at UHA and 5 at SHC) with an automated electronic questionnaire workflow and IBH-specific clinical resources to support primary care providers managing patients who screen positive for symptoms of depression. The PHQ-9 workflow included the following elements (Figure 1): (1) electronic PHQ presented to patients up to 3 days before a scheduled visit, as part of the electronic advance check-in procedure (branching PHQ-2-to-9 for patients without an existing diagnosis of depression or a full PHQ-9 for patients with an existing diagnosis of depression); (2) an automated electronic patient alert containing crisis resources displayed to patients at the time of questionnaire completion if they responded with a high-risk score; (3) reminders in the electronic medical record (EMR) prompting clinical teams to administer PHQs synchronously during visits if patients did not complete a previsit; (4) provider EMR alerts if a patient responded with a high-risk score; and (5) reminders in the EMR prompting providers to document follow-up plans for patients with symptoms of depression and curated decision-support tools in the EMR to assist providers with point-of-care clinical decision-making. Clinical resources at SHC included a triage nurse team, IBH social workers (SWs), and a consulting psychiatrist. Additional resources, such as psychiatrists and therapists, were available in the community for patients in both health systems.

Figure 1. Depression screening program process map. (1) Electronic PHQ questionnaires presented to patients up to 3 days before a scheduled clinic visit, embedded into the electronic check-in procedure; (2) automated electronic patient alerts containing crisis resources displayed to patients at the time of questionnaire completion if they responded with a high-risk score; (3) reminders in the EMR prompting clinical teams to administer depression screening questionnaires during visits if patients did not complete the previsit questionnaire; (4) provider EMR alerts if a patient responded with a high-risk score; (5) reminders in the EMR prompting providers to document follow-up plans for patients who screened positive for symptoms of depression and curated decision-support tools in the EMR to assist providers with point-of-care clinical decision-making. C-SSRS: Columbia-Suicide Severity Rating Scale; EMR: electronic medical record; PHQ: Patient Health Questionnaire; Q#9: question 9 of the PHQ questionnaire ("Over the last two weeks, how often have you been bothered by thoughts that you would be better off dead or of hurting yourself").



Measures

To assess the short-term impact of the screening program, we tracked total, asynchronous, and in-clinic PHQ-9 completion rates using an EMR report. The PHQ-9 has been widely applied and validated for use in primary care settings for the screening of depression with high sensitivity (74%) and specificity (91%) [26,27]. The ninth item evaluates passive thoughts of death or self-injury and has more limited utility as a measure of suicidal risk unless paired with a validated suicide risk assessment instrument and appropriate clinical responses [28]. Specifically, question 9 reads as follows: "Over the last two weeks, how often have you been bothered by thoughts that you would be better off dead or of hurting yourself." The response options were the following: Not at all (0), Several days (1), More than half the days (2), and Nearly every day (3). Patients with a score of 1 or higher on question 9 were included in the analysis. For this study, trained reviewers performed chart reviews of every patient who responded positively to the 9th question on a PHQ-9. Reviewers (MA and NJ) had experience navigating EMRs and were trained by an experienced clinician (AS) to identify relevant details and document them in a shared Excel worksheet.

Patient demographic characteristics were based on the EMR and included sex (male and female), race (White, Black or African American, Asian, and other/unknown), age in years, ethnicity (non-Hispanic, non-Latino and Hispanic/Latino), health insurance, encounter type (telemedicine or office visit), appointment status (completed, canceled, left, or no-show), and PHQ-9 score.

The record reviewers scanned the EMR for patient outcomes and providers' clinical responses to evaluate patient safety and provider adherence to recommended clinical workflows. Clinical response included actions that providers took to manage or improve a patient's depression, including prescribing new antidepressant medications; adjusting the dose of a current antidepressant medication; electronic consultation with a psychiatrist; referral to IBH SWs; referral to a psychiatry specialist; recommending ongoing management with an established behavioral health specialist outside our health system; linkage to the triage nurse for risk assessment; and completing a standardized suicide risk assessment, specifically the Columbia-Suicide Severity Rating Scale (C-SSRS). The C-SSRS is a psychometrically established and the gold standard for evaluating suicidal ideation severity and suicidal behaviors [29]. The scale contains 2 subsets of items with the first subset capturing the past-month severity of suicidal ideation and the second subset evaluating the past 3-month presence of suicide attempts.

Chart reviews for SHC patients included a review of free-text documentation in clinical notes and discrete data captured in standard reports. Chart reviews for UHA were limited to a review of discrete data captured in standard reports and focused review for any adverse patient events, such as suicide attempt or completion, that occurred between the time of questionnaire completion and the patient's clinical encounter. For this reason, clinical response outcomes were only measured in the SHC sample. Patient safety was inferred based on whether a patient attended a subsequent clinical encounter. For patients who were deceased at the time of retrospective chart review, the cause of death was determined based on chart review and classified as "suicide" or "other," such as terminal cancer.

Statistical Analysis

Descriptive statistics were used to summarize participant characteristics using frequency distributions, means, and medians. PHQ-9 categories were made based on the total score and grouped based on validated cut-offs of negative (0-4), mild (≥5), moderate (≥10), moderately severe (≥15), and severe (≥20) [25]. Because of the small sample size, Native Hawaiian, Pacific Islander, and American Indian/Alaska Native were recombined with "Other." Health insurance was categorized into Private, Medicare, Medicaid, or other/unknown. Age was categorized into groups as follows: 12-17, 18-39, 40-59, 60-79, and ≥80 years.

Cross-tabulations and univariate logistic regressions were used for the primary analyses to understand the relationship between the independent variables of various patient demographic characteristics and the dependent variable of completing the PHQ-9 asynchronously or synchronously. In a post hoc analysis, a 2-tailed *t* test was used to compare the mean PHQ-9 score among patients who completed the PHQ-9 asynchronously and synchronously.

Another series of univariate logistic regressions were fitted to evaluate for a difference in clinical response based on completing the PHQ-9 before or during a visit. In these models, we evaluated the relationships between the clinical response outcomes as the dependent variable and 2 different independent variables, including the asynchronous completion of the PHQ-9 and response to the ninth item of the PHQ-9. Bonferroni corrections were applied to mitigate type I errors. All analyses were conducted using Stata (version 17; StataCorp).

Ethical Considerations

This project was determined not to be human subject research by the Stanford Institutional Review Board (IRB-62520) [30]. The data used for this study were deidentified.

Results

Descriptive Characteristics

During the study period, our program sent 202,681 PHQs to patients ahead of their clinic visit, and patients completed approximately 184,700 (91%) of them. Of the total questionnaires sent, 119,389 (58.9%) were completed asynchronously: 43,979 (65.2%) at SHC and 75,410 (55.8%) at UHA.

Of the 184,700 PHQs completed, 1683 (0.9%) patients responded with a score of 1 or higher on question 9 and were included in the sample (Table 1): 456 patients (1.0% of total patients screened) at SHC and 1227 (0.9% of total patients screened) at UHA. Nearly half completed the PHQ-9 asynchronously (n=826, 49.4%). The sample included primarily female (n=1071, 63.6%) and non-Hispanic (n=1293, 76.8%) patients, and the mean age was 46.4 (SD 21.1) years. Most of the sample consisted of White (n=831, 49.4%) or Asian (n=314, 18.7%) patients. Most had private health insurance (n=1025, 60.0%) or Medicare (n=525, 31.2%). A minority (n=71, 4.2%) had Medicaid.

Table . Descriptive characteristics comparing synchronous and asynchronous completion of the Patient Health Questionnaire-9 (PHQ-9).

Characteristics		Synchronous (n=857)	Asynchronous (n=826)	Total (n=1683)
Sex, n (%)				
	Male	301 (49.2)	311 (50.8)	612 (36.4)
	Female	556 (51.9)	515 (48.1)	1071 (63.6)
Age (years), mean (SD)		47.3 (22.3)	45.3 (19.9)	46.4 (21.1)
Race, n (%)				
	American Indian/Alaska Native	6 (66.7)	3 (33.3)	9 (0.5)
	Asian	138 (43.9)	176 (56.1)	314 (18.7)
	Black or African American	66 (60.0)	44 (40.0)	110 (6.5)
	Native Hawaiian or Pacific Islander	13 (54.2)	11 (45.8)	24 (1.4)
	White	432 (52.0)	399 (48.0)	831 (49.4)
	Other/unknown	202 (51.1)	193 (48.9)	395 (23.5)
Ethnicity, n (%)				
	Hispanic/Latino	139 (61.2)	88 (38.8)	227 (13.5)
	Non-Hispanic/non-Latino	623 (48.2)	670 (51.8)	1293 (76.8)
	Other/unknown	95 (58.3)	68 (41.7)	163 (9.7)
Health insurance, n (%)				
	Private	482 (47.0)	543 (53.0)	1025 (61.0)
	Medicare	304 (57.9)	221 (42.1)	525 (31.2)
	Medicaid	33 (46.5)	38 (53.5)	71 (4.2)
	Other/unknown	38 (61.3)	24 (38.7)	62 (3.7)
Preferred language, n (%)				
	English	808 (50.8)	782 (49.2)	1590 (94.5)
	Spanish	15 (65.2)	8 (34.8)	23 (1.4)
	Other	34 (48.6)	36 (51.4)	70 (4.2)
Living status, n (%)				
	Alive	846 (50.8)	818 (49.2)	1664 (98.9)
	Deceased	11 (57.9)	8 (42.1)	19 (1.1)
Suicide attempt, n (%)				
	Yes	—[a]	—	0 (0)
	No	—	—	1683 (100)
Died by suicide, n (%)				
	Yes	—	—	0 (0)
	No	—	—	1683 (100)
PHQ-9 ninth item, n (%)				
	Several days	606 (50.3)	599 (49.7)	1205 (71.6)
	More than half the days	148 (48.5)	157 (51.5)	305 (18.1)
	Nearly every day	103 (59.5)	70 (40.5)	173 (10.3)

[a]Not applicable.

At the time of chart review, 18 (1.1%) patients were deceased. Based on a detailed chart review, none of these patients died by suicide. No patients in the sample were documented to have attempted suicide between the time of questionnaire completion and their clinical encounter. Most patients responded to question 9 with "several days" (n=1205, 71.6%), followed by "more than half the days" (n=305, 18.1%) and "nearly every day" (n=173, 10.3%).

Regarding clinical responses, a minority of patients (n=361, 20.8%) received a new medication or dose adjustment (Table 2). A subset of patients were referred to psychiatry (n=205, 33.7%), SW (n=53, 11.5%), an external specialist (n=96, 21.1%), or a triage nurse (n=7, 1.5%). Some providers placed an eConsult to Psychiatry (n=34, 7.4%), where the provider directly communicated with the psychiatrist about a clinical question and received a response within 3 business days. None were sent to the emergency room for urgent evaluation. Most patients completed a C-SSRS during their visit (n=1176, 74.5%).

Table . Clinical response in the Stanford Health Care (SHC) and University Healthcare Alliance (UHA) samples by asynchronous and synchronous completion of the Patient Health Questionnaire-9 (PHQ-9).

Clinical response			Questionnaire completion		
			Synchronous	Asynchronous	Total
SHC (n=456), n (%)					
	New medication or dose adjustment				
		No	57 (75.0)	295 (80.0)	361 (79.2)
		Yes	19 (25.0)	74 (20.1)	95 (20.8)
	eConsult psychiatry				
		No	73 (94.8)	343 (92.2)	426 (92.6)
		Yes	4 (5.2)	29 (7.8)	34 (7.4)
	Referral psychiatry				
		No	70 (41.7)	324 (75.4)	404 (66.3)
		Yes	98 (58.3)	106 (24.7)	205 (33.7)
	Referral social worker				
		No	69 (87.3)	328 (88.4)	408 (88.5)
		Yes	10 (12.7)	43 (11.6)	53 (11.5)
	Send to ER[a] for emergency evaluation				
		No	76 (100)	369 (100)	456 (100)
		Yes	0 (0)	0 (0)	0 (0)
	Continue with external specialist				
		No	60 (79.0)	289 (78.3)	360 (79.0)
		Yes	16 (21.1)	80 (21.7)	96 (21.1)
	Send to triage RN[b]				
		No	75 (98.7)	75 (98.7)	449 (98.5)
		Yes	6 (1.6)	1 (1.3)	7 (1.5)
UHA and SHC (n=1227), n (%)					
	C-SSRS[c] completed				
		No	276 (33.1)	126 (17.1)	402 (25.5)
		Yes	558 (66.9)	613 (82.9)	1176 (74.5)

[a]ER: emergency room.

[b]RN: registered nurse.

[c]C-SSRS: Columbia-Suicide Severity Rating Scale.

Patient Characteristics Associated With the Modality of Questionnaire Completion: Asynchronous or Synchronous

Several patient characteristics were associated with asynchronous or synchronous completion of the PHQ-9 (Table 3). The relationship between patient race and completion modality did not meet the level of statistical significance designated by the Bonferroni correction, but the trend is worth noting (P=.01). Asian patients were more likely to complete the PHQ-9 asynchronously than White patients (odds ratio [OR] 1.40, 95% CI 1.08-1.82; P=.01). Hispanic and Latino patients were less likely than non-Hispanic or non-Latino patients to complete the PHQ-9 asynchronously (OR 0.60, 95% CI 0.45-0.80; *P*<.001). Patient age was associated with modality of PHQ-9 completion, with patients aged 18 to 79 years having a higher likelihood of completing the PHQ-9 asynchronously than patients aged 80 years or older (*P*<.001). Patients with Medicare insurance were less likely (OR 0.64, 95% CI 0.51-0.79; *P*<.001) to complete the PHQ-9 asynchronously than patients with private insurance. Patients with office visits had lower likelihood of completing the PHQ-9 asynchronously than patients with telemedicine visits (OR 0.24, 95% CI 0.20-0.30). Those with moderate to severe depression symptoms were 1.61 times more likely (95% CI 1.21-2.15; *P*=.001) to complete screening asynchronously than those with no or mild symptomatology. Finally, the mean PHQ-9 score (16.4, SD 5.3) for patients who completed the screening asynchronously was significantly higher than that for patients who completed screening synchronously (mean 15.6, SD 5.7; *P*=.004).

Table . Logistic regression of the association between patient characteristics and whether patients completed the Patient Health Questionnaire-9 (PHQ-9) synchronously or asynchronously (n=1683).

Characteristics		Asynchronous PHQ-9	
		OR[a] (95% CI)	*P* value
Sex			**.30**
	Male	1	
	Female	1.11 (0.91-1.35)	
Race			**.01**
	White	1	
	Black or African American	0.72 (0.48-1.07)	.10
	Asian	1.40 (1.08-1.82)	.01
	Other/unknown	1.07 (0.83-1.37)	.60
Age group (years)			**.03**
	≥80	1	
	60-79	1.17 (0.78-1.77)	.50
	40-59	1.66 (0.10-2.49)	.02
	18-39	1.84 (1.25-2.72)	.002
	12-17	0.06 (0.01-0.24)	<.001
Ethnicity			**<.001**
	Non-Hispanic/non-Latino	1	
	Hispanic/Latino	0.60 (0.45-0.80)	
Health insurance			**<.001**
	Private	1	
	Medicare	0.64 (0.51-0.79)	<.001
	Medicaid	1.03 (0.64-1.68)	.90
	Other/unknown	0.55 (0.33-0.93)	.03
Encounter type			**<.001**
	Telemedicine	1	
	Office visit	0.24 (0.20-0.30)	
Appointment status			**.001**
	Completed	1	
	Cancelled, left, or no-show	5.55 (1.61-19.12)	
PHQ-9 categories			**.001**
	Normal or mild (0-10)	1	
	Moderate to severe (≥10)	1.61 (1.21-2.15)	

[a]OR: odds ratio; reference groups are represented by OR=1 in the table.

Differential Clinical Response to Suicide Risk by Asynchronous or Synchronous Modalities

The relationship between the timing of PHQ-9 completion and provider clinical response was evaluated (Table 4). Regardless of timing of PHQ-9 completion, there was a relationship between the response to question 9 and referral to psychiatry (*P*=.002). Those who reported suicidal thoughts "more than half the days" (OR 1.62, 95% CI 1.06-2.48; *P*=.03) or "nearly every day" (OR 2.38, 95% CI 1.38-4.12; *P*=.002) had a higher likelihood of receiving a referral to psychiatry than those who endorsed question 9 "several days." The likelihood of receiving a referral to psychiatry for those who completed a PHQ-9 asynchronously was lower (OR 0.23, 95% CI 0.16-0.34; *P*<.001) than for patients who completed the screening synchronously. In contrast, the likelihood of a patient completing the C-SSRS during the visit was 2.41 times higher (OR 2.41, 95% CI 1.89-3.06; *P*<.001) for patients who completed the PHQ-9 asynchronously than synchronously.

Table . Logistic regression of the association between completing the Patient Health Questionnaire-9 (PHQ-9) asynchronously or synchronously and clinical response for the Stanford Health Care (SHC; n=456) and University Healthcare Alliance (UHA; n=1227) samples.

Variable	SHC (n=456)												UHA and SHC (n=1683)	
	New medication or dose adjustment		eConsult psychiatry		Referral psychiatry		Referral social worker		Continue with external specialist		Send to triage RN[a]		C-SSRS[b] completed	
	OR[c] (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value	OR (95% CI)	*P* value
Timing of PHQ-9 Completion														
Synchronous	1	.30	1	.40	1	<.001	1	.80	1	.90	1	.80	1	<.001
Asynchronous	0.8 (0.42-1.34)		1.54 (0.53-4.52)		0.23 (0.16-0.34)		0.90 (0.43-1.89)		1.0 (0.60-1.90)		1.24 (0.15-10.45)		2.41 (1.89-3.06)	
Response to the 9th Item of the PHQ-9														
Several days	1	.80	1	.40	1	.002	1	.80	1	.90	1	.20	1	.07
More than half the days	0.80 (0.42-1.51)	.50	1.41 (0.58-3.43)	.40	1.62 (1.06-2.48)	.03	1.28 (0.62-2.63)	.50	1.07 (0.59-1.96)	.80	4.5 (0.89-22.70)	.07	1.09 (0.81-1.49)	.60
Nearly every day	0.98 (0.43-2.22)	.10	2.07 (0.74-5.81)	.20	2.38 (1.38-4.12)	.002	0.94 (0.32-2.81)	.90	1.20 (0.54-2.64)	.70	3.04 (0.31-29.9)	.30	0.68 (0.47-0.97)	.03

[a]RN: registered nurse.

[b]C-SSRS: Columbia-Suicide Severity Rating Scale.

[c]OR: odds ratio.

Discussion

Principal Findings

Shifting depression screening from an in-clinic task to a previsit questionnaire led to a dramatic increase in our system's ability to effectively screen patients for depression without compromising patient safety. The remote delivery of our depression screening and monitoring program saved front line clinical teams time spent administering the PHQs, offered patients a private setting for symptom reporting, and facilitated more intentional clinical responses from providers.

Consistent with previous studies demonstrating that PHQ-9 scores are higher when completed by patients on their own time via a personal device, our patients reported more severe symptoms on asynchronous PHQs compared to in-clinic PHQs [14,15]. In addition, of all patients who completed a PHQ-9, patients with moderate to severe symptoms were more likely to complete the PHQ-9 asynchronously, supporting the idea that patients with severe symptoms may feel more comfortable self-reporting outside of a clinic setting where they are not subject to time pressure or desire to please their provider [31].

Providers were also more likely to conduct a standardized suicide risk assessment when patients completed the PHQ-9 ahead of their visit. Evidence suggests that the ninth item of the PHQ-9 is an insufficient assessment tool for suicide risk and ideation and should be paired with a validated suicide risk assessment instrument [28]. In this study, once a patient completed a PHQ-9 asynchronously and endorsed question 9, the provider was immediately notified via a high-priority alert in the EMR. That previsit notification allowed providers to prepare to address and appropriately prioritize assessing risk of self-harm when agenda-setting for the clinical encounter [32-34]. In contrast, if a provider learned of a high-risk score during a visit, it may be more challenging to prioritize, particularly if that was not a primary reason for the patient presenting to the visit. Mental health concerns can be complex and time-consuming to address. However, these symptoms can also be life-threatening, requiring timely clinical action. Placing a referral to psychiatry is appropriate clinical management of a patient with severe depression. However, when pressed for time or taken by surprise during a visit, and particularly when confronted with severe symptoms, providers may be more likely to delegate further evaluation and management to a psychiatrist. Conversely, when given advance notice of a high-risk score, providers may be better able to prioritize a patient's mental health concern and take the time to more completely assess, triage, and manage a patient's depression at the point of care, reducing the need to refer and providing more timely clinical intervention.

Asynchronous PHQ-9 completion offers some potential advantages, such as increased reporting of severe symptoms and supporting more intentional clinical responses, and should be accessible to all patients seen in health systems offering remote screening. Since suicidal risk can fluctuate over time [35], asynchronous screening programs offer an opportunity to evaluate risk in patients on a more frequent basis. Universal depression screening programs, in general, can reduce bias by addressing important barriers, such as care team discomfort and screening based on a provider's subjective risk assessment [36]. Asynchronous screening can offer a convenient method for vulnerable patient groups with low health care engagement to access resources more directly in primary care [37]. However, technology-enabled programs may disproportionately exclude patients with limited access to digital tools, particularly those with low health or technology literacy [37,38]. We observed that disparities in engagement with PHQ-9 completion existed, despite automation. Specifically, Asian patients were more likely than White patients to complete the PHQ-9 asynchronously, and non-Hispanic or non-Latino patients were more likely to complete screening asynchronously compared to Hispanic or Latino patients. Patients with Medicare insurance were also less likely to complete the PHQ-9 asynchronously, which may be explained by older patients having more of a preference for in-person care than telemedicine [37]. Due to the relative advantages of previsit PHQ-9 completion, health systems may benefit from designing processes to reduce disparities in engagement with electronic tools, such as eCheck-in, and designing in-clinic processes to promote equal opportunities for patients who cannot complete the electronic questionnaire ahead of their visit.

Limitations

This study was limited in scope and only included data for patients who endorsed question 9 of the PHQ-9 during the program's first 6 months. Our patient population had a high rate of technology literacy, with approximately 96% of patients enrolled in the web-based patient portal at SHC and 94% of patients enrolled in the web-based patient portal at UHA. Despite widespread engagement with electronic tools, our results suggest health disparities in previsit remote questionnaire completion. Future research should explore whether certain patient groups are more likely to engage in asynchronous questionnaire completion, particularly patients at high risk for depression, and how to bridge this gap.

Another limitation is the lack of clinical response data for UHA patients. Most clinical documentation is available only in free-text portions of providers' encounter notes, therefore not available discretely in reports and only accessible via manual and detailed chart review, making review of these data on a large scale time prohibitive. Finally, although we believe that no patient attempted or died by suicide in the study population, we only had access to a single data source, the EMR. To increase the likelihood of identifying relevant adverse clinical outcomes related to PHQ-9 completion, during chart review our team placed specific attention on reviewing documentation between the time of questionnaire completion and the patient's associated clinical encounter. A documented clinical encounter or electronic communication from the patient in the period following questionnaire completion provided confirmation of their living status.

Conclusions

This study suggests that screening patients for depression outside of a clinical care setting does not compromise patient safety, may increase honest self-reporting for those with severe symptoms, and can allow for more deliberate clinical evaluation and response by providers. Access to asynchronous depression screening may offer essential shifts in patient-centered healthcare access and provider-driven clinical responses. Disparities in patient characteristics associated with asynchronous questionnaire completion present future opportunities to engage non-White and older patients in reporting severe symptoms. Furthermore, this study demonstrates the benefit of exploring processes that support providers' ability to offer more intentional clinical responses when severe symptoms are identified during clinical encounters. These findings may or may not be limited to the primary care health systems in which the study was conducted. Nonetheless, these conclusions suggest that health systems should consider examining this issue for themselves and potentially reconsider how depression and suicide risk screening are systematically conducted at their institution.

Acknowledgments

The authors acknowledge the invaluable support and contributions of Dr Christopher Sharp, Patrick Cimino and their team for implementing the asynchronous previsit questionnaires. The authors would also like to thank Dr Alan Pinyavat for his critical review and feedback on this manuscript. This project would not have been possible without the support of the Stanford Health

Care (SHC) and University Healthcare Alliance (UHA) population health teams, ambulatory operations leadership, as well as the engagement and dedication of the SHC and UHA faculty and staff.

Conflicts of Interest

None declared.

References

1. Key substance use and mental health indicators in the United States: results from the 2020 national survey on drug use and health. SAMHSA (Substance Abuse and Mental Health Services Administration). 2020. URL: https://www.samhsa.gov/data/sites/default/files/reports/rpt35325/NSDUHFFRPDFWHTMLFiles2020/2020NSDUHFFR1PDFW102121.pdf [accessed 2024-04-08]
2. Nonfatal injury data. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/injury/wisqars/nonfatal.html [accessed 2023-04-02]
3. Mann JJ, Michel CA, Auerbach RP. Improving suicide prevention through evidence-based strategies: a systematic review. Am J Psychiatry 2021 Jul;178(7):611-624. [doi: 10.1176/appi.ajp.2020.20060864] [Medline: 33596680]
4. Kemper AR, Hostutler CA, Beck K, Fontanella CA, Bridge JA. Depression and suicide-risk screening results in pediatric primary care. Pediatrics 2021 Jul;148(1):e2021049999. [doi: 10.1542/peds.2021-049999] [Medline: 34099503]
5. Hawton K, Comabella CCI, Haw C, Saunders K. Risk factors for suicide in individuals with depression: a systematic review. J Affect Disord 2013 May;147(1-3):17-28. [doi: 10.1016/j.jad.2013.01.004] [Medline: 23411024]
6. Cavanagh JTO, Carson AJ, Sharpe M, Lawrie SM. Psychological autopsy studies of suicide: a systematic review. Psychol Med 2003 Apr;33(3):395-405. [doi: 10.1017/s0033291702006943] [Medline: 12701661]
7. Luoma JB, Martin CE, Pearson JL. Contact with mental health and primary care providers before suicide: a review of the evidence. Am J Psychiatry 2002 Jun;159(6):909-916. [doi: 10.1176/appi.ajp.159.6.909] [Medline: 12042175]
8. De Leo D, Draper BM, Snowdon J, Kõlves K. Contacts with health professionals before suicide: missed opportunities for prevention. Compr Psychiatry 2013 Oct;54(7):1117-1123. [doi: 10.1016/j.comppsych.2013.05.007] [Medline: 23768696]
9. Ahmedani BK, Westphal J, Autio K, et al. Variation in patterns of health care before suicide: a population case-control study. Prev Med 2019 Oct;127:105796. [doi: 10.1016/j.ypmed.2019.105796] [Medline: 31400374]
10. Goldman ML, Shah RN, Bernstein CA. Depression and suicide among physician trainees: recommendations for a national response. JAMA Psychiatry 2015 May;72(5):411-412. [doi: 10.1001/jamapsychiatry.2014.3050] [Medline: 25738529]
11. Gaynes BN, West SL, Ford CA, et al. Screening for suicide risk in adults: a summary of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med 2004 May 18;140(10):822-835. [doi: 10.7326/0003-4819-140-10-200405180-00015] [Medline: 15148072]
12. Bolton JM, Gunnell D, Turecki G. Suicide risk assessment and intervention in people with mental illness. BMJ 2015 Nov 9;351:h4978. [doi: 10.1136/bmj.h4978] [Medline: 26552947]
13. Chan S, Li L, Torous J, Gratzer D, Yellowlees PM. Review of use of asynchronous technologies incorporated in mental health care. Curr Psychiatry Rep 2018 Aug 28;20(10):85. [doi: 10.1007/s11920-018-0954-3] [Medline: 30155593]
14. Torous J, Staples P, Shanahan M, et al. Utilizing a personal smartphone custom app to assess the Patient Health Questionnaire-9 (PHQ-9) depressive symptoms in patients with major depressive disorder. JMIR Ment Health 2015;2(1):e8. [doi: 10.2196/mental.3889] [Medline: 26543914]
15. BinDhim NF, Alanazi EM, Aljadhey H, et al. Does a mobile phone depression-screening app motivate mobile phone users with high depressive symptoms to seek a health care professional's help? J Med Internet Res 2016 Jun 27;18(6):e156. [doi: 10.2196/jmir.5726] [Medline: 27349441]
16. Wu Y, Levis B, Riehm KE, et al. Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis. Psychol Med 2020 Jun;50(8):1368-1380. [doi: 10.1017/S0033291719001314] [Medline: 31298180]
17. Gilbody S, Bower P, Fletcher J, Richards D, Sutton AJ. Collaborative care for depression: a cumulative meta-analysis and review of longer-term outcomes. Arch Intern Med 2006 Nov 27;166(21):2314-2321. [doi: 10.1001/archinte.166.21.2314] [Medline: 17130383]
18. Thota AB, Sipe TA, Byard GJ, et al. Collaborative care to improve the management of depressive disorders. Am J Prev Med 2012 May;42(5):525-538. [doi: 10.1016/j.amepre.2012.01.019] [Medline: 22516495]
19. Pfoh ER, Janmey I, Anand A, Martinez KA, Katzan I, Rothberg MB. The impact of systematic depression screening in primary care on depression identification and treatment in a large health care system: a cohort study. J Gen Intern Med 2020 Nov;35(11):3141-3147. [doi: 10.1007/s11606-020-05856-5] [Medline: 32495093]
20. Anxiety and depression household pulse survey. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/covid19/pulse/mental-health.htm [accessed 2022-12-01]
21. Nelson PA, Adams SM. Role of primary care in suicide prevention during the COVID-19 pandemic. J Nurse Pract 2020 Oct;16(9):654-659. [doi: 10.1016/j.nurpra.2020.07.015] [Medline: 32837401]

22. Guidance: screening for suicide risk during telehealth visits. National Action Alliance for Suicide Prevention. URL: https://theactionalliance.org/resource/covid-guidance-screening-suicide-risk-during-telehealth-visits [accessed 2022-09-01]
23. Vannoy SD, Robins LS. Suicide-related discussions with depressed primary care patients in the USA: gender and quality gaps. A mixed methods analysis. BMJ Open 2011 Jan 1;1(2):e000198. [doi: 10.1136/bmjopen-2011-000198] [Medline: 22021884]
24. Simon GE, Rutter CM, Peterson D, et al. Does response on the PHQ-9 depression questionnaire predict subsequent suicide attempt or suicide death? Psychiatr Serv 2013 Dec 1;64(12):1195-1202. [doi: 10.1176/appi.ps.201200587] [Medline: 24036589]
25. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
26. Costantini L, Pasquarella C, Odone A, et al. Screening for depression in primary care with Patient Health Questionnaire-9 (PHQ-9): a systematic review. J Affect Disord 2021 Jan 15;279:473-483. [doi: 10.1016/j.jad.2020.09.131] [Medline: 33126078]
27. Arroll B, Goodyear-Smith F, Crengle S, et al. Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. Ann Fam Med 2010;8(4):348-353. [doi: 10.1370/afm.1139] [Medline: 20644190]
28. Na PJ, Yaramala SR, Kim JA, et al. The PHQ-9 item 9 based screening for suicide risk: a validation study of the Patient Health Questionnaire (PHQ)-9 item 9 with the Columbia-Suicide Severity Rating Scale (C-SSRS). J Affect Disord 2018 May;232:34-40. [doi: 10.1016/j.jad.2018.02.045] [Medline: 29477096]
29. The Lighthouse Project. The Columbia Lighthouse Project. URL: https://cssrs.columbia.edu/ [accessed 2022-12-04]
30. Quality improvement research tools. Stanford Medicine Center for Improvement. URL: https://smci.stanford.edu/improvement-research-tools/ [accessed 2023-08-27]
31. Klitzman R. Pleasing doctors: when it gets in the way. BMJ 2007 Sep 8;335(7618):514. [doi: 10.1136/bmj.39325.646887.94]
32. Akincigil A, Matthews EB. National rates and patterns of depression screening in primary care: results from 2012 and 2013. Psychiatr Serv 2017 Jul 1;68(7):660-666. [doi: 10.1176/appi.ps.201600096] [Medline: 28196461]
33. Kowalski CP, McQuillan DB, Chawla N, et al. 'The Hand on the Doorknob': visit agenda setting by complex patients and their primary care physicians. J Am Board Fam Med 2018;31(1):29-37. [doi: 10.3122/jabfm.2018.01.170167] [Medline: 29330237]
34. Grant RW, Altschuler A, Uratsu CS, et al. Primary care visit preparation and communication for patients with poorly controlled diabetes: a qualitative study of patients and physicians. Prim Care Diabetes 2017 Apr;11(2):148-153. [doi: 10.1016/j.pcd.2016.11.003] [Medline: 27916628]
35. Kleiman EM, Turner BJ, Fedor S, Beale EE, Huffman JC, Nock MK. Examination of real-time fluctuations in suicidal ideation and its risk factors: results from two ecological momentary assessment studies. J Abnorm Psychol 2017 Aug;126(6):726-738. [doi: 10.1037/abn0000273] [Medline: 28481571]
36. Gorman DC, Ham SA, Staab EM, Vinci LM, Laiteerapong N. Medical assistant protocol improves disparities in depression screening rates. Am J Prev Med 2021 Nov;61(5):692-700. [doi: 10.1016/j.amepre.2021.05.010] [Medline: 34284914]
37. Reed ME, Huang J, Graetz I, et al. Patient characteristics associated with choosing a telemedicine visit vs office visit with the same primary care clinicians. JAMA Netw Open 2020 Jun 1;3(6):e205873. [doi: 10.1001/jamanetworkopen.2020.5873] [Medline: 32585018]
38. López L, Green AR, Tan-McGrory A, King R, Betancourt JR. Bridging the digital divide in health care: the role of health information technology in addressing racial and ethnic disparities. Jt Comm J Qual Patient Saf 2011 Oct;37(10):437-445. [doi: 10.1016/s1553-7250(11)37055-9] [Medline: 22013816]

Abbreviations

C-SSRS: Columbia-Suicide Severity Rating Scale
EMR: electronic medical record
IBH: Integrated Behavioral Health
OR: odds ratio
PHQ-9: Patient Health Questionnaire-9
SHC: Stanford Health Care
SW: social worker
UHA: University Healthcare Alliance

Edited by J Torous, R Wong; submitted 22.06.23; peer-reviewed by D Edge, J Lopez-Castroman; revised version received 10.02.24; accepted 27.02.24; published 01.05.24.

Please cite as:
Sattler A, Dunn J, Albarran M, Berger C, Calugar A, Carper J, Chirravuri L, Jawad N, Zein M, McGovern M
Asynchronous Versus Synchronous Screening for Depression and Suicidality in a Primary Health Care System: Quality Improvement Study
JMIR Ment Health 2024;11:e50192
URL: https://mental.jmir.org/2024/1/e50192
doi:10.2196/50192

Original Paper

Identification of Predictors of Mood Disorder Misdiagnosis and Subsequent Help-Seeking Behavior in Individuals With Depressive Symptoms: Gradient-Boosted Tree Machine Learning Approach

Jiri Benacek[1], PgDip; Nimotalai Lawal[1], MEng; Tommy Ong[1], MEng; Jakub Tomasik[1], PhD; Nayra A Martin-Key[1], PhD; Erin L Funnell[1,2], BSc; Giles Barton-Owen[2], MA; Tony Olmert[1], MD; Dan Cowell[2], BSc; Sabine Bahn[1,2], MD, PhD

[1]Department of Chemical Engineering and Biotechnology, Cambridge Centre for Neuropsychiatric Research, University of Cambridge, Cambridge, United Kingdom

[2]Psyomics Ltd, Cambridge, United Kingdom

Corresponding Author:
Sabine Bahn, MD, PhD
Department of Chemical Engineering and Biotechnology
Cambridge Centre for Neuropsychiatric Research
University of Cambridge
Philippa Fawcett Drive
Cambridge, CB3 0AS
United Kingdom
Phone: 44 1223334151
Email: sb209@cam.ac.uk

Abstract

Background: Misdiagnosis and delayed help-seeking cause significant burden for individuals with mood disorders such as major depressive disorder and bipolar disorder. Misdiagnosis can lead to inappropriate treatment, while delayed help-seeking can result in more severe symptoms, functional impairment, and poor treatment response. Such challenges are common in individuals with major depressive disorder and bipolar disorder due to the overlap of symptoms with other mental and physical health conditions, as well as, stigma and insufficient understanding of these disorders.

Objective: In this study, we aimed to identify factors that may contribute to mood disorder misdiagnosis and delayed help-seeking.

Methods: Participants with current depressive symptoms were recruited online and data were collected using an extensive digital mental health questionnaire, with the World Health Organization World Mental Health Composite International Diagnostic Interview delivered via telephone. A series of predictive gradient-boosted tree algorithms were trained and validated to identify the most important predictors of misdiagnosis and subsequent help-seeking in misdiagnosed individuals.

Results: The analysis included data from 924 symptomatic individuals for predicting misdiagnosis and from a subset of 379 misdiagnosed participants who provided follow-up information when predicting help-seeking. Models achieved good predictive power, with area under the receiver operating characteristic curve of 0.75 and 0.71 for misdiagnosis and help-seeking, respectively. The most predictive features with respect to misdiagnosis were high severity of depressed mood, instability of self-image, the involvement of a psychiatrist in diagnosing depression, higher age at depression diagnosis, and reckless spending. Regarding help-seeking behavior, the strongest predictors included shorter time elapsed since last speaking to a general practitioner about mental health, sleep problems disrupting daily tasks, taking antidepressant medication, and being diagnosed with depression at younger ages.

Conclusions: This study provides a novel, machine learning–based approach to understand the interplay of factors that may contribute to the misdiagnosis and subsequent help-seeking in patients experiencing low mood. The present findings can inform the development of targeted interventions to improve early detection and appropriate treatment of individuals with mood disorders.

(JMIR Ment Health 2024;11:e50738) doi:10.2196/50738

KEYWORDS

misdiagnosis; help-seeking; gradient-boosted trees; machine learning; depression; bipolar disorder; diagnose; diagnosis; mood; mental health; mental disorder; mental disorders; depression; depressive; predict; predictive; prediction; depressed; algorithm; algorithms

Introduction

Mood disorders are debilitating psychiatric conditions that negatively affect a person's emotional state. They result in impaired ability to function and complete daily tasks, and an increased risk of self-harm and suicide [1]. Two of the most common mood disorders are major depressive disorder (MDD) and bipolar disorder (BD), which affect approximately 3.4% and 0.5% of the global population, respectively, at any given time [2]. Beyond the impact on the affected individuals, there are also economic and social consequences such as lost productivity, increased health care costs, and costs incurred by unpaid carers. In the United Kingdom alone, the economic burden of managing MDD and BD is estimated at £7.5 billion (US $9.55 billion) and £5.2 billion (US $6.62 billion), respectively [3], with a significant portion of this burden attributed to underdiagnosis and high rates of misdiagnosis of mood disorders.

Although misdiagnosis is prevalent in all areas of medicine, the heterogeneous nature of mental illness and lack of objective diagnosis make it more common for mental health conditions [4]. The diagnosis of mental health disorders is currently based on assessing patient symptom profiles using diagnostic manuals such as the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5) [5] or the International Statistical Classification of Diseases and Related Health Problems, 11th Revision (ICD-11) [6]. As such, diagnosis relies heavily on symptom reporting and patients who do not recognize and thus do not report their symptoms or present with complex symptoms are more likely to be misdiagnosed [7]. For example, issues with symptom reporting are considered a major cause of BD misdiagnosis [8], with many patients with BD only seeking medical help during depressive episodes [9], which makes mania more difficult to identify. Consequently, as many as 78% of mood disorder diagnoses are missed in primary care [10], including approximately 40% of patients with BD who are initially misdiagnosed with MDD [11]. This, in turn, leads to incorrect treatment of BD with antidepressants which have lower efficacy than mood stabilizers in alleviating bipolar symptoms and have been associated with prolonged episodes of mania and accelerated cycling between manic and depressive states [12,13]. Understanding factors that lead to misdiagnosis could guide the development of more effective means for early identification and intervention in individuals at high risk.

An additional barrier to receiving a correct diagnosis and necessary care is the reluctance of affected individuals to speak to medical professionals about their mental health. The European Study of the Epidemiology of Mental Disorders carried out across 6 countries found that only 25.4% of respondents spoke to a medical professional about their mental health problems [14]. Likewise, active engagement with mental health services is consistently low, with almost 75% of patients experiencing a mental illness in England receiving no treatment [15]. One of the reasons for the low rates of help-seeking in individuals experiencing mental health symptoms is concerns of potential public and self-internalized stigma. Consequently, individuals struggling with their mental health often turn to coping mechanisms such as social withdrawal, secrecy, and label avoidance [16,17] rather than seeking help [18]. Therefore, it is imperative to recognize barriers to help-seeking in mental health to facilitate early and accurate diagnosis in un and misdiagnosed individuals.

Although previous studies have investigated factors contributing to the misdiagnosis, poor help-seeking behavior, and barriers to receiving a diagnosis, only a few have used machine learning methods to do so [19]. The use of machine learning in mental health research has increased in recent years, with many studies focusing on detection and diagnosis, treatment and support, public health, and research and clinical administration [19]. While not without limitations, the use of machine learning can offer data-driven insights into complex relationships between high-dimensional data [20,21]. Although other, mostly qualitative investigations have identified the predictors of help-seeking and misdiagnosis by considering factors individually, this study aims to take a more holistic approach. By developing machine learning models based on extensive self-reported patient data, we aim to identify and quantify interdependent predictive factors for the misdiagnosis of mental health disorders, specifically mood disorders, and help-seeking behavior in individuals who may have been misdiagnosed. Identifying such predictive factors could aid in avoiding preventable misdiagnosis, encourage help-seeking, and improve outcomes in patients presenting with depressive symptoms.

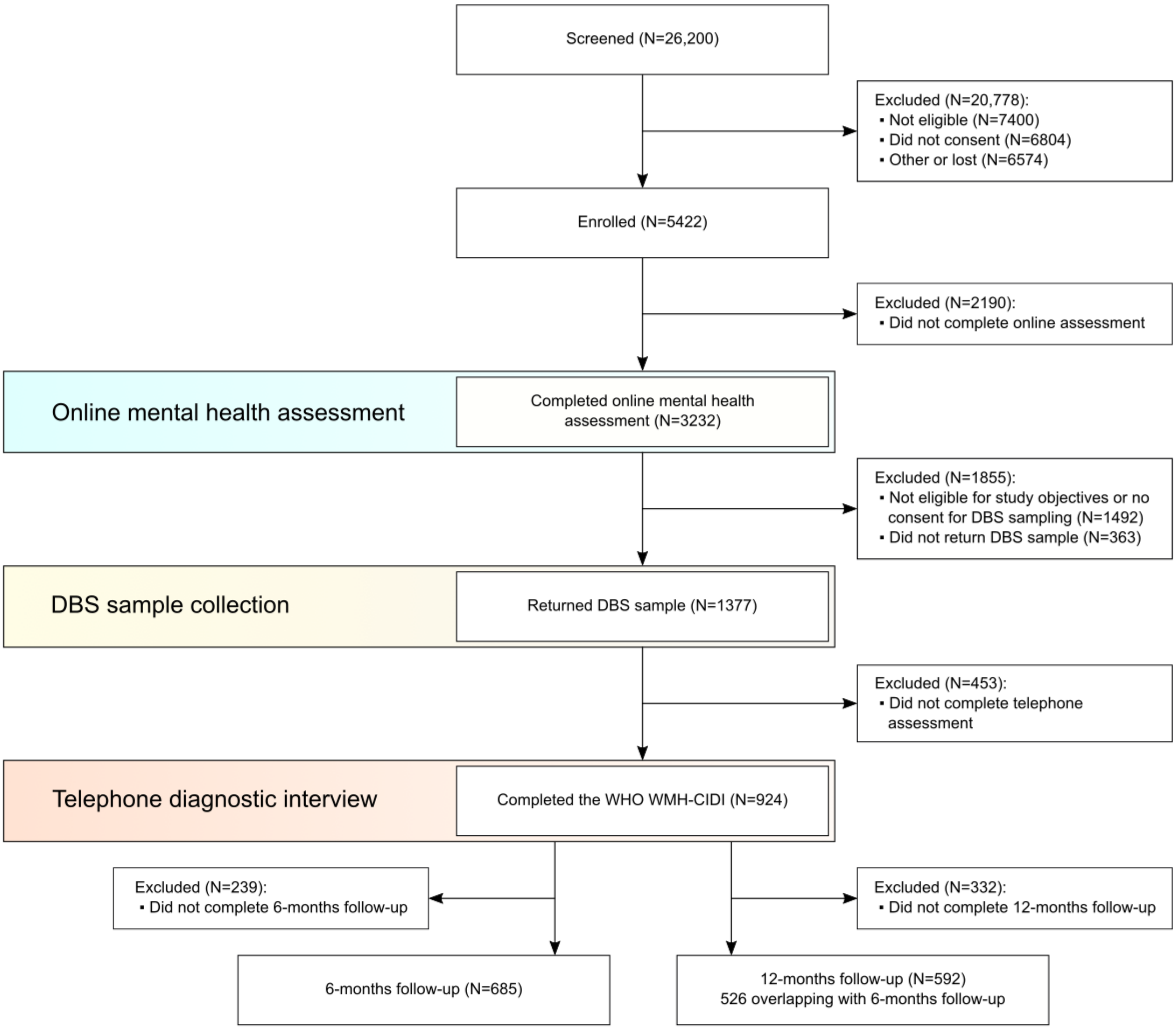
Methods

Data Acquisition

Overview

The data used in this report were collected as part of the Delta Study—a study aiming to facilitate a more accurate and earlier diagnosis of BD and MDD; carried out in the United Kingdom by the Cambridge Centre for Neuropsychiatric Research between 2018 and 2020 [22,23]. The study consisted of an adaptive digital questionnaire, the Composite International Diagnostic Interview (CIDI) [24], and 2 follow-up questionnaires at 6 and 12 months. The stages of the Delta Study are summarized in Figure 1. Participants were recruited nonrandomly through email, the Cambridge Centre for Neuropsychiatric Research (CCNR) website, and paid Facebook advertisements. The eligibility criteria included at least mild depressive symptoms, indicated by a score of ≥5 on the Patient Health Questionnaire-9 (PHQ-9) [25] at the time of recruitment, aged between 18 and 45 years, and residency in the United Kingdom. Participants who indicated current suicidal ideation or intent, were pregnant, or breastfeeding, were excluded.

Figure 1. Delta Study flow diagram [22]. DBS: dried blood spot; WHO: World Health Organization; WMH-CIDI: World Mental Health Composite International Diagnostic Interview.



Adaptive Digital Questionnaire

In total, 3232 participants completed the adaptive digital questionnaire available on the Delta Study digital platform. The questionnaire consisted of 635 questions, divided into six sections: (1) demographic information and personal history; (2) manic and hypomanic symptoms; (3) depressive symptoms; (4) personality profiling; (5) treatment, medication, substance use, and family psychiatric history; and (6) other psychiatric conditions. As the questionnaire was adaptive to answers given by participants, the maximum number of questions an individual could answer was 382, with an average of 284. Within the questionnaire, participants reported their baseline diagnosis, and their current well-being (within the previous 14 days) was quantified using the Warwick-Edinburgh Mental Well-Being Scale (WEMWBS) [26].

Composite International Diagnostic Interview

Participants who completed the web-based mental health questionnaire were invited to complete the CIDI version 3.0 via telephone. The CIDI is a structured diagnostic interview for mental disorders created by the World Health Organization based on the International Classification of Diseases and Related Health Problems, 10th Revision (ICD-10). It was developed primarily for epidemiological studies and has been extensively validated, demonstrating high diagnostic reliability [27]. In this study, only sections pertaining to mood disorder diagnoses were applied, that is, the demographics, depression, and mania modules. Interviewers were trained by CIDI-certified instructors prior to conducting the interviews. In total, 924 participants completed the CIDI and received one of the following diagnoses in their results report: BDI, BDII, subthreshold BD, MDD with subthreshold BD, MDD, or no mood disorder diagnosis (referred to as “low mood”).

Follow-Up Questionnaires

Participants who completed the digital questionnaire were invited to fill out 2 follow-up questionnaires, 6 and 12 months after receiving their results report. The follow-up questionnaires aimed to determine the effects of participation in the Delta Study on participants’ quality of life and record subsequent changes in diagnosis and treatment. A total of 2064 participants completed at least 1 of the follow-up questionnaires, with 1780 respondents at 6 months and 1542 respondents at 12 months.

Outcomes

Overview

For the purposes of this study, 2 dependent variables were defined.

Misdiagnosis

For participants who completed the CIDI, the mood disorder diagnosis reported at baseline was compared to the diagnosis obtained from the CIDI, including patients with no mood disorder diagnosis at baseline who should have been diagnosed. CIDI diagnosis was used as the gold standard, and any mismatch with the baseline diagnosis was defined as misdiagnosis. This definition of misdiagnosis was consistent with previous studies investigating under- and misdiagnosis of mood disorders based on comparing patient-reported diagnoses to the outcomes of structured clinical interviews [28,29].

Help-Seeking Behavior

In the 6- and 12-month follow-up questionnaires, participants were asked: "Have you had an appointment with a GP or psychiatrist to talk about your mental health in the past 6 months?" A positive response to this question at either time point was defined as help-seeking. In order to examine help-seeking in misdiagnosed individuals, only those who were identified as misdiagnosed within outcome 1 were included in the analysis.

Analysis

Overview

Raw data processing and feature engineering were performed in R (version 3.6.3; R Core Team) [30]. Subsequent analyses and modeling were carried out using Python (version 3.9.7; Python Software Foundation) [31]. Main libraries used included Pandas (version 1.5.2; Pandas Development Team) [32] and NumPy version 1.23.5 [33] for data manipulation; scikit-learn version 1.0.2 [34], XGBoost (version 1.6.1; The XGBoost Contributors) [35], and SHAP version 0.41.0 [36] for modeling and interpretation; and Seaborn version 0.12.1 [37] and Matplotlib version 3.6.2 [38] for plotting.

Data Preparation

Prior to analysis, constant and duplicate variables were removed. Answers to questions examining the same symptom or construct were concatenated, and new features were created to represent these aggregated answers. Missing data were imputed where possible (for example, the answer to the question asking "Has anyone suggested you drink less?" was set to 0 for participants who had indicated they do not drink), and otherwise remained nonrandomly missing. Categorical variables were 1-hot encoded, that is, unique dummy variables were created where the presence of each category was denoted by "1," and its absence was represented by "0."

Modeling and Interpretation

This analysis aimed to develop predictive models to identify variables influencing (1) misdiagnosis and (2) help-seeking behavior in participants who were identified as potentially misdiagnosed. A decision tree–based machine learning algorithm Extreme Gradient Boosting (XGBoost) [35] was chosen to train the classification models due to being robust to outliers, agnostic to data distribution, having the ability to handle nonrandom missing data, offering good predictive power, and due to it allowing for good model interpretability. Repeated nested cross-validation (rNCV) was used for model training and evaluation to obtain accurate estimates of model performance in unseen data. rNCV relies on performing a k-fold cross-validation (CV) within each round of another CV. This allows for model-specific hyperparameter optimization in the inner loop, with the final model being trained using the best-performing set of parameters, and later evaluated in the outer loop of rNCV. For this analysis, a 4-fold stratified CV was used in both the inner and outer loops, where 3 of the folds acted as a training set and 1 as a test set. Tuned model parameters included the number of estimators (1 to 100), shrinkage rate (0.1 to 0.3) to prevent overfitting, and tree depth (1 or 2) to allow for first-order interactions between predictors. The training was repeated 100 times, generating a total of 400 models for each of the objectives. Generalized model performance was evaluated by calculating the area under the receiver operating characteristic curve (AUC). The classification cutoff was optimized for the Youden index [39] to balance the true positive and true negative rates and offset potential imbalances between classes. SHAP (Shapley additive explanations) analysis [36], which combines local interpretable model–agnostic explanations (LIME) [40] and Shapley sampling values [41] approaches, was used for model interpretation. Feature occurrence frequency was calculated as the percentage of the models that incorporated a given feature to generate predictions. Reported results represent mean and SD values across the rNCV models.

Ethical Considerations

The study protocol was approved by the University of Cambridge Human Biology Research Ethics Committee (HBREC 2017.11) and all enrolled participants signed a digital informed consent form.

Results

Misdiagnosis

The self-reported baseline diagnosis did not match the diagnosis assigned by CIDI for 471 (50.97%) of the 924 participants who completed the CIDI interview. These participants were therefore considered misdiagnosed. No between-group differences were observed in terms of age, sex, ethnicity, highest achieved education level, or relationship status between the correctly diagnosed and misdiagnosed groups (Table S1 in Multimedia Appendix 1). However, there were significant differences in employment status as well as well-being and PHQ-9 scores, with misdiagnosed individuals, on average, reporting lower well-being and more severe depressive symptoms.

On average, the models correctly classified 70% (SD 9%) of misdiagnosed participants and 71% (SD 9%) of correctly diagnosed participants, with a mean accuracy of 70% (SD 3%) and the out-of-fold AUC of 0.75 (SD 0.03; Figure 2 and Table

S2 in Multimedia Appendix 1). Among the 1045 variables evaluated, the strongest predictors of misdiagnosis were more severe composite depressive symptoms and unstable self-image (Figure 3). Unstable self-image was measured by a 4-level Likert scale question "Is your image and sense of yourself and what you believe in unstable and constantly changing?" The next strongest predictor was the diagnosing clinician, with those who were undiagnosed at baseline or reported a diagnosis by a psychiatrist more likely to be misdiagnosed. The top 10 predictors also included variables related to age at diagnosis of BD and MDD, with late (≥35 years of age) diagnosis or no diagnosis at all, increasing the likelihood of being misdiagnosed (Figure S1 in Multimedia Appendix 1). Misdiagnosed participants were also more likely to recklessly spend money, experienced more frequent intense mood swings or mania in general, had higher weight gain during low mood episodes, and were more sexually active than usual at the time of data collection.

Figure 2. Out-of-fold model performance in predicting misdiagnosis. Green lines represent predictive performance on unseen out-of-fold data for each of the 400 final models. The thick blue line represents the average of all ROC curves. The grey area represents 1 SD. AUC: area under the receiver operating characteristic curve; ROC: receiver operating characteristic.

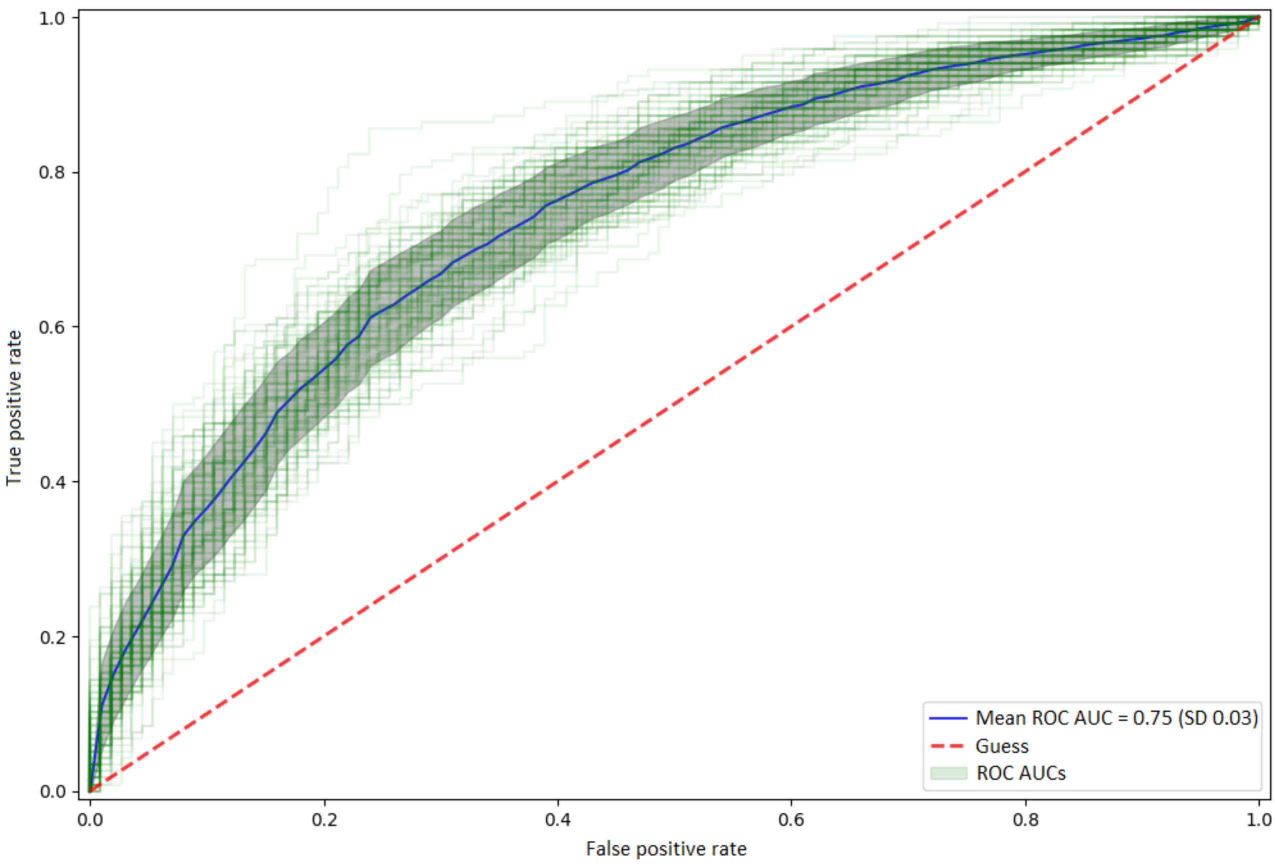
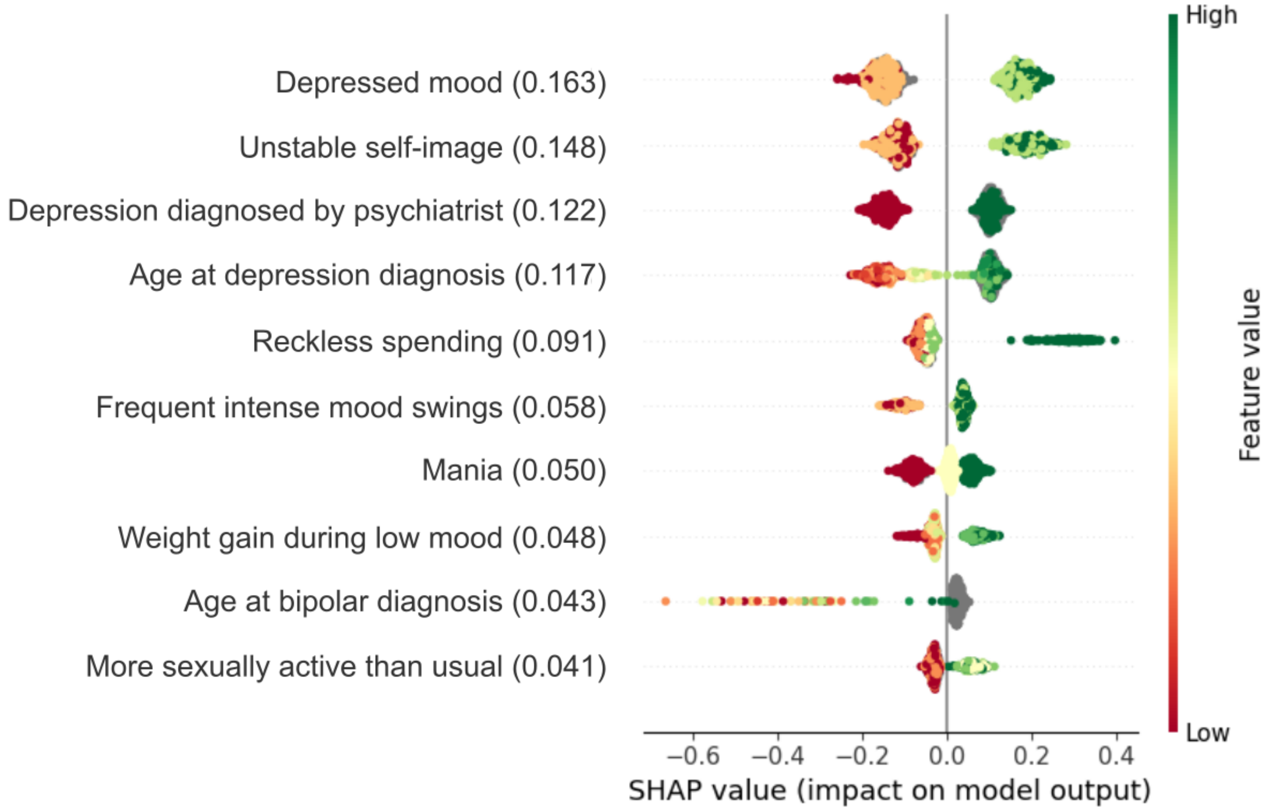


Figure 3. Results for the top 10 variables in the misdiagnosis model. Shown is SHAP analysis of the factors associated with misdiagnosis. The features (y-axis) are ordered by their average feature importance, indicated by the value inside the brackets, across all models. Each colored dot represents a participant, where the color gradient shows the value of the answer (red if low, green if high, and grey if missing), and the corresponding value on the x-axis shows directionality and the impact on model output, as determined using SHAP analysis. Values below 0 show directionality toward being correctly diagnosed, whereas values above 0 show directionality toward misdiagnosis. SHAP: Shapley additive explanations.



Model performance was largely driven by the top 5 predictors, with a steady decline in SHAP scores for subsequent variables. Of the top 10 predictors, 9 were selected in more than 75% (n=300) of the models, suggesting a relatively stable model composition. The exception was a variable related to "being more sexually active than usual," which was selected in 71% (n=284) of the models. More detailed information on feature selection frequency is provided in Figure S3 in Multimedia Appendix 1.

Help-Seeking Behavior

Help-seeking behavior was investigated in 379 participants who were misdiagnosed at the baseline and who had completed at least 1 of the follow-up questionnaires. Of those, 229 (60.42%) participants sought an appointment with a medical professional during the follow-up period to discuss their mental health and were therefore defined as "help-seekers." The help-seeker and non–help-seeker groups differed significantly in the highest achieved education level, relationship status, well-being, and severity of depressive symptoms (Table S3 in Multimedia Appendix 1). Participants more likely to seek help were on average less formally educated, more likely single, reported higher mean severity of symptoms, and worse overall well-being.

The model achieved an AUC of 0.71 (SD 0.04; Figure 4), with a sensitivity of 65% (SD 13%), specificity of 72% (SD 13%), and average accuracy of 67% (SD 4%; Table S4 in Multimedia Appendix 1). The strongest predictor was the shorter time since patients last spoke to a general practitioner (GP) about their mental health at baseline (Figure 5). It was followed by sleep problems disrupting daily tasks and taking prescribed antidepressants, both associated with increased help-seeking. Consistent with this, lower help-seeking was observed in participants who had never been prescribed antidepressants, namely selective serotonin reuptake inhibitors (SSRIs). Furthermore, there was a lower likelihood of help-seeking with higher age at both the first episode of low mood and diagnosis of depression, which was similarly predictive to not having been previously diagnosed with depression. Finally, impaired ability to work, lower well-being scores, feeling worthless, and lower self-rated mental health were associated with help-seeking behavior.

Figure 4. Out-of-fold model performance in predicting help-seeking. Green lines represent predictive performance on unseen out-of-fold data of each of the 400 final models. The thick blue line represents an average of all ROC curves. The grey area represents 1 SD. AUC: area under the receiver operating characteristic curve; ROC: receiver operating characteristic.

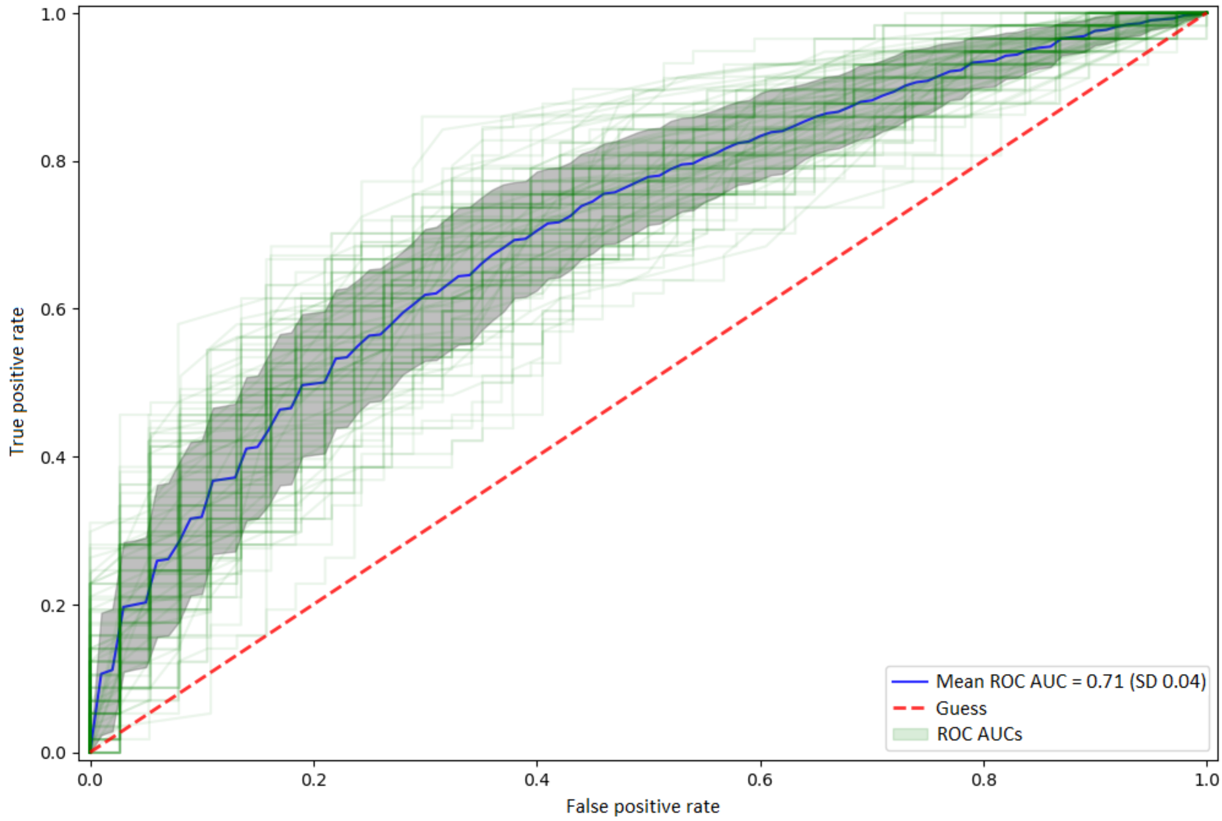
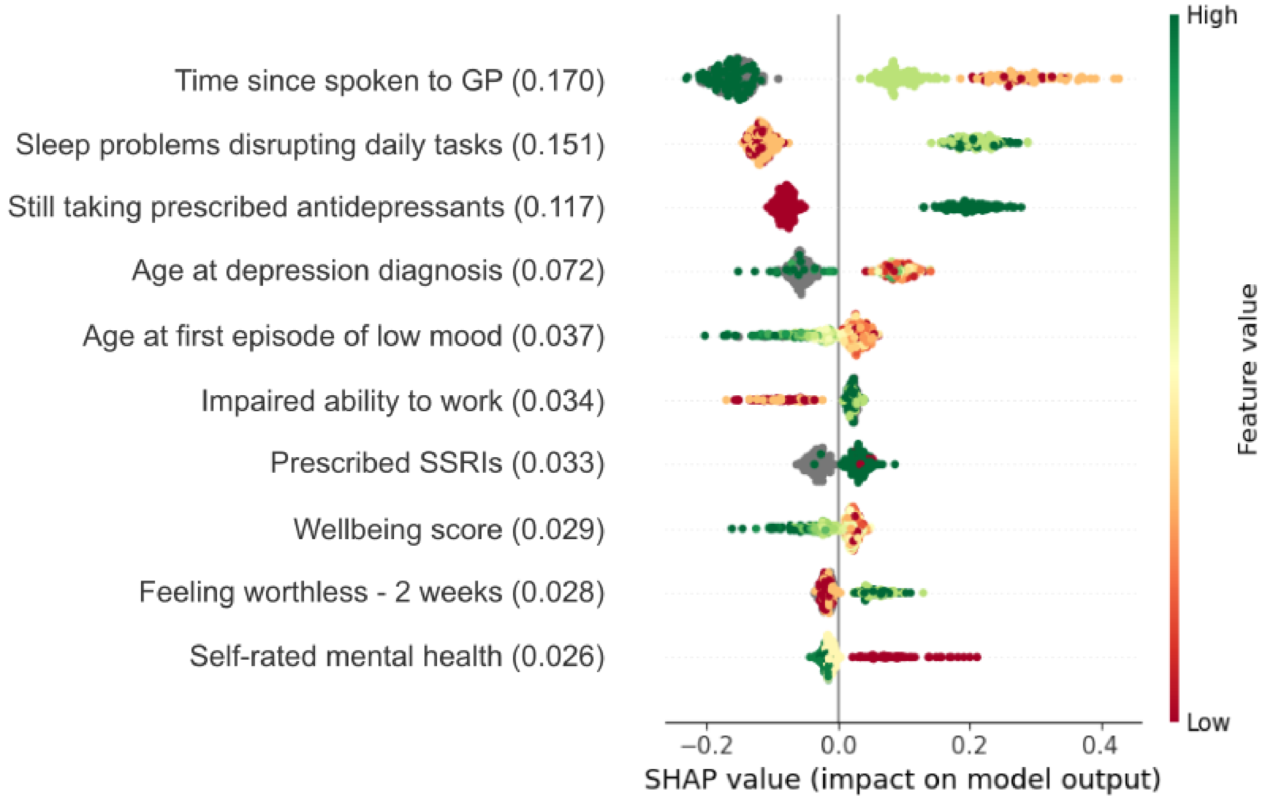


Figure 5. Results for top 10 variables in the help-seeking model in misdiagnosed individuals. Shown is feature SHAP importance (in brackets) and feature SHAP values (data points). SHAP values below 0 show directionality toward low help-seeking (ie, no appointment with GP or psychiatrist to discuss mental health), whereas values above 0 show directionality toward high help-seeking. GP: general practitioner; SHAP: Shapley additive explanations; SSRI: selective serotonin reuptake inhibitor.



The 3 variables, namely, time since last spoken to a GP, sleep problems disrupting daily tasks, and still taking prescribed antidepressants, were selected in nearly all models (Figure S4 in Multimedia Appendix 1), suggesting their high relevance for model predictions. Among other predictors, only age when diagnosed with depression was selected in more than 75% (n=300) of models, with the remaining features only selected in approximately 50% (n=200) of models, indicating their lower relevance.

Discussion

Principal Findings

This study aimed to develop machine learning models to explore factors potentially contributing to misdiagnosis and subsequent help-seeking in individuals experiencing low mood. For this purpose, we used data obtained through an extensive digital questionnaire concerning demographic, personality, and mental health data, as well as, the validated and standardized diagnostic interview, CIDI. Developed models achieved a fair level of predictive power, with AUCs of 0.75 and 0.71 for predicting misdiagnosis and help-seeking, respectively. Below, we discuss the main findings as well as the strengths and limitations of this analysis.

Misdiagnosis

The strongest predictor of misdiagnosis was the severity of depressed mood, with more severe depressive symptoms being associated with a greater risk of being misdiagnosed. This directionality was consistent with other top predictors of misdiagnosis, including unstable self-image, reckless spending, frequent intense mood swings, mania, weight gain during low mood, and being more sexually active than usual. Except for the instability of self-image, these predictors can be divided into either depression or mania or bipolar-related symptoms. Overall, the finding that individuals with more severe mental health symptoms are at a greater risk of being misdiagnosed is surprising, given the opposite could be expected as milder symptoms are harder to detect [42]. Several factors could contribute to this association, including the complexity of diagnosing mental health disorders [43], variability in symptom presentation [44,45], and the high degree of symptom overlap across different diagnoses [5]. A possible explanation for the increased risk of misdiagnosis among individuals with more severe symptoms is that they may present with prominent mood instability, such as that observed in patients with personality disorder, or rapidly cycling symptoms, making accurate diagnosis more challenging [9]. In addition, individuals with more severe symptoms often lack motivation to seek help, hence their symptoms may remain unrecognized for a longer time [46].

In the case of mood disorders, misdiagnosis of individuals with higher depressive symptom severity may result from the fact that patients with BD generally seek medical help during depressive episodes and often present with more severe depressive symptoms than patients with MDD, while underreporting manic phases [47,48]. In fact, less than a third of patients with BD report the presence of reckless behavior, excessive spending, and increased sexual interest or activity [49]. This contributes to approximately 40% of patients with BD receiving an incorrect initial diagnosis of unipolar depression [50]. Also, the association of frequent intense mood swings with mood disorder misdiagnoses may be related to incorrect treatment of depressive symptoms of BD with antidepressants, rather than mood stabilizer medication, which has the potential to induce mania and rapid cycling [51,52].

The second most predictive feature of misdiagnosis identified in this study was unstable self-image. Previous literature has shown that an unstable sense of self is associated with frequent changes in diagnosis, and often linked to complex and unstable personality characteristics [53]. The high ranking of self-image stability could, however, be a result of the high comorbidity rates between BD and other disorders featuring unstable self-image that were not evaluated by the diagnostic interview used in this study, such as borderline personality disorder [54]. This is especially important considering that such disorders may share a high number of similarities with BD, leading to frequent misdiagnoses [55,56]. The 2 additional symptoms that are ranked high in terms of predictive value for misdiagnosis in this analysis regard reckless spending and increased sexual activity, representing reckless or impulsive behavior, which are included in the diagnostic criteria of both BD and borderline personality disorder [17].

Finally, among the top predictors of misdiagnosis were 3 variables related to psychiatric history, including psychiatrist involvement in the diagnosis, age at depression diagnosis, and age at BD diagnosis. Interestingly, the models attributed a higher risk of misdiagnosis to individuals whose depression was diagnosed by a psychiatrist. This may be caused by the fact that patients at high risk of misdiagnosis, such as those with more complex symptom presentation or suspected comorbidities, are usually referred to secondary care, following the National Institute for Health and Care Excellence (NICE) guidelines [57]. However, this finding should be interpreted with caution, as diagnoses made by psychiatrists are generally more accurate than those derived from the CIDI. Also, participants who received a diagnosis of a mood disorder at an older age, or not at all, were more likely to be misdiagnosed. This finding is surprising, as previous literature suggests that the severity and impact of symptoms decline with age, with 86% of patients with BD diagnosed by the age of 25 [58]. However, it is possible that due to milder symptoms, patients who are older may remain undiagnosed for longer periods of time.

Help-Seeking

Analysis of participants with a mismatch between their self-reported formal diagnosis and the CIDI outcome revealed several predictors of help-seeking related to patients' mental health history and symptoms.

The most predictive feature was time since last spoken to a GP at baseline, with patients who had visited their GP more recently being more likely to seek help. Interestingly, that was not the case for the time since last spoken to a psychiatrist, likely due to most participants not being under secondary care and the long waiting times for psychiatric assessment [59]. In line with previous literature [60], these findings indicate that help-seeking was also associated with more severe psychiatric symptoms and having a previous diagnosis of mood disorder. Similarly, participants seeking help reported lower well-being, feeling more worthless, and more functional impairment in carrying out daily tasks and at work caused by symptoms and sleep problems.

Interestingly, while there was not a significant overall age difference between the help-seekers and non–help-seekers, further analyses showed a lower tendency to seek help in individuals who were over 35 years old at initial diagnosis of depression (Figure S2 in Multimedia Appendix 1). The pattern of people who are younger being more likely to seek help is in line with the published literature [61]. Together with the finding that the initial diagnosis at older age was a strong predictor of misdiagnosis [62], this result indicates that patients who are most likely to be misdiagnosed are also the least likely to seek help. Thus, older patients may require more support to tackle potential barriers to help-seeking and receiving a diagnosis, such as stigma and inadequate mental health education [63].

The final set of predictors of help-seeking was related to previous medication. Higher help-seeking was observed in misdiagnosed individuals who were still taking previously prescribed antidepressants, in particular SSRIs, as opposed to misdiagnosed individuals who either had never been prescribed SSRIs or other antidepressant medication or had stopped taking it. The association of antidepressant treatment with help-seeking indicates that the prescribed medication may have been ineffective, as is often the case when attempting to treat depressive episodes of BD with antidepressant monotherapy [64]. Compared with the patients with MDD, the patients with BD respond worse to antidepressant medication, with short-term nonresponse rates of 51.3% in BD versus 31.6% in MDD [65]. This difference is even more pronounced in the long-term, where the loss of response to antidepressants is 3.4 times more frequent in patients with BD, while withdrawal relapse into depression is 4.7 times less frequent in BD compared to patients with MDD [65]. Moreover, individuals with unrecognized BD who are treated with antidepressants sometimes develop symptoms of mania, which in turn may motivate patients or their relatives to seek consultation with a specialist [66].

Limitations

The main limitation of this study is the reliance on CIDI as the gold standard for mood disorder diagnosis. Although the CIDI demonstrates good agreement with structured diagnostic interviews conducted by clinicians [67], future studies should consider either retrospective or longitudinal study designs, and ideally access medical records for more accurate diagnoses, including those beyond mood disorders. Additionally, the study participants were recruited online following strict inclusion criteria and were predominantly White, necessitating further research in traditionally underrepresented ethnic minorities and more representative patient cohorts. Another limitation is the exclusion of individuals with current suicidal ideation, a characteristic that could be an important indicator of misdiagnosis. Finally, the observed associations do not necessarily imply causality, which can only be evaluated through prospective causal inference study designs.

Conclusions

This analysis leveraged comprehensive patient data, a robust machine learning algorithm, and an extensive validation framework, to identify predictors of mood disorder misdiagnosis in individuals experiencing depressive symptoms, and subsequent help-seeking. The results highlight the increased risk for misdiagnosis associated with incomplete symptom profiles, more severe or harder to detect symptoms, and older age. Therefore, comprehensive symptom monitoring outside of depressive episodes, mental health screening at earlier ages, and clinician knowledge of the influence of advanced age on misdiagnosis risk are important considerations for early and accurate diagnosis of mood disorders. Moreover, prior engagement with mental health services, functional impairment in performing daily tasks, and younger age were associated with a higher likelihood of help-seeking. Together, these results add to the growing application of machine learning techniques in examining existing barriers to accessing mental health services [19], and may ultimately lead to the development of novel screening tools or procedures for a comprehensive mental health risk assessment in individuals presenting with mood-related symptoms.

Acknowledgments

We are most grateful to all participants of the Delta Study for their time and efforts. We are also grateful to all members of the Delta Study Service User Advisory Group. We also thank all those involved in designing and conducting the Delta Study for their valuable input including Jason D Cooper, Sung Yeon Sarah Han, Lynn P Farrag, Emily Bell, Lauren V Friend, Sharmelee Thiahulan, Pawel Eljasz, Mark Agius, Neil Hunt, and the CIDI interviewers. This study was funded by the Stanley Medical Research Institute (07R-1888) and Psyomics Ltd. Stanley Medical Research Institute was not involved in any part of the research. Psyomics Ltd was involved in the design and conduct of the study; recruitment, data collection, and management; and review or approval of the study.

Authors' Contributions

SB and DC conceived the Delta Study, conceptualized, and supervised the development of the web-based mental health questionnaire. SB, DC, GBO, and T Olmert contributed to the design of the study. GBO and T Olmert collected the web-based mental health questionnaire data. GBO and JT processed the web-based mental health questionnaire data. JB, NL, and T Ong analyzed the data. SB, JT, NAMK, and ELF advised the analysis. JB and JT wrote the first draft of the study, with contributions from NL, T Ong, NAMK, ELF, and SB. All authors contributed to the final version of the study.

Conflicts of Interest

SB is a director of Psynova Neurotech Ltd and Psyomics Ltd. SB, ELF, and DC have financial interests in Psyomics Ltd. GBO had financial interests in Psyomics Ltd. SB, JT, and T Olmert have received payments from the University of Cambridge for licensing of data from the Delta Study. SB and JT may benefit financially from patents arising from the Delta Study. ELF is a consultant for Psyomics Ltd. All other authors declare no competing interests.

Multimedia Appendix 1

Demographics, model performance metrics, dependence plots, and feature selection frequencies for all objectives.
[DOCX File , 686 KB - mental_v11i1e50738_app1.docx]

References

1. Isometsä E. Suicidal behaviour in mood disorders—who, when, and why? Can J Psychiatry 2014;59(3):120-130 [FREE Full text] [doi: 10.1177/070674371405900303]
2. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
3. McCrone P, Dhanasiri S, Patel A, Knapp M, Lawton-Smith S. Paying the Price: The Cost of Mental Health Care in England to 2026. London: The King's Fund; 2008.
4. Vermani M, Marcus M, Katzman MA. Rates of detection of mood and anxiety disorders in primary care: a descriptive, cross-sectional study. Prim Care Companion CNS Disord 2011;13(2):1-10 [FREE Full text] [doi: 10.4088/PCC.10m01013] [Medline: 21977354]
5. American Psychiatric Association. DSM-5 Task Force. Diagnostic and Statistical Manual of Mental Disorders: DSM-5, 5th Edition. Washington, DC: American Psychiatric Association; 2013.
6. ICD-11: International Classification of Diseases 11th Revision: The Global Standard for Diagnostic Health Information. Geneva: World Health Organisation; 2018.
7. Ayano G, Demelash S, Yohannes Z, Haile K, Tulu M, Assefa D, et al. Misdiagnosis, detection rate, and associated factors of severe psychiatric disorders in specialized psychiatry centers in Ethiopia. Ann Gen Psychiatry 2021;20(1):10 [FREE Full text] [doi: 10.1186/s12991-021-00333-7] [Medline: 33531016]
8. Shen H, Zhang L, Xu C, Zhu J, Chen M, Fang Y. Analysis of misdiagnosis of bipolar disorder in an outpatient setting. Shanghai Arch Psychiatry 2018;30(2):93-101 [FREE Full text] [doi: 10.11919/j.issn.1002-0829.217080] [Medline: 29736129]
9. Young AH, Grunze H. Physical health of patients with bipolar disorder. Acta Psychiatr Scand 2013;127(Suppl 442):3-10 [FREE Full text] [doi: 10.1111/acps.12117] [Medline: 23581787]
10. Frye MA, Calabrese JR, Reed ML, Wagner KD, Lewis L, McNulty J, et al. Use of health care services among persons who screen positive for bipolar disorder. Psychiatr Serv 2005;56(12):1529-1533 [FREE Full text] [doi: 10.1176/appi.ps.56.12.1529] [Medline: 16339614]
11. Singh T, Rajput M. Misdiagnosis of bipolar disorder. Psychiatry (Edgmont) 2006;3(10):57-63 [FREE Full text] [Medline: 20877548]
12. Hirschfeld RM. Differential diagnosis of bipolar disorder and major depressive disorder. J Affect Disord 2014;169(Suppl 1):S12-S16 [FREE Full text] [doi: 10.1016/S0165-0327(14)70004-7] [Medline: 25533909]
13. Bowden CL. A different depression: clinical distinctions between bipolar and unipolar depression. J Affect Disord 2005;84(2-3):117-125. [doi: 10.1016/S0165-0327(03)00194-0] [Medline: 15708408]
14. Alonso J, Lépine JP, ESEMeD/MHEDEA 2000 Scientific Committee. Overview of key data from the European Study of the Epidemiology of Mental Disorders (ESEMeD). J Clin Psychiatry 2007;68(Suppl 2):3-9. [Medline: 17288501]
15. Davies SC. Annual report of the chief medical officer 2013, public mental health priorities: investing in the evidence. Department of Health. 2014. URL: https://assets.publishing.service.gov.uk/media/5a8034e0e5274a2e87db87a0/CMO_web_doc.pdf [accessed 2023-12-19]
16. Schomerus G, Angermeyer MC. Stigma and its impact on help-seeking for mental disorders: what do we know? Epidemiol Psichiatr Soc 2008;17(1):31-37. [doi: 10.1017/s1121189x00002669] [Medline: 18444456]
17. Ben-Zeev D, Young MA, Corrigan PW. DSM-V and the stigma of mental illness. J Ment Health 2010;19(4):318-327. [doi: 10.3109/09638237.2010.492484] [Medline: 20636112]
18. Pattyn E, Verhaeghe M, Sercu C, Bracke P. Public stigma and self-stigma: differential association with attitudes toward formal and informal help seeking. Psychiatr Serv 2014;65(2):232-238 [FREE Full text] [doi: 10.1176/appi.ps.201200561] [Medline: 24233070]
19. Shatte ABR, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. Psychol Med 2019;49(9):1426-1448. [doi: 10.1017/S0033291719000151] [Medline: 30744717]
20. Rahman RA, Omar K, Noah SAM, Danuri MSNM, Al-Garadi MA. Application of machine learning methods in mental health detection: a systematic review. IEEE Access 2020;8:183952-183964 [FREE Full text] [doi: 10.1109/access.2020.3029154]

21. Tiffin PA, Paton LW. Rise of the machines? Machine learning approaches and mental health: opportunities and challenges. Br J Psychiatry 2018;213(3):509-510 [FREE Full text] [doi: 10.1192/bjp.2018.105] [Medline: 30113285]
22. Olmert T, Cooper JD, Han SYS, Barton-Owen G, Farrag L, Bell E, et al. A combined digital and biomarker diagnostic aid for mood disorders (the delta trial): protocol for an observational study. JMIR Res Protoc 2020;9(8):e18453 [FREE Full text] [doi: 10.2196/18453] [Medline: 32773373]
23. Tomasik J, Han SYS, Barton-Owen G, Mirea DM, Martin-Key NA, Rustogi N, et al. A machine learning algorithm to differentiate bipolar disorder from major depressive disorder using an online mental health questionnaire and blood biomarker data. Transl Psychiatry 2021;11(1):41 [FREE Full text] [doi: 10.1038/s41398-020-01181-x] [Medline: 33436544]
24. World Health Organization. The World Health Organization composite international diagnostic interview. In: Kessler RC, Ustun TB, editors. The WHO World Mental Health Surveys: Global Perspectives on the Epidemiology of Mental Disorders. Cambridge: Cambridge University Press; 2008:58-90.
25. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatric Annals 2002;32(9):509-515. [doi: 10.3928/0048-5713-20020901-06]
26. Tennant R, Hiller L, Fishwick R, Platt S, Joseph S, Weich S, et al. The Warwick-Edinburgh Mental Well-Being Scale (WEMWBS): development and UK validation. Health Qual Life Outcomes 2007;5:63 [FREE Full text] [doi: 10.1186/1477-7525-5-63] [Medline: 18042300]
27. Haro JM, Arbabzadeh-Bouchez S, Brugha TS, de Girolamo G, Guyer ME, Jin R, et al. Concordance of the Composite International Diagnostic Interview Version 3.0 (CIDI 3.0) with standardized clinical assessments in the WHO World Mental Health surveys. Int J Methods Psychiatr Res 2006;15(4):167-180 [FREE Full text] [doi: 10.1002/mpr.196] [Medline: 17266013]
28. Zimmerman M, Ruggero CJ, Chelminski I, Young D. Is bipolar disorder overdiagnosed? J Clin Psychiatry 2008;69(6):935-940. [doi: 10.4088/jcp.v69n0608] [Medline: 18466044]
29. McIntyre RS, Patel MD, Masand PS, Harrington A, Gillard P, McElroy SL, et al. The Rapid Mood Screener (RMS): a novel and pragmatic screener for bipolar I disorder. Curr Med Res Opin 2021;37(1):135-144 [FREE Full text] [doi: 10.1080/03007995.2020.1860358] [Medline: 33300813]
30. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2013.
31. Van Rossum G, Drake FL. Python Reference Manual. Amsterdam: Centrum voor Wiskunde en Informatica; 1995.
32. Wes M. Data structures for statistical computing in python. 2010 Presented at: Proceedings of the 9th Python in Science Conference (SciPy 2010); June 28-July 3, 2010; Austin, Texas p. 56-61. [doi: 10.25080/majora-92bf1922-012]
33. Oliphant TE. A Guide to NumPy. USA: Trelgol Publishing; 2006.
34. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2014;12:2825-2830 [FREE Full text]
35. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016 Presented at: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, California, USA p. 785-794. [doi: 10.1145/2939672.2939785]
36. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. 2017 Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, California, USA p. 4765-4774.
37. Waskom M, seaborn development team. mwaskom/seaborn. GitHub. 2020. URL: https://github.com/mwaskom/seaborn [accessed 2023-12-19]
38. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng 2007;9(3):90-95. [doi: 10.1109/mcse.2007.55]
39. Youden WJ. Index for rating diagnostic tests. Cancer 1950;3(1):32-35 [FREE Full text] [doi: 10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3] [Medline: 15405679]
40. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. 2016 Presented at: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, California, USA p. 1135-1144. [doi: 10.1145/2939672.2939778]
41. Shapley LS. A value for n-person games. In: Kuhn HW, editor. Classics in Game Theory. Princeton, NJ: Princeton University Press; 1997:69-79.
42. Coyne JC, Schwenk TL, Fechner-Bates S. Nondetection of depression by primary care physicians reconsidered. Gen Hosp Psychiatry 1995;17(1):3-12. [doi: 10.1016/0163-8343(94)00056-j] [Medline: 7737492]
43. Agteren J, Iasiello M. Advancing our understanding of mental wellbeing and mental health: the call to embrace complexity over simplification. Aust Psychol 2021;55(4):307-316. [doi: 10.1111/ap.12440]
44. Schaakxs R, Comijs HC, Lamers F, Beekman ATF, Penninx BWJH. Age-related variability in the presentation of symptoms of major depressive disorder. Psychol Med 2017;47(3):543-552. [doi: 10.1017/S0033291716002579] [Medline: 27786143]
45. Bernert S, Matschinger H, Alonso J, Haro JM, Brugha TS, Angermeyer MC, ESEMeD / MHEDEA 2000 investigators. Is it always the same? Variability of depressive symptoms across six European countries. Psychiatry Res 2009;168(2):137-144 [FREE Full text] [doi: 10.1016/j.psychres.2008.04.008] [Medline: 19481817]

46. Li XY, Liu Q, Chen P, Ruan J, Gong X, Luo D, et al. Predictors of professional help-seeking intention toward depression among community-dwelling populations: a structural equation modeling analysis. Front Psychiatry 2022;13:801231 [FREE Full text] [doi: 10.3389/fpsyt.2022.801231] [Medline: 35280177]
47. Fletcher K, Parker GB, Manicavasagar V. Coping profiles in bipolar disorder. Compr Psychiatry 2013;54(8):1177-1184. [doi: 10.1016/j.comppsych.2013.05.011] [Medline: 23810079]
48. Lobban F, Taylor K, Murray C, Jones S. Bipolar disorder is a two-edged sword: a qualitative study to understand the positive edge. J Affect Disord 2012;141(2-3):204-212. [doi: 10.1016/j.jad.2012.03.001] [Medline: 22472729]
49. Mitchell PB, Loo CK, Gould BM. Diagnosis and monitoring of bipolar disorder in general practice. Med J Aust 2010;193(S4):S10-S13. [doi: 10.5694/j.1326-5377.2010.tb03890.x] [Medline: 20712553]
50. Bowden CL. Strategies to reduce misdiagnosis of bipolar depression. Psychiatr Serv 2001;52(1):51-55 [FREE Full text] [doi: 10.1176/appi.ps.52.1.51] [Medline: 11141528]
51. El-Mallakh RS, Vöhringer PA, Ostacher MM, Baldassano CF, Holtzman NS, Whitham EA, et al. Antidepressants worsen rapid-cycling course in bipolar depression: a STEP-BD randomized clinical trial. J Affect Disord 2015;184:318-321 [FREE Full text] [doi: 10.1016/j.jad.2015.04.054] [Medline: 26142612]
52. Showraki M, Showraki T. GAD: over-reactive and unstable mood. J Affect Disord Rep 2021;6:100265 [FREE Full text] [doi: 10.1016/j.jadr.2021.100265]
53. Gunderson JG, Herpertz SC, Skodol AE, Torgersen S, Zanarini MC. Borderline personality disorder. Nat Rev Dis Primers 2018;4:18029. [doi: 10.1038/nrdp.2018.29] [Medline: 29795363]
54. Zimmerman M, Morgan TA. The relationship between borderline personality disorder and bipolar disorder. Dialogues Clin Neurosci 2013;15(2):155-169 [FREE Full text] [doi: 10.31887/DCNS.2013.15.2/mzimmerman] [Medline: 24174890]
55. Akiskal HS, Hantouche EG, Allilaire JF. Bipolar II with and without cyclothymic temperament: "dark" and "sunny" expressions of soft bipolarity. J Affect Disord 2003;73(1-2):49-57. [doi: 10.1016/s0165-0327(02)00320-8] [Medline: 12507737]
56. Marsee K, Gross AF. Bipolar disorder or something else? Curr Psychiatry 2013;12(2):43-49 [FREE Full text]
57. National Institute for Health and Care Excellence (Great Britain). Bipolar Disorder: Assessment and Management. London: National Institute for Health and Care Excellence (NICE); 2019.
58. Baldessarini RJ, Tondo L, Vazquez GH, Undurraga J, Bolzani L, Yildiz A, et al. Age at onset versus family history and clinical outcomes in 1,665 international bipolar-I disorder patients. World Psychiatry 2012;11(1):40-46 [FREE Full text] [doi: 10.1016/j.wpsyc.2012.01.006] [Medline: 22295008]
59. Hidden waits force more than three quarters of mental health patients to seek help from emergency services. Royal College of Psychiatrists. 2022. URL: https://www.rcpsych.ac.uk/news-and-features/latest-news/detail/2022/10/10/hidden-waits-force-more-than-three-quarters-of-mental-health-patients-to-seek-help-from-emergency-services [accessed 2023-12-19]
60. Magaard JL, Seeralan T, Schulz H, Brütt AL. Factors associated with help-seeking behaviour among individuals with major depression: a systematic review. PLoS One 2017;12(5):e0176730 [FREE Full text] [doi: 10.1371/journal.pone.0176730] [Medline: 28493904]
61. Tijhuis MA, Peters L, Foets M. An orientation toward help-seeking for emotional problems. Soc Sci Med 1990;31(9):989-995. [doi: 10.1016/0277-9536(90)90108-5] [Medline: 2255971]
62. Polacsek M, Boardman GH, McCann TV. Help-seeking experiences of older adults with a diagnosis of moderate depression. Int J Ment Health Nurs 2019;28(1):278-287 [FREE Full text] [doi: 10.1111/inm.12531] [Medline: 30120874]
63. Conner KO, McKinnon SA, Roker R, Ward CJ, Brown C. Mitigating the stigma of mental illness among older adults living with depression: the benefit of contact with a peer educator. Stig Health 2018;3(2):93-101. [doi: 10.1037/sah0000076]
64. Kiejna A, Rymaszewska J, Hadryś T, Suwalska A, Łojko D, Rybakowski JK. Bipolar or unipolar?—The question for clinicians and researchers. J Affect Disord 2006;93(1-3):177-183. [doi: 10.1016/j.jad.2006.03.008] [Medline: 16678275]
65. Ghaemi SN, Rosenquist KJ, Ko JY, Baldassano CF, Kontos NJ, Baldessarini RJ. Antidepressant treatment in bipolar versus unipolar depression. Am J Psychiatry 2004;161(1):163-165 [FREE Full text] [doi: 10.1176/appi.ajp.161.1.163] [Medline: 14702267]
66. Ghaemi SN. Treatment of rapid-cycling bipolar disorder: are antidepressants mood destabilizers? Am J Psychiatry 2008;165(3):300-302 [FREE Full text] [doi: 10.1176/appi.ajp.2007.07121931] [Medline: 18316425]
67. Kessler RC, Ustün TB. The World Mental Health (WMH) survey initiative version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). Int J Methods Psychiatr Res 2004;13(2):93-121 [FREE Full text] [doi: 10.1002/mpr.168] [Medline: 15297906]

Abbreviations

AUC: area under the receiver operating characteristic curve
BD: bipolar disorder
CCNR: Cambridge Centre for Neuropsychiatric Research
CIDI: Composite International Diagnostic Interview

CV: cross-validation
DSM-5: Diagnostic and Statistical Manual of Mental Disorders, 5th Edition
GP: general practitioner
ICD-10: International Classification of Diseases and Related Health Problems, 10th Revision
ICD-11: International Statistical Classification of Diseases and Related Health Problems, 11th Revision
LIME: local interpretable model-agnostic explanations
MDD: major depressive disorder
NICE: National Institute for Health and Care Excellence
PHQ-9: Patient Health Questionnaire-9
rNCV: repeated nested cross-validation
SHAP: Shapley additive explanations
SSRI: selective serotonin reuptake inhibitor
WEMWBS: Warwick-Edinburgh Mental Well-Being Scale
XGBoost: Extreme Gradient Boosting

Edited by J Torous; submitted 11.07.23; peer-reviewed by A Peckham, B Montezano; comments to author 02.09.23; revised version received 27.10.23; accepted 01.12.23; published 11.01.24.

Please cite as:
Benacek J, Lawal N, Ong T, Tomasik J, Martin-Key NA, Funnell EL, Barton-Owen G, Olmert T, Cowell D, Bahn S
Identification of Predictors of Mood Disorder Misdiagnosis and Subsequent Help-Seeking Behavior in Individuals With Depressive Symptoms: Gradient-Boosted Tree Machine Learning Approach
JMIR Ment Health 2024;11:e50738
URL: https://mental.jmir.org/2024/1/e50738
doi:10.2196/50738
PMID:38206660

Original Paper

Speech Features as Predictors of Momentary Depression Severity in Patients With Depressive Disorder Undergoing Sleep Deprivation Therapy: Ambulatory Assessment Pilot Study

Lisa-Marie Wadle[1], MSc; Ulrich W Ebner-Priemer[1,2], Prof Dr; Jerome C Foo[3,4,5], PhD; Yoshiharu Yamamoto[6], Prof Dr; Fabian Streit[3], PhD; Stephanie H Witt[3], PhD; Josef Frank[3], PhD; Lea Zillich[3], PhD; Matthias F Limberger[1], MA; Ayimnisagul Ablimit[7], MSc; Tanja Schultz[7], Prof Dr; Maria Gilles[2], MD; Marcella Rietschel[3], Prof Dr Med; Lea Sirignano[3], MSc

[1]Mental mHealth Lab, Institute of Sports and Sports Science, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]Department of Psychiatry and Psychotherapy, Central Institute of Mental Health, University of Heidelberg, Mannheim, Germany
[3]Department of Genetic Epidemiology in Psychiatry, Central Institute of Mental Health, University of Heidelberg, Mannheim, Germany
[4]Institute for Psychopharmacology, Central Institute of Mental Health, University of Heidelberg, Mannheim, Germany
[5]Department of Psychiatry, College of Health Sciences, University of Alberta, Edmonton, AB, Canada
[6]Educational Physiology Laboratory, Graduate School of Education, University of Tokyo, Tokyo, Japan
[7]Cognitive Systems Lab, University of Bremen, Bremen, Germany

Corresponding Author:
Lisa-Marie Wadle, MSc
Mental mHealth Lab
Institute of Sports and Sports Science
Karlsruhe Institute of Technology
Hertzstrasse 16, Bldg 06.31
Karlsruhe, 76187
Germany
Phone: 49 72160847543
Email: lisa.wadle@kit.edu

Abstract

Background: The use of mobile devices to continuously monitor objectively extracted parameters of depressive symptomatology is seen as an important step in the understanding and prevention of upcoming depressive episodes. Speech features such as pitch variability, speech pauses, and speech rate are promising indicators, but empirical evidence is limited, given the variability of study designs.

Objective: Previous research studies have found different speech patterns when comparing single speech recordings between patients and healthy controls, but only a few studies have used repeated assessments to compare depressive and nondepressive episodes within the same patient. To our knowledge, no study has used a series of measurements within patients with depression (eg, intensive longitudinal data) to model the dynamic ebb and flow of subjectively reported depression and concomitant speech samples. However, such data are indispensable for detecting and ultimately preventing upcoming episodes.

Methods: In this study, we captured voice samples and momentary affect ratings over the course of 3 weeks in a sample of patients (N=30) with an acute depressive episode receiving stationary care. Patients underwent sleep deprivation therapy, a chronotherapeutic intervention that can rapidly improve depression symptomatology. We hypothesized that within-person variability in depressive and affective momentary states would be reflected in the following 3 speech features: pitch variability, speech pauses, and speech rate. We parametrized them using the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) from open-source Speech and Music Interpretation by Large-Space Extraction (openSMILE; audEERING GmbH) and extracted them from a transcript. We analyzed the speech features along with self-reported momentary affect ratings, using multilevel linear regression analysis. We analyzed an average of 32 (SD 19.83) assessments per patient.

Results: Analyses revealed that pitch variability, speech pauses, and speech rate were associated with depression severity, positive affect, valence, and energetic arousal; furthermore, speech pauses and speech rate were associated with negative affect, and speech pauses were additionally associated with calmness. Specifically, pitch variability was negatively associated with improved momentary states (ie, lower pitch variability was linked to lower depression severity as well as higher positive affect, valence, and energetic arousal). Speech pauses were negatively associated with improved momentary states, whereas speech rate was positively associated with improved momentary states.

Conclusions: Pitch variability, speech pauses, and speech rate are promising features for the development of clinical prediction technologies to improve patient care as well as timely diagnosis and monitoring of treatment response. Our research is a step forward on the path to developing an automated depression monitoring system, facilitating individually tailored treatments and increased patient empowerment.

(JMIR Ment Health 2024;11:e49222) doi:10.2196/49222

KEYWORDS

ambulatory assessment; experience sampling; ecological momentary assessment; speech features; speech pattern; depression; sleep deprivation therapy; mobile phone

Introduction

Background

Depression is one of the most prevalent health disorders worldwide [1,2]. The World Health Organization predicted that depression would be 1 of the 3 leading causes of disease burden by 2030 [3], even before its prevalence increased owing to the COVID-19 pandemic [4]. This disorder has symptoms that include depressed mood, loss of energy and interest, sleep problems, and diminished ability to concentrate [5]; thus, depression imposes a substantial burden on the patients as well as their surroundings, society, and the economy [6]. Most importantly, depression is a chronic disorder, characterized by multiple episodes over years or decades. However, strategies for secondary prevention or early detection of new episodes are missing.

The diagnosis and severity assessment of depression relies mostly on self- or caregiver reports, which are prone to retrospective and social desirability bias [7,8]. In addition, such assessments are time and resource intensive because clinical specialists are needed over the course of treatment and recovery [9]. Moreover, many new episodes remain undiagnosed or untreated, that is, secondary prevention is the main issue [10,11]. To reduce burden, the timely detection and diagnosis of (new) depressive episodes are critical.

In recent years, research has focused on the identification of mental health disorder indicators that can be derived automatically, driven by technological developments [12,13]. In particular, the innovation of the ambulatory assessment research technique has contributed strongly to this endeavor [14]. Different terms have been used for this kind of methodology: *ambulatory assessment* [15], *ecological momentary assessment* [16], *experience sampling* [17], and *digital phenotyping* [18]. Although the terms differ, all approaches use computer-assisted methodology to assess momentary self-reported symptoms (eg, via electronic diaries [ediaries]), behaviors, or physiological processes, or actively or passively collect smartphone and physical data or context information (eg, via wearables) while the participant performs normal daily activities in their natural environment [19]. The main advantages of ambulatory assessment are (1) the ability to collect real-life data in real time, thereby reducing retrospective recall bias and increasing ecological validity; and (2) the ability to collect data continuously (passively), which allows us to capture dynamic changes. Accordingly, ambulatory assessment is a promising tool for the timely detection of upcoming clinical episodes to prevent further clinical deterioration [20-22]. In particular, parameters captured objectively by wearables are useful because they can be assessed passively with a high frequency over prolonged time periods [23].

Promising markers that can be assessed objectively are speech and language, which are also metaphorically called “the mirror of the soul” [24]. Even before objective measurements with ambulatory assessment technology were feasible, clinical observations described the voice of patients with depression as low, slow, and hesitant, with these patients speaking in a monotonous and expressionless manner [24,25]. Voice and speech production may be affected by typical characteristics of the clinical nature of depression; for example, psychomotor retardation, energy loss, and cognitive difficulties also affect the vocal folds, leading to a lower *intensity*, *rate*, and *loudness* of speech, which manifest in a monotone and toneless voice [26-28]. Recent reviews have highlighted the potential of using speech markers to assess a variety of psychiatric disorders [29], especially depression [30]. The use of speech as a marker has several advantages because it can be recorded (1) casually; (2) in a noninvasive manner at people’s homes or in public places (with consent provided); and (3) at low cost because microphones are integrated in many devices such as smartphones, smartwatches, and hearing aids. With the availability of open-source speech analysis software (eg, open-source Speech and Music Interpretation by Large-Space Extraction [openSMILE; audEERING GmbH] and Praat) and advances in automatic speech processing technologies based on machine learning techniques, research and development on the use of acoustic and linguistic features to identify mood disorders in particular [29] have been made possible.

Prior Work

Many studies have successfully discriminated between healthy controls and patients with depression based on speech features [30]. However, understanding within-person (vs between-person) depression-related voice changes is essential in detecting new episodes, that is, the secondary prevention. To the best of our knowledge, only a few studies in samples with clinical (not subclinical) depression have examined the variability of speech features within persons [31-36]. In a

6-week treatment-monitoring study, weekly speech samples were obtained from 35 patients with depression using an interactive voice response system [31]. Patients with an improvement in depressive symptoms showed a significant increase in *pitch* and *pitch variability*, an increase in *speech rate*, and shorter *speech pauses* while speaking at their final assessment compared with their baseline assessment. Importantly, patients whose depressive symptoms did not improve did not show these changes.

The data set of Mundt et al [31] was reanalyzed multiple times [32,34,35]. Quatieri and Malyska [34] integrated additional speech features and identified that lower *pitch variability*, *shimmer*, and *jitter* as well as an increased *harmonics-to-noise ratio* were correlated with lower depression severity. This is in contrast to the study by Mundt et al [31], who found that increased *pitch variability* was associated with lower depression severity, which Quatieri and Malyska [34] attributed to differences in the set of voice samples analyzed (read speech in the study by Mundt et al [31] and conversational speech in the study by Quatieri and Malyska [34] from the same patients).

Trevino et al [32] discussed *speech rate* extraction methods based on the data set of Mundt et al [31] and replicated results regarding *speech rate* in automatically derived phonologically based features. *Speech rate* was negatively correlated with depression scores and the psychomotor retardation item in particular. Moreover, the authors replicated the finding that *speech pauses* were positively correlated with depression severity.

Furthermore, Horwitz et al [35] reanalyzed a subset of data from the study by Mundt et al [31] with a focus on disentangling how speech features relate to the total assessment score and individual symptom items. The authors found a positive correlation between *pitch variability* and depression scores and a slower *speech rate* with increasing depression severity. Notably, they analyzed a different speech task and a different depression assessment in comparison with Mundt et al [31].

Mundt et al [33] replicated their results from Mundt et al [31] in a larger study. Here, 105 patients were observed in a 4-week randomized placebo-controlled study. Again, analyses entailed a comparison of the final and baseline assessments. For patients benefiting from the treatment, *total pause time* was lower, *pitch* was higher (*pitch variability* was not assessed), and *speech rate* was higher. For patients who did not benefit from the treatment, only *speech rate* increased; however, it increased significantly less than in patients benefiting from the treatment.

Yang et al [36] analyzed clinical interviews recorded in 7-week intervals. In contrast to Mundt et al [31], they did not find a change in *pitch variability* with a change in depression severity in the patients but rather in the interviewers. The authors also found shorter *switching pauses* between patient and interviewer (ie, both interlocutors) with lower depression severity.

Although not completely consistent, these findings support the assumption that voice features change within individuals when depression severity changes. However, although data were collected at multiple time points during the study (eg, weekly), except in the study by Yang et al [36], the analysis was limited to a comparison between the baseline and final assessments. However, given that the goal is to detect and ultimately prevent new depressive episodes and deterioration, it is essential to understand within-person trajectories of voice features and how they are associated with momentary states with increased granularity. In this study, we used a naturalistic data set where a rapidly acting antidepressant treatment (ie, sleep deprivation therapy [SDT] [37]) was applied to patients experiencing a depressive episode. The antidepressant effect vanishes in most of the cases after recovery sleep. Baseline, the treatment effect of SDT, and relapse can be measured in a matter of 4 days, making it a preferable setting to study within-person fluctuations.

Aims and Hypotheses

To investigate the within-person relationship between fluctuations in depression severity and fluctuations in speech features, we used a longitudinal data set with an average of 32 (SD 19.83) assessments per patient. All patients had experienced an acute depressive episode and undergone SDT [37], a chronotherapeutic intervention that can rapidly improve depression symptomatology. The main advantage of this therapeutic is that we maximize the variance of affective states within the data set and ensure sufficient within-person fluctuations over time. As the amount of speech features is immense, resulting in alpha error inflation, we focused on 3 speech features with high face validity that have shown first hints in past research [31-36]. Specifically, we hypothesized that (1) changes in *pitch variability*, (2) shorter *speech pauses*, and (3) higher *speech rate* are associated with lower depression severity. In addition, we assessed the associations of these features with additional momentary affective states (ie, positive affect, negative affect, valence, energetic arousal, and calmness). We hypothesized that the associations of speech features with negative affect are similar to those for depression severity and that the associations of speech features with the other momentary affective states listed follow the opposite pattern.

Methods

Sample

We used a data set that was collected as part of a pilot study (Sleep Deprivation and Gene Expression [SLEDGE II]; German Clinical Trials Register: DRKS00022025) gathering digital phenotypes and multiomics data in a clinical sample undergoing SDT at the Central Institute of Mental Health in Mannheim, Germany. A total of 30 inpatients experiencing acute depressive episodes were enrolled in the study. The patients were diagnosed according to the *International Classification of Diseases, Tenth Revision* (ICD-10), codes by the senior clinician at admittance to the hospital. All patients received treatment as usual, which also included SDT (for a list of medications, refer to Textbox S1 in Multimedia Appendix 1). Exclusion criteria were comorbid substance use disorders or personality disorders. From this sample of 30 patients, the complete data sets of 8 (27%) patients were excluded from the final analyses (n=4, 50% did not record any videos; n=1, 13% did not say anything during the videos [23 videos]; n=2, 25% had no sound recorded in the videos owing to technical issues [30 recordings]; and n=1, 13%

recorded only 2 videos); thus, the final sample consisted of 22 (73%) patients (n=12, 55% male) aged between 18 and 63 (mean 33.5, SD 12.4; median 29, IQR 23.25-42.75) years.

Ethical Considerations

The study was approved by the Ethics Committee II of the Medical Faculty Mannheim, University of Heidelberg (2013-563N-MA). All patients received detailed information about the aims and procedures of the study and provided informed consent. Patients could withdraw from the study at any time and did not receive any compensation for participation. Data was deidentified to ensure privacy.

Study Procedure

Patients were given a study smartphone (Nokia 4.2 or Samsung Galaxy J7) at the beginning of the study (day 0), instructed on how to use it, and (if necessary) performed test runs supervised by the study personnel. A telephone number for technical support and an information sheet regarding the ambulatory assessment procedure were handed out. Data were collected using movisensXS software (movisens GmbH) [38]. Patients underwent SDT as part of their depression treatment, which involves staying awake for approximately 36 hours. Treatment effect and relapse can be measured in a matter of 4 days [37], thus ensuring a maximum of within-person variance in the data set. After at least 1 day of baseline assessment (day 0), SDT was conducted on day 1. Patients stayed awake from 6 AM on day 1 to 6 PM on day 2. Recovery sleep was allowed from 6 PM on day 2 until 1 AM on day 3. Data were collected before, during, and after SDT for up to 26 days. In the first week of the study, smartphones sent prompts 3 times per day (morning, afternoon, and evening); in addition, self-initiated assessments were possible to report specific events or to catch up with missed assessments. To reduce the burden on patients, the sampling schema was altered to 2 prompts per day (morning and evening). With each prompt, patients were requested to fill out items concerning their affective state and to record a selfie video reporting how they felt currently. Patients returned the smartphone at the end of the study. The study personnel uploaded the data from the smartphones to the movisensXS platform [38] and then downloaded the data for analysis.

Ambulatory Assessment: eDiary Ratings and Selfie Videos

The data set contains 3 sets of momentary assessments in the form of eDiary ratings at each prompt (Textboxes S2-S4 in Multimedia Appendix 1): (1) the short version of the Allgemeine Depressionsskala (ADS-K) [39] adapted to momentary assessment with 14 items on depressive mood rated on a scale ranging from 0=*rarely* to 3=*mostly* (we left out the item regarding sleep from the original questionnaire because its inclusion was not reasonable in the momentary assessment design); (2) a total of 15 positive (cheerful, content, energetic, enthusiastic, relaxed, and happy) and negative (lonely, sad, insecure, anxious, depressed, low-spirited, guilty, distrustful, and irritable) affect items [40] rated on a 5-point Likert scale ranging from 1=*not at all* to 5=*very much*; and (3) a 6-item short version of the Multidimensional Mood Questionnaire (MDMQ) [41] capturing time-varying momentary fluctuations in daily life on the affect dimensions of valence (*unwell* to *well* and *discontent* to *content*), energetic arousal (*without energy* to *full of energy* and *tired* to *awake*), and calmness (*tense* to *relaxed* and *agitated* to *calm*). The items were presented on visual analog scales with 2 poles and a slider from 0 to 100. For each of the constructs, we computed mean values per scale, resulting in 6 outcome variables (depressive symptoms, positive affect, negative affect, valence, energetic arousal, and calmness). For the ADS-K, we also report sum scores as described in the tool's manual; however, to increase comparability among outcomes, we used the mean value for analyses. If necessary, we recoded items such that higher values indicated a (1) higher intensity of depressive symptoms, (2) higher positive affect, (3) higher negative affect, (4) higher positive valence, (5) higher energetic arousal, and (6) higher calmness.

In addition to the aforementioned eDiary ratings, patients were requested to record selfie videos with the following instructions: "Please keep the camera stable during the recording and record your whole face. Please describe in 10-20 seconds how you currently feel."

Clinical Assessments

The Montgomery–Åsberg Depression Rating Scale (MADRS) [42] was completed in the morning at 4 time points (baseline, morning before sleep deprivation, 1 week after sleep deprivation, and 2 weeks after sleep deprivation) and once at midday (the day after sleep deprivation night). The MADRS is a 10-item expert assessment of depressive symptom severity over the past week, with items rated on a 7-point scale ranging from 0 to 6; higher scores indicate higher severity.

Data Preprocessing

The data set contained 899 selfie videos in mp4 format. The full set of videos of 4 (13%) of the 30 patients had to be excluded owing to the reasons mentioned previously (55/899, 6.1%) and additional 2 videos had to be excluded because of technical damage (2/899, 0.02%). As our research questions focused on audio data (not visual data), we extracted the audio tracks of the remaining 842 (93.66%) from the original 899 selfie videos using the *ffmpeg* package in Python and archived them as wav files (sampling rate: 48 kHz; mono=1 channel). We excluded test runs (14/842, 1.7%), accidental short recordings with no content (29/842, 3.4%), recordings during which the microphone was masked by the patient (27/842, 3.2%), and assessments in which 1 of the 2 corresponding assessments (speech or affective state) was missing (18/842, 2.1%). In addition, if 2 consecutive assessments were <15 minutes apart from each other, only the first assessment was kept unless its audio quality was insufficient or only the second assessment included assessments of affective states; in such cases, the second assessment was kept (21/842, 2.5%). We also excluded recordings with background noise that restricted speech intelligibility (9/842, 1.1%) or that included the speech of third parties (8/842, 1%). We filtered the remaining recordings (716/842, 85%) using DeepFilterNet2 [43] to remove background noise.

Acoustic Features

For our main analyses, we focused on the acoustic features *pitch variability*, *speech pauses*, and *speech rate* (Table 1). We restricted the number of features to limit α error inflation and selected specifically these 3 features because they revealed sufficient empirical support to warrant an explicit hypothesis. We extracted acoustic features of the final recordings (n=716) using the functionals (v02) of the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [44] of the open-source toolkit openSMILE implemented in Python [45,46]. eGeMAPS is a minimalistic set of acoustic features recommended for clinical speech analysis; it helps to guarantee comparability between studies, given the proliferation of speech features. Features related to frequency, energy, spectrum, and tempo are included in the set. *Pitch variability* is represented by the SD of the logarithmic fundamental frequency (F0) on a semitone frequency scale starting at 27.5 Hz and measured in hertz. F0 is the lowest frequency of the speech signal and is perceived as *pitch*. *Speech pauses* are approximated as the mean length of unvoiced regions (F0=0) measured in seconds. With respect to *speech rate*, a transcription of the recordings is necessary, which we obtained using an automatic speech recognition system according to published procedures [47]. We corrected the transcripts manually. To determine *speech rate*, we calculated the ratio of words divided by the duration of the voice sample.

Table 1. Overview of extracted speech features.

Speech feature	Technical feature	Explanation
Pitch variability	F0semitoneFrom27.5Hz_sma3nz_stddevNorm	SD of the F0 perceived as the extent to which a person's *pitch* changes (in Hz)
Speech pauses	MeanUnvoicedSegmentLength	Mean of the length of unvoiced regions approximating silent parts of the speech sample (in seconds)
Speech rate	Words per second	Ratio of words counted on the basis of the automatically transcribed and manually corrected text divided by the duration of the speech sample

Beside our main analyses based on *pitch variability*, *speech pauses*, and *speech rate*, we decided to integrate further eGeMAPS features in an exploratory analysis. These features have been recommended in the context of affective states in particular because they contain additional cepstral and dynamic features [44]. We included the following features in the exploratory analyses: for voiced and unvoiced regions together, the mean and SD of the mel-frequency cepstral coefficients (MFCCs) 1 to 4 and spectral flux difference of the spectra of 2 consecutive frames; for voiced regions, the formant 2 to 3 bandwidths along with spectral flux and MFCCs 1 to 4; and for unvoiced regions, the mean and SD of the spectral flux [44].

Statistical Analysis

In addition to the *mean*, *SD*, and *range*, we present *min* and *max* as the mean of all patients' minimum and maximum scores, respectively, of each parameter throughout the whole study. Moreover, following the recommendations by Snijders and Bosker [48], we computed Pearson correlation analyses with person-mean–centered variables to evaluate the relationship between affective scores and speech features. To generate person-mean–centered variables, we subtracted the individual's mean from their score, which represents the variation around the individual's mean.

To evaluate psychometric properties, we calculated McDonald ω as the reliability coefficient using the *multilevelTools* package in R. For the MDMQ subscales, we used the *misty* package in R to calculate the Spearman-Brown corrected correlation coefficients because the subscales consist of only 2 items [49]. For the MADRS score at the time of inclusion, we calculated Cronbach α using the *psych* package in R.

To analyze the within-person association of speech features and subjectively evaluated affective states, we used multilevel modeling [48] using the *nlme* package in R. Multilevel modeling offers two specific advantages for the given data: (1) separation of within-person effects from between-person effects and (2) allowing and considering different numbers of assessments per patient. Before the analyses, we centered time-variant level-1 predictors (*pitch variability*, *speech pauses*, and *speech rate*) at the person level and included the predictors *time* and $time^2$ in minutes (each centered at 2 PM) as covariates. To facilitate the comparison of the magnitude of effects among different predictors, we report standardized beta coefficients (standardized β) according to the recommendations by Hox and van de Schoot [50] following the equation: standardized $\beta = \beta \times (SD_{predictor} / SD_{outcome})$. We further calculated Hox R^2 values according to the recommendation by Hox and Maas [51] following the equation: $R^2_{Hox} = (\sigma^2_{null} - \sigma^2_{model}) / \sigma^2_{null}$. We set the α level at 5% and applied Bonferroni corrections for exploratory analyses (α_{adj}=.002). We performed all analyses in R (version 4.2.1, 2022-06-23).

Our analyses can be split into 4 parts: the calculation of intraclass correlation coefficients (ICCs); separate models with all speech features as predictors and all affective scores as outcomes; combined models with all speech features as simultaneous predictors; and exploratory analyses, including additional speech features. Specifically, we first descriptively investigated whether our study procedure resulted in sufficient within-person variance. For this purpose, we calculated ICCs, including all momentary affective ratings and speech recordings, regardless of whether they were assessed before, during, or after SDT. In general, the ICC indicates the amount of between-person variance in unconditional (null) models. The 2-level models analyzed contained repeated measures (level 1) that were nested within patients (level 2). The second step contained our main analysis: we calculated separate models for each speech feature (*pitch variability* [model set 1], *speech pauses* [model set 2], and *speech rate* [model set 3]) and each affective state (depression severity [ADS-K], positive affect, negative affect, valence, energetic arousal, and calmness), resulting in 18 models. In the third step, to evaluate the relative significance of *pitch variability*, *speech pauses*, and *speech rate*, we constructed combined models for each of the affective scores, including all 3 features simultaneously (6 models). In the fourth step, exploratory analyses were conducted with the inclusion of 24 additional speech features from eGeMAPS (Textbox S5 in Multimedia Appendix 1). These features were used as predictors for each of the affective scores separately.

Results

Descriptive Statistics

We included 716 speech-state pairs (mean 32, SD 19.83 per patient) in the final analysis. The mean MADRS score at the time of inclusion assessment was 30.1 (SD 5.8). This corresponds to 18 (82%) patients with moderate depression and 4 patients (18%) with severe depression out of 22 patients at study inclusion.

Regarding depressive symptoms (ADS-K; scale 0-3), patients had a mean score of 1.2 (SD 0.6; min 0.7, max 2.0) and a mean sum score of 16.9 (SD 8.1; min 9.6, max 26.1). At inclusion, the mean ADS-K score was 1.4 (SD 0.6; range 0.4-2.8), and the mean sum score was 20.0 (SD 8.4; range 6-39). For positive and negative affect (scale 1-5), the mean scores were 2.1 (SD 0.8; min 1.3, max 3.1) and 2.3 (SD 1.0; min 1.4, max 3.9), respectively; on the MDMQ (scale 1-100) valence subscale, the mean score was 44.9 (SD 21.5; min 9.4, max 67.5); on the energetic arousal subscale, the mean score was 41.7 (SD 21.0; min 16.4, max 62.7); and on the calmness subscale, the mean score was 43.8 (SD 22.8; min 6.9, max 70.7). The ICCs were 0.47 for the ADS-K, 0.45 for positive affect, 0.59 for negative affect, 0.27 for energetic arousal, 0.25 for valence and 0.40 for calmness, that is, the following amount of variance in the momentary assessments can be attributed to within-person fluctuations: 53% for the ADS-K, 55% for positive affect, 41% for negative affect, 73% for energetic arousal, 75% for valence, and 60% for calmness.

Regarding speech features, the mean *pitch variability* was 0.32 Hz (SD 0.09; min 0.14, max 0.44), the mean *speech pause* length was 0.26 seconds (SD 0.12; min 0.17, max 0.47), and the mean *speech rate* was 1.77 words per second (SD 0.57; min 1.16, max 2.75). The ICCs were 0.66 for *pitch variability*, 0.36 for *speech pauses*, and 0.57 for *speech rate*. This corresponds to the following amount of variance in the speech feature assessments that can be attributed to within-person fluctuations: 34% for *pitch variability*, 64% for *speech pauses*, and 43% for *speech rate*.

Correlational analyses (Figure S1 in Multimedia Appendix 1) included between 698 and 716 observations depending upon the specific pairing. We found correlations among and between affective scores and speech features, except for *pitch variability* and *speech rate*, neither of which correlated with negative affect and calmness; in addition, there was no correlation between *pitch variability* and *speech rate*. Specifically, ADS-K scores correlated negatively with positive affect, all MDMQ subscales, and *speech rate* and correlated positively with negative affect, *pitch variability*, and *speech pauses*. Negative affect showed the same pattern, except for the pairings with *pitch variability* and *speech rate*, for which no correlations were found. Regarding positive affect, we found the opposite correlation pattern, that is, positive correlations with all MDMQ subscales and *speech rate* and negative correlations with *pitch variability* and *speech pauses*. The MDMQ subscales showed the same relationships as positive affect, except for the pairing between calmness and *pitch variability* and *speech rate*, for which no correlations were found. Within speech features, we found a negative correlation between *pitch variability* and *speech pauses*, no correlation between *pitch variability* and *speech rate*, and a negative correlation between *speech pauses* and *speech rate.* Overall, correlations among affective scores were strong (r>0.5). Correlations among speech features as well as between affective scores and speech features were weak (r<0.2), except for a strong negative correlation between *speech pauses* and *speech rate.*

The psychometric properties for momentary affective ratings were good to excellent. Specifically, McDonald ω values [52] were 0.87 (within-person) and 0.90 (between-person) for depressive symptoms (ADS-K), 0.87 (within-person) and 0.95 (between-person) for positive affect, and 0.87 (within-person) and 0.96 (between-person) for negative affect. The Spearman-Brown coefficients were 0.83 (within-person) and 0.94 (between-person) for valence, 0.74 (within-person) and 0.89 (between-person) for energetic arousal, and 0.74 (within-person) and 0.89 (between-person) for calmness. Cronbach α for the MADRS score at the time of inclusion was acceptable (.67).

Association Between Speech Features and Momentary Affective Scores

Overview

In Tables 2 and 3, we present the fixed effects of *pitch variability*, *speech pauses*, and *speech rate* separately for each affective state. Details, including the effects of time and time^2, are presented in Table S1 in Multimedia Appendix 1.

Table 2. Multilevel linear regression analysis to predict depression and positive and negative affect: fixed effects of pitch variability, speech pauses, and speech rate.

Predictors	Outcome														
	ADS-K[a]					Positive affect					Negative affect				
	β	Standardized β	SE	R^2_{Hox}, %	P value	β	Standardized β	SE	R^2_{Hox}, %	P value	β	Standardized β	SE	R^2_{Hox}, %	P value
Model set 1															
Intercept	1.27	N/A[b]	0.10	N/A	<.001	2.10	N/A	0.13	N/A	<.001	2.45	N/A	0.16	N/A	<.001
Pitch variability	.88	.14	0.32	1	.007	−1.50	−.18	0.42	1	<.001	.85	.08	0.43	1	.05
Model set 2															
Intercept	1.27	N/A	0.10	N/A	<.001	2.09	N/A	0.13	N/A	<.001	2.46	N/A	0.16	N/A	<.001
Speech pauses	.52	.10	0.18	1	.005	−1.16	−.17	0.24	17	<.001	.76	.09	0.25	2	.002
Model set 3															
Intercept	1.27	N/A	0.10	N/A	<.001	2.10	N/A	0.13	N/A	<.001	2.45	N/A	0.16	N/A	<.001
Speech rate	−.11	−.10	0.05	<1	.02	.26	.18	0.06	2	<.001	−.13	−.08	0.07	1	.04

[a]ADS-K: Allgemeine Depressionsskala.

[b]N/A: not applicable.

Table 3. Multilevel linear regression analysis to predict valence, energetic arousal, and calmness: fixed effects of pitch variability, speech pauses, and speech rate.

Predictors	Outcome														
	Valence					Energetic arousal					Calmness				
	β	Standardized β	SE	R^2_{Hox}, %	P value	β	Standardized β	SE	R^2_{Hox}, %	P value	β	Standardized β	SE	R^2_{Hox}, %	P value
Model set 1															
Intercept	43.72	N/A[a]	2.70	N/A	<.001	42.82	N/A	2.71	N/A	<.001	40.97	N/A	3.39	N/A	<.001
Pitch variability	−36.50	−.16	13.61	1	.008	−33.21	−.15	12.48	1	<.001	−11.52	−.05	12.82	<1	.37
Model set 2															
Intercept	43.26	N/A	2.69	N/A	<.001	42.71	N/A	2.71	N/A	<.001	40.58	N/A	3.39	N/A	<.001
Speech pauses	−34.06	−.19	7.71	3	<.001	−14.06	−.08	7.14	1	.049	−24.27	−.12	7.27	5	<.001
Model set 3															
Intercept	43.56	N/A	2.70	N/A	<.001	42.77	N/A	2.71	N/A	<.001	40.86	N/A	3.39	N/A	<.001
Speech rate	6.49	.17	2.03	2	.001	4.13	.11	1.87	1	.03	3.43	.09	1.91	5	.07

[a]N/A: not applicable.

ADS-K Scores

In the column entitled *ADS-K* (Table 2), we report the results of all models with ADS-K scores as the outcome. *Pitch variability* (standardized β=.14; *P*=.007), *speech pauses* (standardized β=.10; *P*=.005), and *speech rate* (standardized β=−.10; *P*=.02) were significantly associated with the ADS-K score, indicating that higher *pitch variability*, longer *speech pauses*, and lower *speech rate* are associated with more severe depressive symptomatology.

Positive and Negative Affect

In the columns entitled *Positive affect* and *Negative affect* (Table 2), we show results for positive affect and negative affect, respectively, as outcomes. *Pitch variability* (standardized β=−.18; *P*<.001), *speech pauses* (standardized β=−.17; *P*<.001), and *speech rate* (standardized β=.18; *P*<.001) were significantly associated with positive affect, indicating that lower *pitch variability*, shorter *speech pauses*, and higher *speech rate* are associated with higher positive affect. The associations between negative affect and speech features were in the opposite direction of the associations between positive affect and the speech features just presented: *speech pauses* (standardized β=.09; *P*=.002) and *speech rate* (standardized β=−.08; *P*=.04) were significantly associated with negative affect, indicating that longer *speech pauses* and lower *speech rate* are associated with higher negative affect. We further found a trend with respect to the association between *pitch variability* and negative affect, but this result was not statistically significant (standardized β=.08; *P*=.05). In addition, we found trends with respect to the associations between negative affect and time and negative affect and time², specifically in the models that included *pitch variability* (time: standardized β=.04; *P*=.08), *speech pauses* (time: standardized β=.04; *P*=.08; time²: standardized β<.01; *P*=.06), and *speech rate* (time: standardized β=.04; *P*=.09), but these results were not statistically significant.

MDMQ Results

In the columns entitled *Valence*, *Energetic arousal*, and *Calmness* (Table 3), we present the results for the MDMQ. *Pitch variability* (standardized β=−.16; *P*=.008), *speech pauses* (standardized β=−.19; *P*<.001), and *speech rate* (standardized β=.17; *P*=.001) were significantly associated with valence, indicating that lower *pitch variability*, shorter *speech pauses*, and higher *speech rate* are associated with higher (ie, positive) valence. In the model that included valence and *speech pauses*, we found a significant association between time² and valence (standardized β<.001; *P*=.03). In addition, we found trends with respect to the associations between valence and time², specifically in the models that included *pitch variability* (time: standardized β<.01; *P*=.098) and *speech rate* (time: standardized β<.01; *P*=.07), but these results were not statistically significant. Moreover, *pitch variability* (standardized β=−.15; *P*<.001), *speech pauses* (standardized β=−.08; *P*=.049), and *speech rate* (standardized β=.11; *P*=.03) were significantly associated with energetic arousal, indicating that lower *pitch variability*, shorter *speech pauses*, and higher *speech rate* are associated with higher energetic arousal. In all model combinations of energetic arousal and each speech feature, we found significant associations between time and energetic arousal (standardized β=−.11; *P*<.001) and time² and energetic arousal (standardized β<.01; *P*<.001). Furthermore, *speech pauses* (standardized β=−.12; *P*<.001) were significantly associated with calmness, indicating that shorter *speech pauses* are associated with greater calmness. In all model combinations of calmness and each speech feature, we found significant associations between time² and calmness (standardized β<.01; *P*= .013 for *pitch variability*, *P*=.003 for *speech pauses*; *P*=.009 for *speech rate*). In addition, we found a trend with respect to the association between *speech rate* and calmness (standardized β=.09; *P*=.07), but this result was not statistically significant.

Combined Models

In Tables 4 and 5, we display the results for the combined models that included all 3 speech features. In the model of ADS-K scores, associations with *pitch variability* (standardized β=.17; *P*<.001) and *speech pauses* (standardized β=.12; *P*=.01) remained statistically significant. Regarding positive affect, associations with *pitch variability* (standardized β=−.23; *P*<.001) and *speech pauses* (standardized β=−.19; *P*<.001) remained statistically significant. We further found a trend regarding the association between positive affect and time (standardized β=−.05; *P*=.09), but this result was not statistically significant. Regarding negative affect, associations with *pitch variability* (standardized β=.12; *P*=.008), *speech pauses* (standardized β=.12; *P*=.005), time (standardized β=.05; *P*=.03), and time² (standardized β<.01; *P*=.03) remained statistically significant. In the model of valence, associations with *pitch variability* (standardized β=−.22; *P*<.001), *speech pauses* (standardized β=.22; *P*<.001), and time² (standardized β<.01; *P*=.01) remained statistically significant. Regarding energetic arousal, associations with *pitch variability* (standardized β=−.17; *P*=.003), time (standardized β=.12; *P*<.001), and time² (standardized β<.01; *P*<.001) remained statistically significant. Regarding calmness, associations with *speech pauses* (standardized β=−.17; *P*=.002) and time² (standardized β<.01; *P*=.002) remained statistically significant. We further found a trend for the association between calmness and *pitch variability* (standardized β=.09; *P*=.097), but this result was not statistically significant.

Table 4. Multilevel linear regression analysis to predict momentary depression, positive affect, and negative affect: fixed effects of the combined models that included pitch variability, speech pauses, speech rate, time, and time².

Predictors[a]	Outcome											
	ADS-K[b]				Positive affect				Negative affect			
	β	Standardized β	SE	*P* value	β	Standardized β	SE	*P* value	β	Standardized β	SE	*P* value
Intercept	1.28	N/A[c]	0.10	<.001	2.08	N/A	0.13	<.001	2.47	N/A	0.16	<.001
Time	<.01	.02	<0.01	.42	<−.01	−.05	<0.01	.09	<.01	.05	<0.01	.03
Time²	<.01	<.001	<0.01	.44	<.01	<.001	<0.01	.31	<.01	<.01	<0.01	.03
Pitch variability	1.11	.17	0.33	<.001	−1.96	−.23	0.43	<.001	1.19	.12	0.45	.008
Speech pauses	.64	.12	0.26	.01	−1.29	−.19	0.33	<.001	.99	.12	0.35	.005
Speech rate	<−.01	<.001	0.07	.99	.04	.03	0.09	.66	.04	.02	0.09	.68

[a]R^2_{Hox} for ADS-K=2%, for positive affect=6%, and for negative affect=2%.

[b]ADS-K: Allgemeine Depressionsskala.

[c]N/A: not applicable.

Table 5. Multilevel linear regression analysis to predict momentary valence, energetic arousal, and calmness: fixed effects of the combined models that included pitch variability, speech pauses, speech rate, time, and time².

Predictors[a]	Outcome											
	Valence				Energetic arousal				Calmness			
	β	Standardized β	SE	*P* value	β	Standardized β	SE	*P* value	β	Standardized β	SE	*P* value
Intercept	42.95	N/A[b]	2.68	<.001	42.48	N/A	2.71	<.001	40.45	N/A	3.38	<.001
Time	<.01	.03	<0.01	.48	<−.01	.12	<0.01	<.001	<−.01	.01	<0.01	.89
Time²	<.01	<.01	<0.01	.01	<.01	<.01	<0.01	<.001	<.01	<.01	<0.01	.002
Pitch variability	−49.01	−.22	13.76	<.001	−37.74	−.17	12.78	.003	−21.75	.09	13.07	.097
Speech pauses	−41.01	.22	10.73	<.001	−12.97	.07	9.97	.19	−32.53	−.17	10.20	.002
Speech rate	−.64	.02	2.76	.82	1.96	.05	2.56	.44	−2.28	.06	2.62	.38

[a] R^2_{Hox} for valence=4%, for energetic arousal=5%, and for calmness=2%.

[b] N/A: not applicable.

Exploratory Analysis

Analyzing additional speech features, we found significant associations of the *equivalent sound level*, the mean of *spectral flux*, and the mean of *spectral flux of voiced regions only*, individually, with all affective scores (Table S2 in Multimedia Appendix 1). With respect to *equivalent sound level*, this indicates that louder voice samples were linked to improved affective states (ADS-K: standardized β=−.30; positive affect: standardized β=.34; negative affect: standardized β=−.21; valence: standardized β=.29; energetic arousal: standardized β=.26; and calmness: standardized β=.19); with respect to the mean of *spectral flux*, this indicates that a faster change in the spectrum was linked to better affective states (ADS-K: standardized β=−.22, positive affect: standardized β=.28, negative affect: standardized β=−.15, valence: standardized β=.21, energetic arousal: standardized β=.17, and calmness: standardized β=.27); and with respect to the mean of *spectral flux of voiced regions only*, this indicates that a faster change in the spectrum in voiced regions was linked to better affective states (ADS-K: standardized β=−.23, positive affect: standardized β=.28, negative affect: standardized β=−.15, valence: standardized β=.20, energetic arousal: standardized β=.20, and calmness: standardized β=.16). Regarding the additional speech features, the following significant associations were found: the mean of *spectral flux of unvoiced regions only* was associated with positive affect, indicating that a faster change in the spectrum in unvoiced regions was linked to improved positive affect (standardized β=.13); and the mean of the *MFCC 2 of voiced regions only* was significantly associated with energetic arousal, indicating that a higher mean was linked to lower energetic arousal (standardized β=−.15). Furthermore, we revealed a significant association between the SD of the *MFCC 4 of voiced regions only* ADS-K scores (standardized β=.13) as well as positive affect (standardized β=−.10) and negative affect (standardized β=.09). Specifically, smaller SDs were linked to higher positive affect, reduced negative affect, and lower ADS-K scores.

Discussion

Principal Findings

This is the first study to investigate whether speech features are associated with depression severity and momentary affective states in a longitudinal data set of patients with a depressive episode undergoing SDT. Our findings showed that lower *pitch variability*, higher *speech rate*, and shorter *speech pauses* were associated with better momentary states (ie, lower depression severity; higher positive affect and lower negative affect; and higher positive valence, energetic arousal, and calmness), supporting prior clinical observations with innovative methods applied to an intensive longitudinal data set.

Lower depression severity was accompanied by shorter *speech pauses*. This is in line with past research findings reporting that shorter *speech pauses* were associated with lower depression severity [31-33,36]. Our findings extend prior results because we also found an association between *speech pauses* and affective states more broadly, not limited to depressed mood. Regarding *speech rate*, we revealed associations with depression severity and all other affective state scales except for calmness. In particular, we found that higher *speech rate* was associated with lower depression symptomatology and lower negative affect, higher positive affect, higher positive valence, and higher energetic arousal. This is in line with prior research [31-33,35], in which a higher *speech rate* was found for patients who benefited from treatment.

Regarding *pitch variability*, we found support for our hypothesis that *pitch variability* changes with depression severity; more precisely, lower *pitch variability* was associated with lower depression symptomatology. This is in line with the studies by Quatieri and Malyska [34] and Horwitz et al [35], where a positive correlation between *pitch variability* and depression severity was found. However, the results reported in the studies by Mundt et al [31] and Yang et al [36] contrasted with ours and those found in the studies by Quatieri and Malyska [34] and Horwitz et al [35], that is, that higher *pitch variability* was associated with lower depression severity. A possible explanation for contradictory results in major depression are the heterogeneity of (1) the depression phenotype per se because diagnosis criteria include >400 possible symptom combinations [53,54]; and (2) the questionnaires, assessment approaches, statistical analyses, and speech feature extraction tools used in these studies. The within-person research design approach underlying our data set addressed the heterogeneity of the depression phenotype at least partially. Furthermore, we analyzed free speech collected naturally in a selfie task, whereas in the study by Mundt et al [31], read speech was used in the analyses. In line with what is suggested in the study by Quatieri and Malyska [34], this could also be a reason for the contradictory results. However, because assessing within-person fluctuations in daily life increases ecological validity, we regard our results as an important contribution.

Observing the full picture of associations, we note that the results for all 3 speech features are similar and do not provide evidence of specific associations (eg, association of 1 specific speech feature with 1 specific momentary affective state), showing no distinct patterns of momentary states for each speech feature. This is reasonable because the constructs overlap in content (eg, patients experiencing depression experience higher negative affect and lower positive affect).

In terms of the combined models evaluating the relative importance of the features, we found that in the 4 models (ADS-K, valence, positive affect, and negative affect) both *pitch variability* and *speech pauses* remained significant, whereas *speech rate* did not. *Pitch variability* remained the only significant parameter in the model of energetic arousal, and *speech pauses* remained the only significant parameter in the model of calmness. This suggests that *pitch variability* and *speech pauses* are speech features rather independent of each other, whereas the high correlation between *speech pauses* and *speech rate* might account for the fact that only 1 of these features (in this case, *speech pauses)* remained a significant predictor.

Limitations

First, this study examined a limited set of 3 speech features. Instead of applying brute force methods involving thousands of technical speech features, we selected speech features based on previous work and with high face validity, restricting the scope of our analysis. Although we did expand our scope of features in the exploratory analysis, it is very likely that other configurations and features (eg, the *ComParE feature set* containing 6373 features [55]) might also be predictive of affective states. Future work is needed to compare theory-driven approaches with brute force data-driven machine learning methods to find the best possible combination of speech features also considering aspects of computational power. However, selecting the features on a theoretical basis and restricting their pure number limits alpha error inflation and should increase replicability.

Second, although the sample size of our study was limited, this was a true within-person design with many data points per patient. In addition, we regard this study as a pilot study providing important indications regarding feasibility in a clinical context. As some patients dropped out of the study, and some recordings had to be excluded, in future studies, data collection needs to be integrated better into clinical routines. Moreover, the instructions for patients may need to be revised to reduce the likelihood of missing data and recording errors. However, the data set at hand is still unique in the relatively high number of assessments per patient and the applied SDT, which yielded meaningful variation in the depression severity within a short time period. From a theoretical perspective, it is crucial to emphasize that to uncover existing relations among variables, meaningful variance in both parameters is needed.

Third and last, selfie videos were recorded in a clinical environment, which may limit generalizability to other contexts. In future studies, ambulant patients could be integrated and other environments explored to evaluate the replicability of the results. However, our approach, which involved sampling free speech, offers higher ecological validity to reading standardized text paragraphs because it provides a closer representation of people's everyday lives. The development of passive sensing will be helpful in this context (ie, the random assessment of audio bits in an ecological environment). To date, automated passive voice recordings in nonprotected environments have been restricted in 2-party consent states, such as Germany. However, in single-party consent states, a few speech-related applications can be used *in the wild* (eg, the Electronically Activated Recorder [56]). Although the development of technical devices is ongoing, future studies will have to consider ethical issues related to voice recording in natural settings (eg, ensuring that no third parties who did not give informed consent are recorded).

Conclusions

Our study provides evidence that fluctuations in the speech features *pitch variability*, *speech pauses*, and *speech rate* are associated with fluctuations in depression severity and other momentary affect states. Notably, the data were collected from clinically diagnosed patients (no subclinical sample or staged emotions) experiencing an acute depressive episode. A particularly important advantage is that our longitudinal ambulatory assessment data set ensured a maximum of within-person dynamics of depressive parameters within a short time period by applying a sleep deprivation intervention design. This is of great importance because future technology will try to predict upcoming depressive episodes on an individual level and will need information on within-person trajectories. For the development of such tailored precision medicine tools, *pitch variability*, *speech pauses*, and *speech rate* present promising features. Our research is a step forward on the path to developing an automated depression monitoring system, facilitating individually tailored treatments and increased patient empowerment.

Acknowledgments

This study was funded by the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation; GRK2739/1, project 447089431) Research Training Group: KD²School—Designing Adaptive Systems for Economic Decisions; the Network of European Funding for Neuroscience Research (ERA-NET NEURON): “Impact of Early Life Metabolic and Psychosocial Stress on Susceptibility to Mental Disorders; From Converging Epigenetic Signatures to Novel Targets for Therapeutic Intervention (EMBED)” (01EW1904); and the Bundesministerium für Bildung und Forschung (BMBF; Federal Ministry of Education and Research)–funded e:Med program *Target-OXY* (FKZ: 031L0190A). The authors acknowledge support by the Karlsruhe Institute of Technology Publication Fund.

Authors' Contributions

MR, JCF, JF, SHW, LS, MG, YY, and FS planned the investigation and developed the sampling scheme. MG and LS were responsible for data collection. L-MW preprocessed the data, carried out statistical analysis, and interpreted the results. UWE-P and MFL contributed to the analysis and interpretation of data. TS and AA contributed to acoustic analysis. L-MW drafted the manuscript with contributions from UWE-P and LS. All authors revised and edited the manuscript critically and had final approval of the version to be published.

Conflicts of Interest

UWE-P reports consultancy for Boehringer Ingelheim and speaker honorarium from Angelini Pharma, both of which had no influence over the content of this paper. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Supplementary material.
[PDF File (Adobe PDF File), 269 KB - mental_v11i1e49222_app1.pdf]

References

1. GBD 2019 Diseases and Injuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct 17;396(10258):1204-1222 [FREE Full text] [doi: 10.1016/S0140-6736(20)30925-9] [Medline: 33069326]
2. Streit F, Zillich L, Frank J, Kleineidam L, Wagner M, Baune BT, NAKO Investigators, et al. Lifetime and current depression in the German National Cohort (NAKO). World J Biol Psychiatry 2023 Dec;24(10):865-880. [doi: 10.1080/15622975.2021.2014152] [Medline: 34870540]
3. Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med 2006 Nov;3(11):e442 [FREE Full text] [doi: 10.1371/journal.pmed.0030442] [Medline: 17132052]
4. COVID-19 Mental Disorders Collaborators. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 2021 Nov 06;398(10312):1700-1712 [FREE Full text] [doi: 10.1016/S0140-6736(21)02143-7] [Medline: 34634250]
5. American Psychiatric Association. Diagnostisches und Statistisches Manual Psychischer Störungen DSM-5. Göttingen, Germany: Hogrefe Verlag; Dec 2, 2014.
6. Sobocki P, Jönsson B, Angst J, Rehnberg C. Cost of depression in Europe. J Ment Health Policy Econ 2006 Jun;9(2):87-98. [Medline: 17007486]
7. Ben-Zeev D, Young MA. Accuracy of hospitalized depressed patients' and healthy controls' retrospective symptom reports: an experience sampling study. J Nerv Ment Dis 2010 Apr;198(4):280-285. [doi: 10.1097/NMD.0b013e3181d6141f] [Medline: 20386257]
8. Eaton WW, Neufeld K, Chen LS, Cai G. A comparison of self-report and clinical diagnostic interviews for depression: diagnostic interview schedule and schedules for clinical assessment in neuropsychiatry in the Baltimore epidemiologic catchment area follow-up. Arch Gen Psychiatry 2000 Mar;57(3):217-222. [doi: 10.1001/archpsyc.57.3.217] [Medline: 10711906]
9. Abd-Alrazaq A, AlSaad R, Aziz S, Ahmed A, Denecke K, Househ M, et al. Wearable artificial intelligence for anxiety and depression: scoping review. J Med Internet Res 2023 Jan 19;25:e42672 [FREE Full text] [doi: 10.2196/42672] [Medline: 36656625]
10. Kohn R, Saxena S, Levav I, Saraceno B. The treatment gap in mental health care. Bull World Health Organ 2004 Nov;82(11):858-866 [FREE Full text] [Medline: 15640922]
11. Williams SZ, Chung GS, Muennig PA. Undiagnosed depression: a community diagnosis. SSM Popul Health 2017 Dec;3:633-638 [FREE Full text] [doi: 10.1016/j.ssmph.2017.07.012] [Medline: 29349251]
12. De Angel V, Lewis S, White K, Oetzmann C, Leightley D, Oprea E, et al. Digital health tools for the passive monitoring of depression: a systematic review of methods. NPJ Digit Med 2022 Jan 11;5(1):3 [FREE Full text] [doi: 10.1038/s41746-021-00548-8] [Medline: 35017634]

13. Torous J, Staples P, Onnela JP. Realizing the potential of mobile mental health: new methods for new data in psychiatry. Curr Psychiatry Rep 2015 Aug;17(8):602 [FREE Full text] [doi: 10.1007/s11920-015-0602-0] [Medline: 26073363]
14. Zarate D, Stavropoulos V, Ball M, de Sena Collier G, Jacobson NC. Exploring the digital footprint of depression: a PRISMA systematic literature review of the empirical evidence. BMC Psychiatry 2022 Jun 22;22(1):421 [FREE Full text] [doi: 10.1186/s12888-022-04013-y] [Medline: 35733121]
15. Fahrenberg J, Myrtek M, Pawlik K, Perrez M. Ambulatory assessment - monitoring behavior in daily life settings. Eur J Psychol Assessment 2007 Jan;23(4):206-213. [doi: 10.1027/1015-5759.23.4.206]
16. Stone AA, Shiffman S. Ecological momentary assessment (EMA) in behavorial medicine. Ann Behav Med 1994;16(3):199-202. [doi: 10.1093/abm/16.3.199]
17. Csikszentmihalyi M, Larson R. Validity and reliability of the experience-sampling method. J Nerv Ment Dis 1987 Sep;175(9):526-536. [doi: 10.1097/00005053-198709000-00004] [Medline: 3655778]
18. Torous J, Kiang MV, Lorme J, Onnela JP. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR Ment Health 2016 May 05;3(2):e16 [FREE Full text] [doi: 10.2196/mental.5165] [Medline: 27150677]
19. Ebner-Priemer UW, Trull TJ. Ecological momentary assessment of mood disorders and mood dysregulation. Psychol Assess 2009 Dec;21(4):463-475. [doi: 10.1037/a0017075] [Medline: 19947781]
20. Ebner-Priemer U, Santangelo P. Digital phenotyping: hype or hope? Lancet Psychiatry 2020 Apr;7(4):297-299 [FREE Full text] [doi: 10.1016/s2215-0366(19)30380-3]
21. Onnela JP, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology 2016 Jun;41(7):1691-1696 [FREE Full text] [doi: 10.1038/npp.2016.7] [Medline: 26818126]
22. Trull TJ, Ebner-Priemer U. The role of ambulatory assessment in psychological science. Curr Dir Psychol Sci 2014 Dec;23(6):466-470 [FREE Full text] [doi: 10.1177/0963721414550706] [Medline: 25530686]
23. Torous J, Onnela JP, Keshavan M. New dimensions and new tools to realize the potential of RDoC: digital phenotyping via smartphones and connected devices. Transl Psychiatry 2017 Mar 07;7(3):e1053 [FREE Full text] [doi: 10.1038/tp.2017.25] [Medline: 28267146]
24. Sundberg J. Expressivity in singing. A review of some recent investigations. Logopedics Phoniatrics Vocology 2009 Jul 11;23(3):121-127. [doi: 10.1080/140154398434130]
25. Kraepelin E. Manic-Depressive Insanity and Paranoia. Sacramento, CA: Creative Media Partners, LLC; 1921.
26. France DJ, Shiavi RG, Silverman S, Silverman M, Wilkes DM. Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans Biomed Eng 2000 Jul;47(7):829-837. [doi: 10.1109/10.846676] [Medline: 10916253]
27. Hashim NW, Wilkes M, Salomon R, Meggs J, France DJ. Evaluation of voice acoustics as predictors of clinical depression scores. J Voice 2017 Mar;31(2):256.e1-256.e6. [doi: 10.1016/j.jvoice.2016.06.006] [Medline: 27473933]
28. Smith M, Dietrich BJ, Bai EW, Bockholt HJ. Vocal pattern detection of depression among older adults. Int J Ment Health Nurs 2020 Jun;29(3):440-449. [doi: 10.1111/inm.12678] [Medline: 31811697]
29. Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders using speech: a systematic review. Laryngoscope Investig Otolaryngol 2020 Feb;5(1):96-116 [FREE Full text] [doi: 10.1002/lio2.354] [Medline: 32128436]
30. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. Speech Commun 2015 Jul;71:10-49. [doi: 10.1016/j.specom.2015.03.004]
31. Mundt JC, Snyder PJ, Cannizzaro MS, Chappie K, Geralts DS. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. J Neurolinguistics 2007 Jan;20(1):50-64 [FREE Full text] [doi: 10.1016/j.jneuroling.2006.04.001] [Medline: 21253440]
32. Trevino AC, Quatieri TF, Malyska N. Phonologically-based biomarkers for major depressive disorder. EURASIP J Adv Signal Process 2011 Aug 16;2011(1):42. [doi: 10.1186/1687-6180-2011-42]
33. Mundt JC, Vogel AP, Feltner DE, Lenderking WR. Vocal acoustic biomarkers of depression severity and treatment response. Biol Psychiatry 2012 Oct 01;72(7):580-587 [FREE Full text] [doi: 10.1016/j.biopsych.2012.03.015] [Medline: 22541039]
34. Quatieri TF, Malyska N. Vocal-source biomarkers for depression: a link to psychomotor activity. In: Proceedigs of Interspeech 2012. 2012 Presented at: Interspeech 2012; September 9-13, 2012; Portland, OR. [doi: 10.21437/interspeech.2012-311]
35. Horwitz R, Quatieri TF, Helfer BS, Yu B, Williamson JR, Mundt J. On the relative importance of vocal source, system, and prosody in human depression. In: Proceedings of the IEEE International Conference on Body Sensor Networks. 2013 Presented at: IEEE International Conference on Body Sensor Networks; May 6-9, 2013; Cambridge, MA. [doi: 10.1109/BSN.2013.6575522]
36. Yang Y, Fairbairn C, Cohn JF. Detecting depression severity from vocal prosody. IEEE Trans Affect Comput 2013;4(2):142-150 [FREE Full text] [doi: 10.1109/T-AFFC.2012.38] [Medline: 26985326]
37. Wirz-Justice A, Benedetti F. Perspectives in affective disorders: clocks and sleep. Eur J Neurosci 2020 Jan;51(1):346-365. [doi: 10.1111/ejn.14362] [Medline: 30702783]
38. movisens GmbH. movisens. URL: https://movisens.com/en/ [accessed 2023-05-02]
39. Hautzinger M. Die CES-D Skala. Ein depressionsinstrument für untersuchungen in der allgemeinbevölkerung. Diagnostica 1988;34:167-173.

40. Myin-Germeys I, Peeters F, Havermans R, Nicolson NA, DeVries MW, Delespaul P, et al. Emotional reactivity to daily life stress in psychosis and affective disorder: an experience sampling study. Acta Psychiatr Scand 2003 Feb;107(2):124-131. [doi: 10.1034/j.1600-0447.2003.02025.x] [Medline: 12534438]
41. Wilhelm P, Schoebi D. Assessing mood in daily life. Eur J Psychol Assess 2007 Jan;23(4):258-267. [doi: 10.1027/1015-5759.23.4.258]
42. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry 1979 Apr;134:382-389. [doi: 10.1192/bjp.134.4.382] [Medline: 444788]
43. Schröter H, Maier AN, Escalante-B AN, Rosenkranz T. Deepfilternet2: towards real-time speech enhancement on embedded devices for full-band audio. In: Proceedings of the International Workshop on Acoustic Signal Enhancement (IWAENC). 2022 Presented at: International Workshop on Acoustic Signal Enhancement (IWAENC); September 5-8, 2022; Bamberg, Germany. [doi: 10.1109/iwaenc53105.2022]
44. Eyben F, Scherer KR, Schuller BW, Sundberg J, Andre E, Busso C, et al. The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans Affect Comput 2016 Apr 1;7(2):190-202. [doi: 10.1109/taffc.2015.2457417]
45. openSMILE Python. GitHub. URL: https://github.com/audeering/opensmile-python [accessed 2023-05-02]
46. Eyben F, Wöllmer M, Schuller B. Opensmile: the Munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on Multimedia. 2010 Presented at: MM '10; October 25-29, 2010; Firenze, Italy. [doi: 10.1145/1873951.1874246]
47. Abulimiti A, Weiner J, Schultz T. Automatic speech recognition for ILSE-interviews: longitudinal conversational speech recordings covering aging and cognitive decline. In: Proceedings of the Interspeech 2020. 2020 Presented at: Interspeech 2020; October 25-29, 2020; Shanghai, China. [doi: 10.21437/interspeech.2020-2829]
48. Snijders TA, Bosker RJ. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. Thousand Oaks, CA: SAGE Publications; 1999.
49. Eisinga R, Grotenhuis MT, Pelzer B. The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? Int J Public Health 2013 Aug;58(4):637-642. [doi: 10.1007/s00038-012-0416-3] [Medline: 23089674]
50. Hox J, van de Schoot R. Robust methods for multilevel analysis. In: Scott MA, Simonoff JF, Marx BD, editors. The SAGE Handbook of Multilevel Modeling. Thousand Oaks, CA: SAGE Publications; 2013.
51. Maas CJ, Hox JJ. Sample sizes for multilevel modeling. Computer Intensieve Statistiek. URL: https://www.joophox.net/publist/simnorm1.pdf [accessed 2023-12-20]
52. Geldhof GJ, Preacher KJ, Zyphur MJ. Reliability estimation in a multilevel confirmatory factor analysis framework. Psychol Methods 2014 Mar;19(1):72-91. [doi: 10.1037/a0032138] [Medline: 23646988]
53. Goldberg D. The heterogeneity of "major depression". World Psychiatry 2011 Oct;10(3):226-228 [FREE Full text] [doi: 10.1002/j.2051-5545.2011.tb00061.x] [Medline: 21991283]
54. Ostergaard SD, Jensen SO, Bech P. The heterogeneity of the depressive syndrome: when numbers get serious. Acta Psychiatr Scand 2011 Dec;124(6):495-496. [doi: 10.1111/j.1600-0447.2011.01744.x] [Medline: 21838736]
55. Schuller B, Steidl S, Batliner A, Vinciarelli A, Scherer K, Ringeval F, et al. The INTERSPEECH 2016 computational paralinguistics challenge: deception, sincerity and native language. In: Proceedings of the Interspeech 2016. 2016 Presented at: Interspeech '16; September 8-12, 2016; San Francisco, CA. [doi: 10.21437/interspeech.2013-56]
56. Mehl MR. The lay assessment of subclinical depression in daily life. Psychol Assess 2006 Sep;18(3):340-345. [doi: 10.1037/1040-3590.18.3.340] [Medline: 16953737]

Abbreviations

ADS-K: Allgemeine Depressionsskala
eGeMAPS: extended Geneva Minimalistic Acoustic Parameter Set
F0: fundamental frequency
ICC: intraclass correlation coefficient
ICD-10: International Classification of Diseases, Tenth Revision
MADRS: Montgomery–Åsberg Depression Rating Scale
MDMQ: Multidimensional Mood Questionnaire
MFCC: mel-frequency cepstral coefficient
openSMILE: open-source Speech and Music Interpretation by Large-Space Extraction
SDT: sleep deprivation therapy
SLEDGE II: Sleep Deprivation and Gene Expression

Edited by J Torous; submitted 22.05.23; peer-reviewed by D Zarate, O Ajilore; comments to author 03.10.23; accepted 21.10.23; published 18.01.24.

Please cite as:
Wadle LM, Ebner-Priemer UW, Foo JC, Yamamoto Y, Streit F, Witt SH, Frank J, Zillich L, Limberger MF, Ablimit A, Schultz T, Gilles M, Rietschel M, Sirignano L
Speech Features as Predictors of Momentary Depression Severity in Patients With Depressive Disorder Undergoing Sleep Deprivation Therapy: Ambulatory Assessment Pilot Study
JMIR Ment Health 2024;11:e49222
URL: https://mental.jmir.org/2024/1/e49222
doi:10.2196/49222
PMID:38236637

Original Paper

eHealth in the Management of Depressive Episodes in Catalonia's Primary Care From 2017 to 2022: Retrospective Observational Study

Aïna Fuster-Casanovas[1,2], RPh, MSc; Queralt Miró Catalina[2,3], MSc; Josep Vidal-Alaball[2,3,4], MD, MPH, PhD; Anna Escalé-Besa[2,3,4], MD, MSc; Carme Carrión[1,5,6], PhD

[1]eHealth Lab Research Group, School of Health Sciences and eHealth Centre, Universitat Oberta de Catalunya, Barcelona, Spain

[2]Unitat de Suport a la Recerca de la Catalunya Central, Fundació Institut Universitari per a la Recerca a l'Atenció Primària de Salut Jordi Gol i Gurina, Sant Fruitós de Bages, Spain

[3]Health Promotion in Rural Areas Research Group, Gerència d'Atenció Primària i a la Comunitat de Catalunya Central, Institut Català de la Salut, Sant Fruitós de Bages, Spain

[4]Faculty of Medicine, University of Vic-Central, University of Catalonia, Vic, Spain

[5]Network for Research on Chronicity, Primary Care, and Health Promotion, Barcelona, Spain

[6]School of Medicine, Universitat de Girona, Girona, Spain

Corresponding Author:
Josep Vidal-Alaball, MD, MPH, PhD
Unitat de Suport a la Recerca de la Catalunya Central
Fundació Institut Universitari per a la Recerca a l'Atenció Primària de Salut Jordi Gol i Gurina
Carrer Pica d'Estats, 13-15
Sant Fruitós de Bages, 08272
Spain
Phone: 34 936 93 00 40
Email: jvidal.cc.ics@gencat.cat

Abstract

Background: The reasons for mental health consultations are becoming increasingly relevant in primary care. The Catalan health care system is undergoing a process of digital transformation, where eHealth is becoming increasingly relevant in routine clinical practice.

Objective: This study aimed to analyze the approach to depressive episodes and the role of eHealth in the Catalan health care system from 2017 to 2022.

Methods: A retrospective observational study was conducted on diagnostic codes related to depressive episodes and mood disorders between 2017 and 2022 using data from the Catalan Institute of Health. The sociodemographic evolution and prevalence of depression and mood disorders in Catalonia were analyzed between 2017 and 2022. Sociodemographic variables were analyzed using absolute frequency and percentage. The prevalence of depressive episodes was calculated, highlighting the year-to-year changes. The use of eHealth for related consultations was assessed by comparing the percentages of eHealth and face-to-face consultations. A comparison of sociodemographic variables based on attendance type was conducted. Additionally, a logistic regression model was used to explore factors influencing face-to-face attendance. The analysis used R software (version 4.2.1), with all differences examined using 95% CIs.

Results: From 2017 to 2022, there was an 86.6% increase in the prevalence of depression and mood disorders, with women consistently more affected (20,950/31,197, 67.2% in 2017 and 22,078/33,169, 66.6% in 2022). In 2022, a significant rise in depression diagnoses was observed in rural areas (difference 0.71%, 95% CI 0.04%-1.43%), contrasting with a significant decrease in urban settings (difference –0.7%, 95% CI –1.35% to –0.05%). There was a significant increase in antidepressant use in 2022 compared to 2017 (difference 2.4%, 95% CI 1.87%-3.06%) and the proportion of eHealth visits rose from 4.34% (1240/28,561) in 2017 to 26.3% (8501/32,267) in 2022. Logistic regression analysis indicated that men (odds ratio [OR] 1.06, 95% CI 1.04-1.09) and younger individuals had a higher likelihood of eHealth consultations in 2022. Furthermore, individuals using eHealth consultations were more likely to use antidepressants (OR 1.54, 95% CI 1.50-1.57) and anxiolytics (OR 1.06, 95% CI 1.03-1.09).

Conclusions: The prevalence of depression in Catalonia has significantly increased in the last 6 years, likely influenced by the COVID-19 pandemic. Despite ongoing digital transformation since 2011, eHealth usage remained limited as of 2017. During the lockdown period, eHealth accounted for nearly half of all health care consultations, representing a quarter of consultations by 2022. In the immediate aftermath of the COVID-19 pandemic, emerging evidence suggests a significant role of eHealth in managing depression-related consultations, along with an apparent likelihood of patients being prescribed antidepressants and anxiolytics. Further research is needed to understand the long-term impact of eHealth on diagnostic practices and medication use.

(JMIR Ment Health 2024;11:e52816) doi:10.2196/52816

KEYWORDS

eHealth; depression; depressive disorder; primary health care; mental health patient; patient; patients; healthcare system; digital transformation; mental disorder; mental disorders; diagnostic; clinical practice; clinical practices; retrospective; observational; regression; digital tool; digital tools

Introduction

Depression is a mental disorder that is a major health problem due to its high prevalence, direct repercussions on people's quality of life, and strong social impact [1,2]. According to the World Health Organization, it is estimated that 5% of the world's population suffers from depression [3].

Catalonia, an autonomous community of Spain, had a total of 7,747,709 citizens at the beginning of 2022 [4]. The Catalan health care system, characterized by being public, universal, comprehensive, and equitable, provides health coverage to 7.6 million people [5]. At the level of the Catalan territory, in 2015, according to data from the Catalonia health survey, 16.6% of the population older than 15 years suffered from anxiety or depression, which affected more women (20.8%) than men (12.2%) [6]. In 2020, 5 years later and taking into account the outbreak of the SARS-CoV-2 pandemic, an increase in emotional distress and moderate and major depression was observed. Specifically, 27.19% of women and 17% of men suffered from emotional distress, and 12.2% of women and 5.7% of men had moderate or major depression in Catalonia. Therefore, a decrease in the emotional well-being of the Catalan population has been observed [7].

Primary care (PC), one of the gateways to the health care system, is key in the detection, management, and follow-up of illnesses such as depression. High-quality PC is the foundation of a leading health care system and fundamental to optimizing the performance of the health care system, fulfilling the following 5 dimensions of the quintuple aim: (1) enhancing the care experience, (2) improving population health, (3) reducing costs, (4) care team well-being, and (5) advancing health equity [8-10]. PC has, among other characteristics, longitudinality and universal accessibility that make it the only area designed to be used by all people throughout their lives. As early as 1994, Starfield [11] showed how people living in countries with high-quality PC had better health indices, a more equitable distribution of available resources, and a more efficient health care system.

The reasons for mental health consultations are becoming increasingly relevant in PC, especially among the younger population in the wake of the COVID-19 pandemic [12,13]. Since there is no standardized protocol for mental health care in Catalonia, the approach to these issues varies significantly. This variability could be attributed to differences in professionals' training or their sensitivity toward handling such problems [14]. Based on the report by the Public Health Agency of Catalonia regarding drug consumption with potential misuse, it was found that antidepressants were the most commonly used group of drugs by the population in 2021. Among most antidepressant drugs, the consumption ratio by women is between 2 to 4 times higher than that by men [15].

Over the last decade, information and communication technology is being introduced into the health care systems of different industrialized countries [16]. Following this trend, the Catalan health care system has been immersed in a digital transformation process. Currently, eHealth within PC in Catalonia involves communication both between providers and patients and amongst providers themselves. These interactions occur via different modes, including telephone or video calls and electronic consultations. It is important to emphasize that although the Catalan health care system had already incorporated technology before, the pandemic served as a clear catalyst for the adoption of eHealth in Catalonia. Regarding the typology of eHealth visits, the majority are conducted via remote consultations (referred to as eConsulta in Catalonia) or phone calls, with video calls being relatively infrequent [17].

Since 2011, several cross-cutting projects of significant importance have been developed in Catalonia, incorporating digital health as a key component [18]. The design of the Catalonia shared medical record began with the aim of being able to share patient information among the different health providers, as well as a personal folder (La Meva Salut) to give citizens access to their personal information, such as current medication plan, diagnoses, vaccines administered, clinical reports, test results, and examinations. Also, within this personal folder, in 2015, eConsulta was developed as an asynchronous digital communication tool involving health care professionals and patients, allowing the population to send queries at any time to their PC doctor or nurse and receive a response within a maximum of 48 hours on working days [19,20]. In this sense, eConsulta represents a significant advancement in eHealth within the Catalan public health system. Prior to the pandemic, eConsulta usage had a monthly growth rate of 7%; from March 15 to May 2020, an exponential growth rate was observed [21]. A study focusing on the profile of health care professionals who utilized remote consultations like eConsulta before the pandemic revealed that physicians engaging with eConsulta were typically aged between 45 and 64 years, exceeded the 80th percentile in the Quality of Care Index, had a high level of accessibility to their patients, participated in educational activities, and operated within a health team framework in urban settings with a high socioeconomic status [22].

In the context of Catalan PC, eHealth has been widely implemented and adopted in routine clinical practice, even more so in the context of COVID-19 [23-25]. However, there are other digital health tools with great potential, such as mobile health or artificial intelligence, that have not yet been introduced into the public health care system in Catalonia.

In order to improve the management of depressive episodes in PC in the Catalan health care system and to obtain higher effectiveness rates close to the potential efficacy of the available treatments, the aim of this study was to analyze the evolution of the prevalence of depression and mood disorders in PC from 2017 to 2022, to examine the sociodemographic profile of the affected population, and to investigate the role played by eHealth in this context and assess its impact on consultations related to depression and mood disorders.

Methods

Study Design

This study was a retrospective observational study of diagnostic codes related to depressive episodes and mood disorders between 2017 to 2022 from the Catalan Institute of Health.

Sample

We analyzed the entire population of Catalonia that visited PC centers of the Catalan Institute of Health who had a face-to-face or eHealth consultations (eConsulta, telephone, or video consultations) associated with the selected diagnoses of depression and mood disorders in the period of January 2017 to December 2022. The Catalan Institute of Health is the main provider of health services in the public system. It manages approximately 80% of the PC teams in Catalonia and provides health care coverage to approximately 5.8 million people [16,21,26,27].

To obtain the study sample, the diagnostic codes were selected according to those typified in the International Statistical Classification of Diseases and Related Health Problems (ICD-10), which is the classification used by PC professionals when recording a diagnosis in the computerized PC clinic [28]. All diagnostic codes related to depression (mild, moderate, and major) and mood disorders (eg, cyclothymia and dysthymia) that are usually detected, managed, or followed up in Catalan PC were selected (Table S1 in Multimedia Appendix 1). Mood disorders are marked by persistent clinical manifestations. In the case of dysthymia, it necessitates at least 2 years of a consistently low mood, whereas cyclothymia involves oscillations between a depressed mood and euphoria, without fulfilling the criteria for major depression [29]. Therefore, these diagnoses were included because they are not reactive disorders of short duration. By including them, there is a reduced risk of missing cases of mild depressive disorders.

Finally, we excluded all consultations that did not have an associated diagnosis code among those selected. The part of the population of Catalonia whose public health coverage was not provided by the Catalan Institute of Health was also excluded.

The database was obtained through the Information System for the Development of Research in Primary Care [30].

Variables

To study sociodemographic characteristics, we considered age, sex, drug use, recurrent depressive episode diagnoses, rurality, and the socioeconomic situation recorded through the MedeA index [31]. A variable was created to distinguish between individuals with recurrent depressive episode diagnoses (where depression had been diagnosed previously in their medical history) and those without such recurrent diagnoses. The MedeA is a deprivation index linked to each census tract of the population. This assessment focuses on the barriers to access employment, education, culture, and social development. The aim is to evaluate them at a level that is considered acceptable within the surrounding society or region. The assessment comprises subindicators for employment and education. It is only available for urban areas, which are defined as municipalities with more than 10,000 inhabitants and a population density of more than 150 inhabitants per km^2. Other areas were considered rural. The MedeA index is ranked in quintiles, from MedeA urban 1 indicating low deprivation to MedeA urban 5 indicating high deprivation [31]. The MedeA index was used to categorize rurality by grouping urban and rural groups.

Statistical Analysis

To observe the evolution of the profile of people diagnosed with depression and mood disorders, sociodemographic variables were described in 2017 and 2022. Categorical variables were described by absolute frequency and percentage. The difference between the years was calculated with percentage points.

To observe the evolution of the prevalence of depressive episodes and mood disorders over the years for the total sample, the prevalent cases for each year were divided with respect to the total population assigned to the Catalan Institute of Health PC centers throughout Catalonia. The percentage change was calculated to determine the year-to-year evolution of these prevalences.

To observe the use of eHealth in consultations related to depression and mood disorders over the years, the percentage of telematic and face-to-face consultations resulting in new diagnoses per year was calculated.

A comparison of the sociodemographic variables as a function of face-to-face attendance was performed. Categorical variables were described by absolute frequency and percentage. The difference was calculated as percentage points.

Lastly, a logistic regression model was applied to observe how the studied variables affected face-to-face attendance. The model incorporated the variables of sex, age, rurality, recurrence, antidepressants, and anxiolytics. The MedeA variable was not introduced into this model since the rurality variable was used.

Analyses were performed with R version 4.2.1 (R Foundation for Statistical Computing). All differences were examined using CIs, and a confidence level of 95% was established.

Ethics Approval

The study protocol was approved by the University Institute for Primary Care Research (IDIAP) Jordi Gol Health Care Ethics Committee (Code 23/013-P).

Results

Description of Sociodemographic Profiles of New Cases From 2017 to 2022

An analysis of the sociodemographic profile of new cases during 2017 (n=31,197) and 2022 (n=33,169) was performed to observe changes in the characteristics of this population group during the study period (Table 1).

Table 1. Comparison of demographic characteristics of new cases in 2017 and 2022.

	2017 (n=31,197), n (%)	2022 (n=33,169), n (%)	Absolute difference (%) from 2017 to 2022 (95% CI)
Age (years)			
0-15	689 (2.21)	922 (2.78)	0.57 (0.33 to 0.81)[a]
16-24	1437 (4.61)	2415 (7.28)	2.67 (2.31 to 3.04)[a]
25-34	2727 (8.74)	3712 (11.2)	2.46 (1.98 to 2.91)[a]
35-44	4924 (15.8)	4765 (14.4)	–1.4 (–1.97 to –0.86)
45-54	5653 (18.1)	6101 (18.4)	0.3 (–0.32 to 0.87)
55-64	5355 (17.2)	5754 (17.3)	0.1 (–0.41 to 0.76)
65-74	4094 (13.1)	3860 (11.6)	–1.5 (–1.99 to –0.87)[a]
75-84	4171 (13.4)	3670 (11.1)	–2.3 (–2.81 to –1.79)[a]
≥85	2147 (6.88)	1970 (5.94)	–0.94 (–1.32 to –0.56)[a]
Gender			
Women	20,950 (67.2)	22,078 (66.6)	–0.6 (–1.30 to 0.15)
Men	10,222 (32.8)	11,057 (33.4)	—[b]
MedeA index			
Rural	7312 (25.6)	8499 (26.33)	0.71 (0.04 to 1.43)[a]
Urban 1	6036 (21.1)	6592 (20.4)	–0.7 (–1.35 to –0.05)[a]
Urban 2	4269 (14.94)	4806 (14.9)	0 (–0.62 to 0.51)
Urban 3	5644 (19.76)	6419 (19.9)	0.1 (–0.50 to 0.77)
Urban 4	5300 (18.55)	5951 (18.4)	–0.2 (–0.73 to 0.50)
Recurrent			
No	29,144 (93.4)	29,967 (90.3)	–3.07 (–3.49 to –2.65)[a]
Yes	2053 (6.58)	3202 (9.65)	3.07 (2.65 to 3.49)[a]
Face-to-face			
eHealth	1240 (4.34)	8501 (26.3)	22 (21.46 to 22.54)[a]
Face-to-face	27,321 (95.7)	23,766 (73.7)	—
Rurality			
Rural	7312 (25.6)	8499 (26.3)	0.7 (0.04 to 1.43)[a]
Urban	21,249 (74.4)	23,768 (73.7)	—
Antidepressants			
Yes	25,353 (81.3)	27,777 (83.7)	2.4 (1.87 to 3.06)[a]
No	5844 (18.7)	5392 (16.3)	—
Anxiolytics			
Yes	15,615 (50.1)	16,519 (49.8)	–0.3 (–1.02 to 052)
No	15,582 (49.9)	16,650 (50.2)	—

[a]Statistically significant difference.

[b]Given the symmetry of the binary variables, the difference and 95% CI have only been expressed for one of the categories.

The results showed that in both 2017 and 2022, the 45-54 years age range had the most diagnoses related to depressive episodes and mood disorders. However, in 2022 there was a significant increase compared to 2017 in diagnoses in the younger age ranges (0-15 years: difference 0.57%, 95% CI 0.33%-0.81%; 16-24 years: difference 2.67%, 95% CI 2.31%-3.04%; 25-34 years: difference 2.46%, 95% CI 1.98%-2.91%). A significant reduction in diagnoses was also observed in the older age ranges

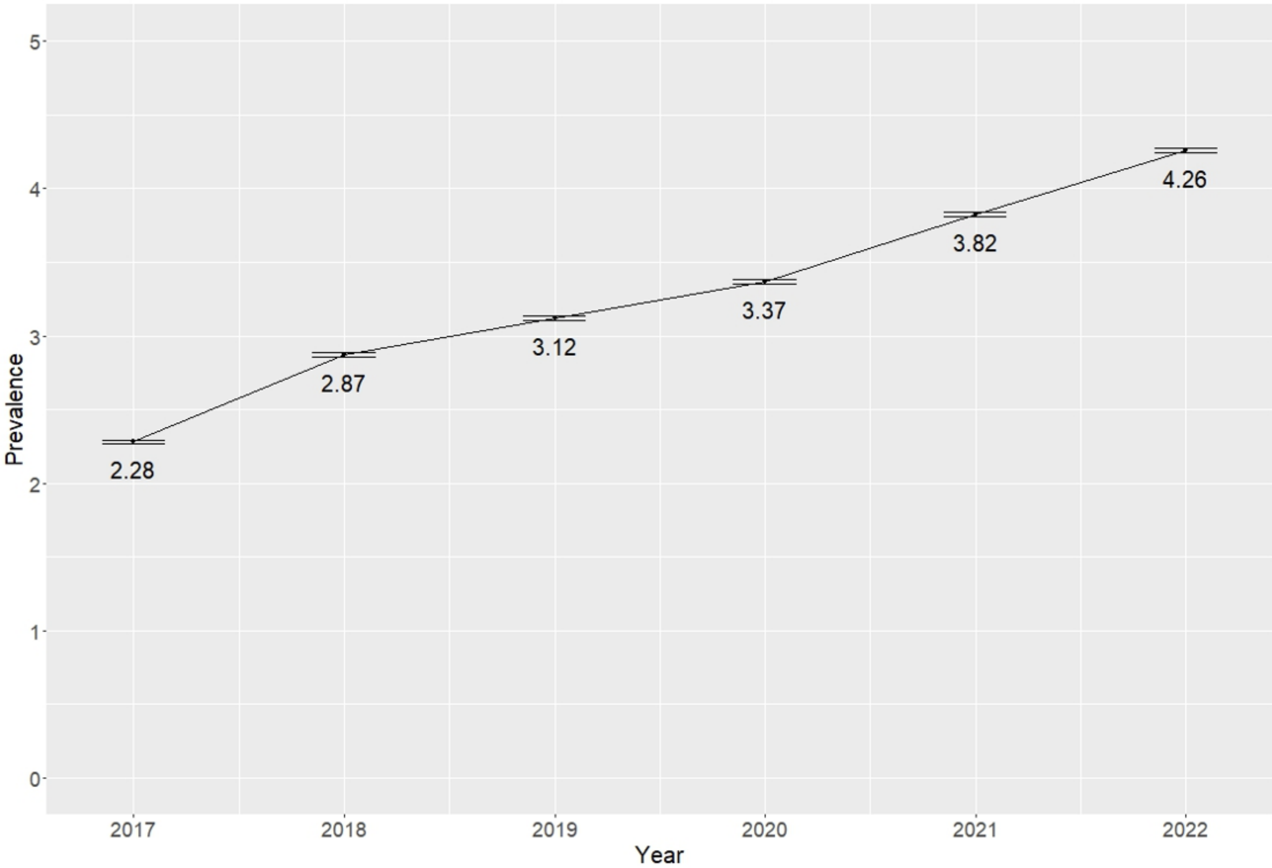
(65-74 years: difference –1.5%, 95% CI –1.99% to –0.87%; 75-84 years: difference –2.3%, 95% CI –2.81 to –1.79) and among those older than 85 years (difference –0.94%, 95% CI –1.32% to –0.56%).

Regarding gender, it was observed that in both 2017 and 2022, women were the most affected gender (20,950/31,172, 67.2% and 22,078/33,135, 66.6%, respectively). Regarding the MedeA index, a significant increase in diagnoses was observed in rural settings in 2022 with respect to 2017 (difference 0.71%, 95% CI 0.04%-1.43%), while a significant decrease was observed in urban settings, specifically in the urban population with low deprivation (difference –0.7%, 95% CI –1.35% to –0.05%). Regarding recurrent depressive episodes, there was a pronounced and significant increase in those diagnosed as recurrent compared to those diagnosed with a depressive episode for the first time (difference 3.07%, 95% CI 2.65%-3.49%). Regarding the use of eHealth for managing these illnesses, a notable shift was observed in the pattern of face-to-face attendance for the management of depressive episodes and mood disorders. A significant increase was observed in the use of eHealth as the approach for these diagnostic codes (difference 22%, 95% CI 21.46%-22.54%). Finally, regarding the most common medication used for depressive episodes and mood disorders, a significant increase in the use of antidepressants was observed in 2022 compared to 2017 (difference 2.4%, 95% CI 1.87%-3.06%), while no differences were observed in the use of anxiolytics.

Evolution of the Prevalence of Depressive Episodes and Mood Disorders by Year

The prevalence per year of all selected diagnostic codes was calculated during the period from 2017 to 2022, considering the total population assigned to the Catalan Institute of Health Primary Care Centers throughout Catalonia (Figure 1). Since 2017, an increase in diagnoses was observed. The steepest increase was in 2018, which went from 2.3% (95% CI 2.27%-2.29%) during 2017 to 2.9% (95% CI 2.86%-2.88%) in 2018. From 2019 to 2022, a progressive increase in prevalence was observed, reaching 4.3% (95% CI 4.24%-4.27%) in 2022. In short, the prevalence of depressive episodes and mood disorders increased by 86.6% during the study period. The results can be seen in Table S2 in Multimedia Appendix 1.

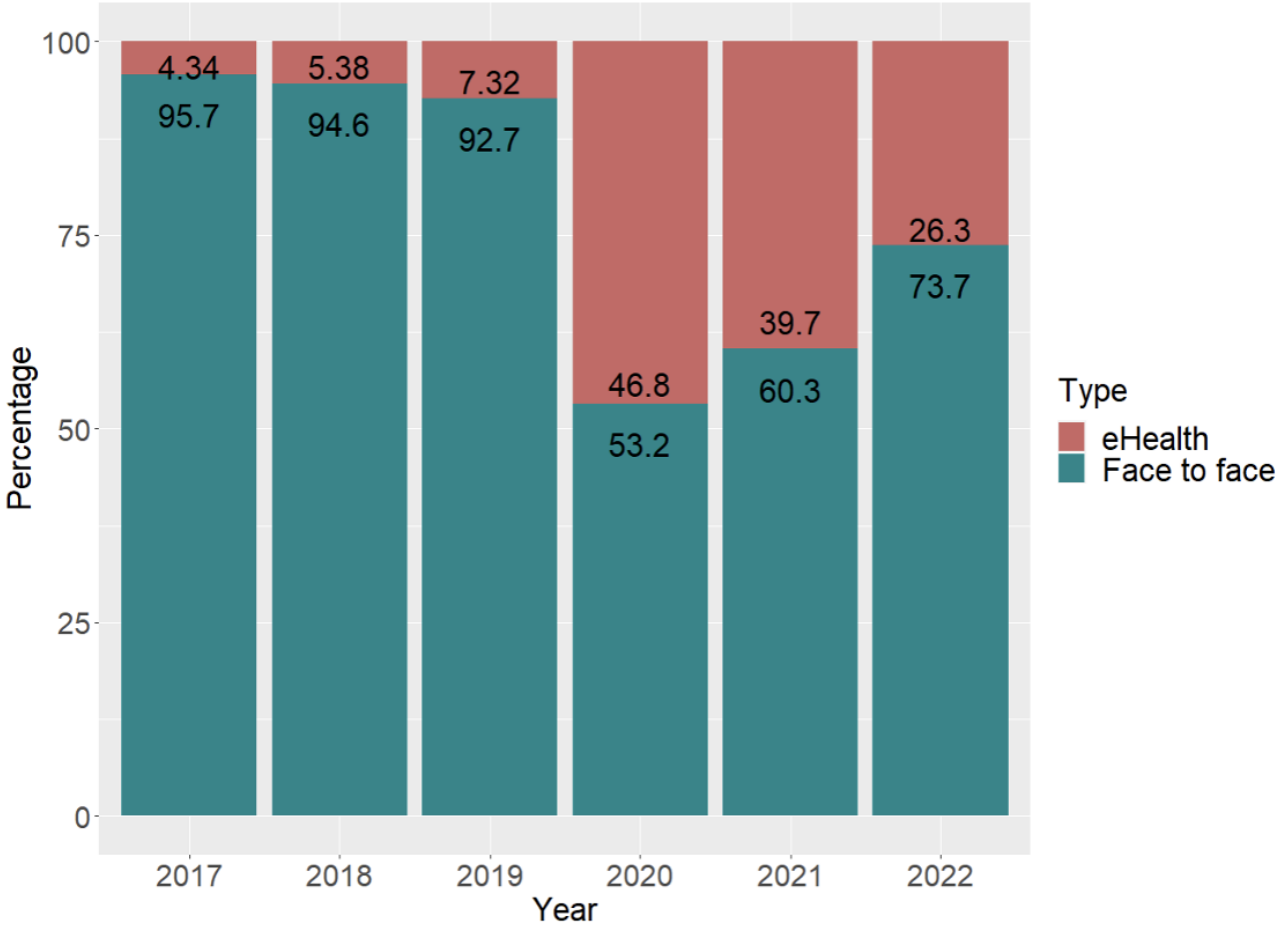
Figure 1. Depression prevalence (%) with 95% CIs between years.



Use of eHealth for Consultations Related to Depression and Mood Disorders by Year

The percentage of face-to-face and eHealth visits for each year of new diagnoses was calculated to analyze the use of eHealth for consultations related to depression and mood disorders in the public health care system (Figure 2). It was noted that in 2017 eHealth consultations represented 4.34% (1240/28,561) of all depressive episode and mood disorder visits, with an increase during 2018 and 2019. As of 2020, coinciding with the pandemic lockdown, eHealth consultations accounted for 46.8% (11,122/23,790), representing a more than 6-fold increase over the previous year. In 2021, there was some recovery of face-to-face consultations, but nevertheless, eHealth consultations still accounted for 39.7% (12,119/30,561) of total consultations. Finally, in 2022, eHealth consultations accounted for 26.3% (8501/32,267), representing just over a quarter of all consultations for diagnoses of depression and mood disorders.

Figure 2. Percentage of telematic and face-to-face consultations from 2017 to 2022.



Comparison of Prevalent Cases in 2022 According to Face-to-Face Attendance

A comparative analysis of the sociodemographic variables of the prevalent 2022 cases was performed according to face-to-face consultations (Table 2). Regarding age, it was observed that the populations that used eHealth consultations the most were those aged 0 to 15 years (difference –1.18%, 95% CI –1.36% to –0.98%), 16 to 24 years (difference –0.64%, 95% CI –0.89% to 0.38%), and those aged 85 years and older (difference –4.37%, 95% CI –4.68% to –4.05). In contrast, in 2022, it was observed that the populations that used face-to-face consultations the most were people aged 35 to 44 years (difference 0.7%, 95% CI 0.33%-1.10%), 45 to 54 years (difference 0.7%, 95% CI 0.27%-1.11%), 55 to 64 years (difference 0.8%, 95% CI 0.43%-1.24%), 65 to 74 years (difference 3.1%, 95% CI 2.84%-3.53%), and 75 to 84 years (difference 0.59%, 95% CI 0.59%-1.31%).

Significant differences were also observed in the MedeA index. People living in urban settings with less deprivation (urban 1) made more use of eHealth (difference –1.9%, 95% CI –2.35% to –1.44%), while those with higher deprivation in urban settings (urban 3 and urban 4) used face-to-face consultations more often (difference 0.9%, 95% CI 0.48%-1.34% and difference 2.4%, 95% CI 2.01%-2.84%, respectively). Regarding recurrence, it was observed that individuals with recurrent depressive episodes sought consultation more frequently via eHealth (difference –2.46%, 95% CI –2.77% to –2.15%). In terms of rurality, individuals living in rural areas demonstrated a higher frequency of eHealth usage (difference –1.4%, 95% CI –1.94% to –0.97%). Finally, individuals receiving treatment with antidepressants or anxiolytics showed a higher tendency to use eHealth services for their health consultations (difference –10.1%, 95% CI –10.59% to –9.50% and difference –4%, 95% CI –4.53% to –3.51%, respectively).

Table 2. Comparative table of prevalent cases in 2022 according to attendance type.

	Non–face-to-face (n=38,719), n (%)	Face-to-face (n=188,598), n (%)	Absolute difference (%) between attendance types (95% CI)
Age (years)			
0-15	1246 (3.22)	3850 (2.04)	–1.18 (–1.36 to –0.98)[a]
16-24	2219 (5.73)	9599 (5.09)	–0.64 (–0.89 to –0.38)[a]
25-34	3615 (9.34)	17,240 (9.14)	–0.2 (–0.51 to 0.12)
35-44	5616 (14.5)	28,711 (15.2)	0.7 (0.33 to 1.10)[a]
45-54	6859 (17.7)	34,716 (18.4)	0.7 (0.27 to 1.11)[a]
55-64	6388 (16.5)	32,688 (17.3)	0.8 (0.43 to 1.24)[a]
65-74	4086 (10.6)	25,913 (13.7)	3.1 (2.84 to 3.53)[a]
75-84	4742 (12.2)	24,894 (13.2)	1 (0.59 to 1.31)[a]
≥85	3948 (10.2)	10,987 (5.83)	–4.37 (–4.69 to –4.05)[a]
Gender			
Women	25,802 (66.7)	126,652 (67.2)	0.5 (0.00 to 1.03)
Men	12,886 (33.3)	61,832 (32.8)	—[b]
MedeA index			
Rural	10,418 (26.9)	48,003 (25.5)	–1.4 (–1.94 to –0.97)[a]
Urban 1	8687 (22.4)	38,726 (20.5)	–1.9 (–2.35 to –1.44)[a]
Urban 2	5838 (15.1)	28,469 (15.1)	0 (–0.37 to 0.41)
Urban 3	7363 (19)	37,588 (19.9)	0.9 (0.48 to 1.34)[a]
Urban 4	6413 (16.6)	35,812 (19)	2.4 (2.01 to 2.84)[a]
Recurrent			
No	35,209 (90.9)	176,141 (93.39)	—
Yes	3510 (9.07)	12,457 (6.61)	–2.46 (–2.77 to –2.15)[a]
Rurality			
Rural	10,418 (26.9)	48,003 (25.5)	–1.4 (–1.94 to –0.97)[a]
Urban	28,301 (73.1)	140,595 (74.5)	—
Antidepressants			
Yes	20,464 (52.9)	80,734 (42.8)	–10.1 (–10.59 to –9.50)[a]
No	18,255 (47.1)	107,864 (57.2)	—
Anxiolytics			
Yes	12,819 (33.1)	54,855 (29.1)	–4 (–4.53 to –3.51)[a]
No	25,900 (66.9)	133,743 (70.9)	—

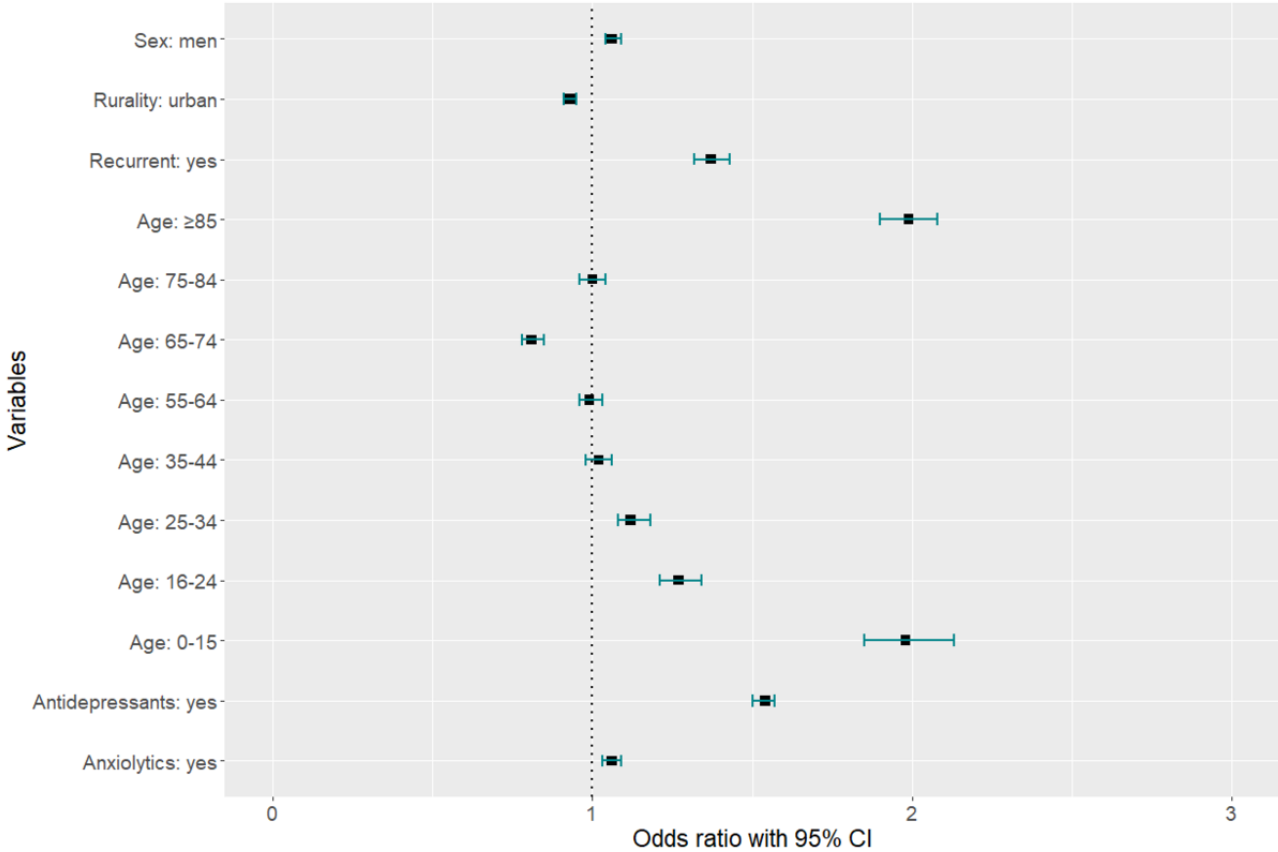
[a]Statistically significant differences.

[b]Given the symmetry of the binary variables, the difference and CI have only been expressed for one of the categories.

Finally, a logistic regression model was used to observe the impact of the studied variables from the 2022 attendance data (Figure 3). Men were 1.06 times more likely (95% CI 1.04-1.09) to use eHealth in consultations than women, and those suffering from recurrent depressive episodes were 1.37 times more likely (95% CI 1.32-1.43) to use eHealth in consultations. It was also observed that the youngest age ranges (0-15 years: odds ratio [OR] 1.98, 95% CI 1.85-2.13; 16-24 years: OR 1.27, 95% CI 1.21-1.34; 25-34 years: OR 1.12, 95% CI 1.08-1.18) were more likely to use eHealth in consultations, as was also the case among those aged 85 and older (OR 1.99, 95% CI 1.90-2.08). In contrast, people living in urban settings were 7% less likely (OR 0.93, 95% CI 0.91-0.95) to have an eHealth consultation. Finally, people who consulted more through eHealth were 1.54

times (95% CI 1.50-1.57) more likely to take an antidepressant and 1.06 times (95% CI 1.03-1.09) more likely to take an anxiolytic. The results can be seen in Table S3 in Multimedia Appendix 1.

Figure 3. Logistic regression results. The reference categories were as follows: women, aged 45-54, nonrecurrent, rural, no antidepressants, and no anxiolytics. The black squares indicate the odds ratio estimate, and the horizontal lines indicate the 95% CI.



Discussion

Principal Findings

The aim of the present study was to analyze the context of depression within the Catalan health care system. Consequently, the prevalence of depression and mood disorders has been successfully analyzed over the years, along with the sociodemographic profile of this population. Additionally, the role played by eHealth and its impact on PC for these conditions has been studied.

Regarding the prevalence of depression and mood disorders in Catalan PC, an increase of 86.6% was observed during the study period. One possible reason explaining this increase in prevalence could be an uptick in the detection of these cases within the public health care system, possibly indicating increased awareness among both professionals and citizens. In 2006 in Catalonia, the Primary Care Support Programme was implemented to coordinate PC with mental health services, providing a budget to hire more psychiatrists and psychologists in order to reinforce PC [32,33]. However, it was not until 2017 that implementation was completed in all PC centers in Catalonia. In this context, despite the existence of critical reports regarding the operation of this support program, its initiation might have positively influenced professionals' awareness, potentially contributing to an increase in detections in PC since then. The results of the study revealed a more pronounced increase (25.82%) in 2018, coinciding with the end of this program's operation.

Another possible reason explaining the increase in prevalence is the COVID-19 pandemic. The data analyzed cover the period from before the pandemic to the early postpandemic period, so the results may indeed be influenced by the consequences of lockdown and the isolation of individuals. In this context, the pandemic has generated great concern and an even more marked increase in mental health awareness, especially among the younger population. As the context of social isolation has strongly impacted the mental health of young people, as it is a crucial period for the development of social skills, it is noteworthy that from the prepandemic context in 2019, the overall prevalence increased by 36.52% in 2022 [34,35]. However, in line with the statistics already published, it is worth mentioning that women are the most affected group [6,7]. Over the past decade, the importance of sex- and gender-stratified biomedical research has grown significantly, as demonstrated by the observations of Oertelt-Prigione et al [36] of an increasing number of publications in medicine addressing this issue. It is now crucial to incorporate this perspective into the clinical approach to depression. Understanding how individuals' behaviors align with social expectations, coping strategies, and help-seeking tendencies can contribute to a more comprehensive approach to treatment. Regarding the medicalization of these conditions, it has been observed that approximately 80% of new diagnoses of depression and mood disorders in both 2017 and 2022 were prescribed antidepressant drugs. That is, most people who were diagnosed for the first time were associated with antidepressants. In contrast, the prevalence of antidepressant use has been found to be around 40% to 50% in the entire prevalent population. The results suggest that although most people who are diagnosed with depression or mood disorders are prescribed medication, in most cases treatment is withdrawn, as indicated by clinical practice guidelines [37,38].

Regarding the use of eHealth for consultations on depression and mood disorders in the PC context in Catalonia, it has been observed that the use of eHealth is increasingly being integrated into daily clinical practices in Catalonia [18]. In fact, while in 2017 diagnoses made through eHealth consultations accounted for only 4.34% of consultations, in 2022 they accounted for 26.34%. Without a doubt, the pandemic has also been a clear catalyst in the use of eHealth. Although the volume of eHealth consultations in 2020 was not equivalent to that in 2022, the results from the latter year may suggest a shift in the pattern of consultations related to depression and mood disorders. In this sense, eHealth is gaining prominence in routine practice, as has been suggested by other studies conducted in Catalonia [17,39,40]. Nevertheless, the incorporation of eHealth into the health care system must be accompanied by training for the professionals who must use it. In other words, it is not just about providing infrastructure and resources for technology operation. It is also crucial to understand when the use of these tools aligns with the care process and equip the health care professional to make clinical and therapeutic decisions appropriately, as they hold responsibility for the procedure and its outcomes [41]. The results of the present study have also shown that one of the impacts of eHealth on depressive episodes and mood disorders is on the prescription of drugs. People treated with eHealth were prescribed 54% more antidepressant drugs and 6% more anxiolytics. A systematic review by Han et al [42] observed the impact of eHealth on antibiotic prescribing in PC. Although it was not possible to conclude with sufficient confidence that eHealth significantly impacts the prescribing of these drugs, 4 of the 12 studies analyzed reported higher prescribing rates through eHealth [42]. In a study conducted by Wabe et al [43] during the pandemic in Australia's PC, they compared the prescription of medication through face-to-face consultation versus eHealth methods. The findings revealed that prescribing through face-to-face consultations was more prevalent for all drug groups classified in the World Health Organization Anatomical Therapeutic Chemical Classification list [44], except for the group that includes nervous system drugs, such as antidepressants and anxiolytics. In the context of PC in Catalonia, it would be interesting to examine whether the results obtained related to the increased prescription of drugs via eHealth can be applied to other medical conditions. In this regard, the development strategies used in the approach to eHealth will be crucial to ensure consistency in clinical practice.

Strengths and Limitations

The most relevant strength of this study is the size of the sample, since it was possible to work with the consultations related to depressive episodes and mood disorders of almost 5.8 million people. As a result, this study has yielded robust and realistic findings regarding the prevalence of depression, the use of eHealth, and the associated implications. Another strength of this study lies in the comprehensive analysis of the role and implications of eHealth in depression-related consultations. It not only highlights the current utilization of eHealth in depression consultations but also underscores the need to develop strategies that promote consistency between in-person and eHealth visits.

The main limitation of this analysis is that the data were collected only from the Catalan Institute of Health within Catalonia's public health care system. As a result, there is a lack of information from individuals who have sought treatment through private health care providers. Another limitation is that, since this was a cross-sectional study, it was not possible to follow up on each individual user, and therefore, it was not possible to establish a causal relationship between the variables studied, although the relationship between them was observed. However, due to the absence of subdivision in the original database regarding telematic consultations, specifically whether they were synchronous or asynchronous, a detailed breakdown of the proportions of telematic consultations could not be provided. Furthermore, it is essential to acknowledge the limitation that the data analyzed may reflect the impact of the lockdown imposed during de COVID-19 pandemic. The pandemic led to the isolation of individuals, which, in turn, resulted in a notable increase in diagnoses through eHealth and prescriptions of antidepressants and anxiolytics when comparing the period before COVID-19 to the postlockdown situation. In this context, the often-mandatory use of eHealth consultations during the lockdown, along with the isolation of the population, may have contributed to the increase in the detection rates and medication usage reflected in the results. Therefore, it would be valuable to replicate this study in the coming years to reassess the prevalence and sociodemographic characteristics of depression as well as the role of eHealth in addressing the observed diagnostic and pharmacological trends. This follow-up study would enable us to determine whether the findings observed in this study represent short-term consequences of COVID-19 or if they have more lasting implications over time.

Conclusions

The prevalence of depression in Catalonia has significantly increased from 2017 to 2022, with the pandemic likely having a profound impact. Although the Catalan health care system has been undergoing digital transformation since 2011, eHealth usage was limited in 2017. During the lockdown, it accounted for nearly half of the care, and by 2022 it represented a quarter of the consultations. In the short post-COVID-19 era, evidence suggests that eHealth plays a role in consultations related to depression and appears to increase the likelihood of taking antidepressants and anxiolytics. However, future studies in a context further removed from the pandemic could provide a more accurate indication of the role of eHealth in diagnosing and treating depression and mood disorders.

Acknowledgments

We would like to thank Manuel Moreno for extracting the study variables from the original database. This project was supported by the Department of Health of the Generalitat de Catalunya and received funding in the 2021 grant call from the Strategic Plan for Research and Innovation in Health (PERIS) for the 2022-2024 period. The project falls under the research projects focused on primary care, with the file code SLT021/21/000002.

Data Availability

The data sets generated and analyzed during the current study are not publicly available because the manuscript was based on confidential and sensitive health data but are available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Selected diagnostic codes, depression prevalence, and logistic regression.
[DOCX File , 27 KB - mental_v11i1e52816_app1.docx]

References

1. Depression. World Health Organization. 2021. URL: https://www.who.int/health-topics/depression#tab=tab_3 [accessed 2022-11-13]
2. Depressió. Generalitat de Catalunya. 2022. URL: https://canalsalut.gencat.cat/ca/salut-a-z/d/depressio/ [accessed 2022-11-03]
3. Depressive disorder (depression). World Health Organization. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/depression [accessed 2022-11-03]
4. Població. Generalitat de Catalunya. 2022. URL: https://www.idescat.cat/tema/xifpo [accessed 2023-05-23]
5. LLEI 15/1990, de 9 de juliol, d'ordenació sanitària de Catalunya. Generalitat de Catalunya. 1990. URL: https://portaljuridic.gencat.cat/ca/document-del-pjur/?documentId=48426 [accessed 2024-01-08]
6. Garcia O, Medina A, Schiaffino A. Comportaments relacionats amb la salut, l’estat de salut i l’ús de serveis sanitaris a Catalunya: informe dels principals resultats 2015. Generalitat de Catalunya. URL: http://tinyurl.com/9tmz3fxv [accessed 2024-01-11]
7. L’estat de salut, els comportaments relacionats amb la salut i l’ús de serveis sanitaris a Catalunya 2021. Generalitat de Catalunya. URL: http://tinyurl.com/53ctyy42 [accessed 2024-01-08]
8. Triple aim and population health. Institute for Healthcare Improvement. 2022. URL: https://www.ihi.org/Engage/Initiatives/TripleAim/Pages/default.aspx [accessed 2022-11-17]
9. Nundy S, Cooper LA, Mate KS. The quintuple aim for health care improvement: a new imperative to advance health equity. JAMA 2022 Feb 08;327(6):521-522. [doi: 10.1001/jama.2021.25181] [Medline: 35061006]
10. Bodenheimer T, Sinsky C. From triple to quadruple aim: care of the patient requires care of the provider. Ann Fam Med 2014;12(6):573-576 [FREE Full text] [doi: 10.1370/afm.1713] [Medline: 25384822]
11. Starfield B. Is primary care essential? Lancet 1994 Oct 22;344(8930):1129-1133 [FREE Full text] [doi: 10.1016/s0140-6736(94)90634-3] [Medline: 7934497]
12. La COVID-19 y la necesidad de actuar en relación con la salud mental. Naciones Unidas. 2020. URL: https://www.un.org/sites/un2.un.org/files/policy_brief_-_covid_and_mental_health_spanish.pdf [accessed 2024-01-08]
13. Efectes del confinament en la salut mental (42/2021). Generalitat de Catalunya. 2021. URL: https://scientiasalut.gencat.cat/handle/11351/5576 [accessed 2024-01-08]
14. Fernández de Sanmamed J, González Y, Mazo V, Pons G, Robles I, Serrano E, et al. Al voltant del model d'atenció a la salut mental a l'atenció primària de salut. Fòrum Català d'Atenció Primària. 2020. URL: https://focap.files.wordpress.com/2020/07/al-voltant-del-model-datencic3b3-a-la-salut-mental-a-laps.pdf [accessed 2024-01-08]
15. Informe sobre consum problemàtic i conseqüències: consum de fàrmacs amb potencial d'abús. Generalitat de Catalunya. 2021. URL: https://scientiasalut.gencat.cat/handle/11351/6359 [accessed 2024-01-08]
16. Baltaxe E, Cano I, Herranz C, Barberan-Garcia A, Hernandez C, Alonso A, et al. Evaluation of integrated care services in Catalonia: population-based and service-based real-life deployment protocols. BMC Health Serv Res 2019 Jun 11;19(1):370 [FREE Full text] [doi: 10.1186/s12913-019-4174-2] [Medline: 31185997]
17. Lopez Segui F, Hernandez Guillamet G, Pifarré Arolas H, Marin-Gomez FX, Ruiz Comellas A, Ramirez Morros AM, et al. Characterization and identification of variations in types of primary care visits before and during the COVID-19 pandemic in Catalonia: big data analysis study. J Med Internet Res 2021 Sep 14;23(9):e29622 [FREE Full text] [doi: 10.2196/29622] [Medline: 34313600]
18. Plade salut de Catalunya 2011-2015. Generalitat de Catalunya. URL: http://tinyurl.com/yc3zpa63 [accessed 2024-01-08]

19. La meva salut. Generalitat de Catalunya. 2022. URL: https://catsalut.gencat.cat/ca/serveis-sanitaris/la-meva-salut/ [accessed 2022-11-28]
20. eConsulta. Generalitat de Catalunya. 2021. URL: http://tinyurl.com/mv2bb96x [accessed 2022-11-07]
21. López Seguí F, Walsh S, Solans O, Adroher Mas C, Ferraro G, García-Altés A, et al. Teleconsultation between patients and health care professionals in the Catalan primary care service: message annotation analysis in a retrospective cross-sectional study. J Med Internet Res 2020 Sep 17;22(9):e19149 [FREE Full text] [doi: 10.2196/19149] [Medline: 32687477]
22. Fernández OS, Seguí FL, Vidal-Alaball J, Bonet Simo JM, Vian OH, Cabo PR, et al. Primary care doctor characteristics that determine the use of teleconsultations in the Catalan public health system: retrospective descriptive cross-sectional study. JMIR Med Inform 2020 Jan 31;8(1):e16484 [FREE Full text] [doi: 10.2196/16484] [Medline: 32012061]
23. Solans O, Vidal-Alaball J, Roig Cabo P, Mora N, Coma E, Bonet Simó JM, et al. Characteristics of citizens and their use of teleconsultations in primary care in the Catalan public health system before and during the COVID-19 pandemic: retrospective descriptive cross-sectional study. J Med Internet Res 2021 May 27;23(5):e28629 [FREE Full text] [doi: 10.2196/28629] [Medline: 33970867]
24. Vidal-Alaball J, López Seguí F. Ha llegado para quedarse: beneficios e inconvenientes de la eConsulta. Atención Primaria Práctica 2020 Jan;2(1-2):100043. [doi: 10.1016/j.appr.2020.100043]
25. López Seguí F, Vidal-Alaball J, Sagarra Castro M, García-Altés A, García Cuyàs F. General practitioners' perceptions of whether teleconsultations reduce the number of face-to-face visits in the Catalan public primary care system: retrospective cross-sectional study. J Med Internet Res 2020 Mar 16;22(3):e14478 [FREE Full text] [doi: 10.2196/14478] [Medline: 32175914]
26. Institut CDLS. Institut Català de la Salut memòria 2020. Generalitat de Catalunya. 2021. URL: http://tinyurl.com/bdeda5js [accessed 2024-01-08]
27. Entitats proveïdores d'atenció primària. Generalitat de Catalunya. URL: http://tinyurl.com/bdess5nf [accessed 2023-07-04]
28. CIM-10-MC/SCP. Generalitat de Catalunya. 2022. URL: http://tinyurl.com/3fdak2z2 [accessed 2022-11-13]
29. Markon KE, Fossati A, Somma A, Krueger B. Guía de Consulta de los Criterios Diagnósticos del DSM-5. Washington, DC: American Psychiatric Association; 2014.
30. SIDIAP. URL: https://www.sidiap.org/index.php/ca/ [accessed 2024-01-11]
31. Domínguez-Berjón MF, Borrell C, Cano-Serral G, Esnaola S, Nolasco A, Pasarín MI, et al. Constructing a deprivation index based on census data in large Spanish cities (the MEDEA project). Gac Sanit 2008;22(3):179-187 [FREE Full text] [doi: 10.1157/13123961] [Medline: 18579042]
32. Pla director de salut mental i addiccions. Generalitat de Catalunya. 2017. URL: http://tinyurl.com/3kbp3d7d [accessed 2024-01-08]
33. Consideracions sobre l'atenció a la salut mental en l'àmbit sanitari d'Atenció Primària. Metges de Catalunya. 2021. URL: http://tinyurl.com/3pssnvdm [accessed 2024-01-08]
34. Salud mental e infancia en el escenario de la COVID-19. Propuestas de UNICEF España. UNICEF España. 2020. URL: http://tinyurl.com/strk8nry [accessed 2024-01-08]
35. Pla de salut 2021-2025. Generalitat de Catalunya. 2021. URL: https://salutweb.gencat.cat/ca/departament/pla-salut/ [accessed 2024-01-08]
36. Oertelt-Prigione S, Parol R, Krohn S, Preissner R, Regitz-Zagrosek V. Analysis of sex and gender-specific research reveals a common increase in publications and marked differences between disciplines. BMC Med 2010 Nov 10;8:70 [FREE Full text] [doi: 10.1186/1741-7015-8-70] [Medline: 21067576]
37. Depression in adults: treatment and management NICE guideline. National Institute for Health and Care Excellence. 2022. URL: https://www.nice.org.uk/guidance/ng222 [accessed 2024-01-08]
38. Guía de práctica clínica sobre el manejo de la depresión en el adulto. Ministerio de Sanidad, Servicios Sociales e Igualdad. 2014. URL: http://tinyurl.com/2yy6a3t9 [accessed 2024-01-08]
39. Coma E, Miró Q, Medina M, Marin-Gomez FX, Cos X, Benítez M, et al. Association between the reduction of face-to-face appointments and the control of patients with type 2 diabetes mellitus during the Covid-19 pandemic in Catalonia. Diabetes Res Clin Pract 2021 Dec;182:109127 [FREE Full text] [doi: 10.1016/j.diabres.2021.109127] [Medline: 34752800]
40. López Seguí F, Ander Egg Aguilar R, de Maeztu G, García-Altés A, García Cuyàs F, Walsh S, et al. Teleconsultations between patients and healthcare professionals in primary care in Catalonia: the evaluation of text classification algorithms using supervised machine learning. Int J Environ Res Public Health 2020 Feb 09;17(3):1093 [FREE Full text] [doi: 10.3390/ijerph17031093] [Medline: 32050435]
41. Telemedicina: com i quan utilitzar-la en la pràctica assistencial. Quad la Bona Praxi. 2021. URL: https://www.comb.cat/pdf/publicacions/qbp/praxi37.pdf [accessed 2024-01-08]
42. Han SM, Greenfield G, Majeed A, Hayhoe B. Impact of remote consultations on antibiotic prescribing in primary health care: systematic review. J Med Internet Res 2020 Nov 09;22(11):e23482 [FREE Full text] [doi: 10.2196/23482] [Medline: 33031045]
43. Wabe N, Thomas J, Sezgin G, Sheikh MK, Gault E, Georgiou A. Medication prescribing in face-to-face versus telehealth consultations during the COVID-19 pandemic in Australian general practice: a retrospective observational study. BJGP Open 2021 Nov 24;6(1):BJGPO.2021.0132. [doi: 10.3399/bjgpo.2021.0132]

44. Anatomical therapeutic chemical (ATC) classification. World Health Organization. URL: https://www.who.int/tools/atc-ddd-toolkit/atc-classification [accessed 2023-06-28]

Abbreviations

ICD-10: International Statistical Classification of Diseases and Related Health Problems
OR: odds ratio
PC: primary care

Edited by J Torous; submitted 16.09.23; peer-reviewed by A Baena Garcia, C Asuzu; comments to author 08.12.23; revised version received 12.12.23; accepted 28.12.23; published 18.01.24.

Please cite as:
Fuster-Casanovas A, Miró Catalina Q, Vidal-Alaball J, Escalé-Besa A, Carrión C
eHealth in the Management of Depressive Episodes in Catalonia's Primary Care From 2017 to 2022: Retrospective Observational Study
JMIR Ment Health 2024;11:e52816
URL: https://mental.jmir.org/2024/1/e52816
doi:10.2196/52816
PMID:38236631

Original Paper

Incorporating a Stepped Care Approach Into Internet-Based Cognitive Behavioral Therapy for Depression: Randomized Controlled Trial

Jasleen Kaur Jagayat[1*], MSc; Anchan Kumar[2], MD; Yijia Shao[2,3], MPS; Amrita Pannu[2], MD; Charmy Patel[2], MPH, MACP; Amirhossein Shirazi[2,4], MD, PhD; Mohsen Omrani[2,4], MD, PhD; Nazanin Alavi[1,2*], MD

[1]Centre for Neuroscience Studies, Queen's University, Kingston, ON, Canada
[2]Department of Psychiatry, Queen's University, Kingston, ON, Canada
[3]Pastoral Studies, Toronto School of Theology (TST), Emmanuel College, University of Toronto, Toronto, ON, Canada
[4]Online Psychotherapy Tool, Toronto, ON, Canada
[*]these authors contributed equally

Corresponding Author:
Jasleen Kaur Jagayat, MSc
Centre for Neuroscience Studies
Queen's University
18 Stuart Street
Kingston, ON, K7L 3N6
Canada
Phone: 1 613 533 2600
Email: jasleenjagayat@gmail.com

Abstract

Background: Depression is a hidden burden, yet it is a leading cause of disability worldwide. Despite the adverse effects of depression, fewer than one-third of patients receive care. Internet-based cognitive behavioral therapy (i-CBT) is an effective treatment for depression, and combining i-CBT with supervised care could make the therapy scalable and effective. A stepped care model is a framework for beginning treatment with an effective and low-intensity intervention while adapting care based on the patient's needs.

Objective: This study investigated the efficacy of a stepped care i-CBT model for depression based on changes in self-reported depressive symptoms.

Methods: In this single-blinded, randomized controlled trial, participants were allocated to either the i-CBT–only group (28/56, 50%) or the i-CBT with stepped care group (28/56, 50%). Both groups received a 13-week i-CBT program tailored for depression. The i-CBT program was provided through a secure, online mental health clinic called the Online Psychotherapy Tool. Participants read through the sessions and completed the assignments related to each session. Participants in the stepped care group received additional interventions from their care provider based on standard questionnaire scores (ie, Patient Health Questionnaire–9 [PHQ-9], Quick Inventory of Depressive Symptomatology [QIDS], and Quality of Life Enjoyment and Satisfaction Questionnaire–Short Form) and their assignment responses. From lowest to highest intensity, the additional interventions included SMS text messages, phone calls, video calls, or a video call with a psychiatrist.

Results: For this study, 56 participants were recruited to complete an i-CBT program (n=28, 50%; mean age 37.9; SD 13.08 y; 7/28, 27% were men) or an i-CBT with stepped care program (n=28, 50%; mean age 40.6; SD 14.28 y; 11/28, 42% were men). The results of this study indicate that the i-CBT program was effective in significantly reducing depressive symptoms, as measured by the PHQ-9 ($F_{4,80}$=9.95; P<.001) and QIDS ($F_{2,28}$=5.73; P=.008); however, there were no significant differences in the reduction of depressive symptoms between the 2 groups (PHQ-9: $F_{4,80}$=0.43; P=.78; QIDS: $F_{2,28}$=3.05; P=.06). The stepped care group was not significantly better in reducing depressive symptoms than the i-CBT group (PHQ-9, P=.79; QIDS, P=.06). Although there were no significant differences observed between the number of participants who completed the program between the groups (χ^2_1=2.6; P=.10), participants in the stepped care group, on average, participated in more sessions than those who prematurely terminated participation in the i-CBT group (t_{55}=−2; P=.03; 95% CI –4.83 to –0.002).

Conclusions: Implementing a stepped care approach in i-CBT is an effective treatment for depression, and the stepped care model can assist patients to complete more sessions in their treatment.

Trial Registration: Clinicaltrials.gov NCT04747873; https://clinicaltrials.gov/study/NCT04747873

(JMIR Ment Health 2024;11:e51704) doi:10.2196/51704

KEYWORDS

internet-based cognitive behavioral therapy; i-CBT; major depressive disorder; MDD; stepped care; digital mental health care; mobile phone

Introduction

Background and Rationale

Depression is a leading cause of disability, affecting approximately 3.8% of the population worldwide [1,2]. Major depressive disorder (MDD) is characterized by persistent feelings of sadness, negative mood, or loss of interest in life activities [3]. Detrimental and persistent changes in appetite, sleep, energy, and cognition may accompany these feelings. In addition to its deleterious impacts on mental health, depression is associated with increased morbidity, decreased quality of life, and reduced work productivity [4-6]. Despite the negative consequences of depression, only one-third of individuals receive treatment, and of those, only 3 in 5 people receive sufficient care [7,8].

Cognitive behavioral therapy (CBT) is an effective treatment for MDD [9-11]. It is a form of psychotherapy that focuses on cognitive restructuring strategies and behavioral activation techniques to help individuals with depression overcome their depressive symptoms and modify ineffective thinking patterns. Although effective, CBT on its own presents some challenges such as accessibility issues, lack of follow-ups, and increased costs [12-14]. Over the past 2 decades, issues of accessibility are being increasingly addressed through digital modalities, such as internet-based CBT (i-CBT) [11,15,16]. i-CBT allows for the delivery of CBT through digital media including computers and smartphones. It involves using web-based sessions with interactive components and practices to teach individuals ways to improve their mental health. Many studies have shown the efficacy of i-CBT for depression and anxiety, with results comparable with those of traditional in-person CBT [11,16-20]. The digital format allows for increased accessibility and convenience when delivering and receiving psychotherapy while remaining effective [21,22]; however, the digital format comes with limitations, including low adherence. Adherence rates are variable and can be affected by multiple factors including the digital modalities used and patient-therapist interactions [23-25]. Furthermore, attrition rates in CBT are the highest in patients with depression [26], and adherence can be compromised as disease severity increases [27]. With the benefit of being versatile and adaptable, i-CBT may be adapted to include support to address the issues of adherence.

The stepped care model is an approach to providing individuals with adaptable and effective care that is adjusted based on symptom severity [28-31]. This dynamic approach allows care to be adapted as required throughout an individual's treatment plan to best tailor care to the individual's needs. Stepped care models aim to provide the most effective and efficient care by minimizing resources and costs for care. The least intensive intervention, such as self-help materials or psychoeducation, is first provided to patients in a stepped care model. The patient may then advance to the next level of care, which may include more extensive interventions such as increased support or medication if they do not improve or just partially improve. Until the patient gets the desired result or reaches the highest level of care, this process is repeated. This stepped care model allows for efficient allocation of resources, increases access to care, can reduce wait times, can lower costs, and can increase versatility in treatment options [28,31]. This model is effective in treating mental health disorders including MDD and focuses on a patient-centered approach to care [21,31].

In the stepped care model, all patients receive some form of treatment for their symptoms, and based on the symptom severity and progress through treatment, additional care is added to their treatment plan. The addition of therapist-guided support to i-CBT has demonstrated high levels of patient satisfaction and a general decrease in symptom severity [32]. It is also noted that dropout rates decrease as increased therapeutic support is provided to patients. A meta-analysis that gathered dropout rates of CBT based on treatment support levels found that without support, the dropout rate was 74%; with administrative support, the dropout rate was 38%; and with therapeutic support, the patient dropout rate was 28% [16]. This suggests that the type of support provided to patients is important. Therapist-guided support in an i-CBT setting can also include text (ie, messaging and emails) or audio-based (ie, telephone and video) formats, and the mode of therapy may influence treatment outcome [22,33,34]. Incorporating text and call interventions into online therapy can provide an additional support outlet for patients throughout their treatment by providing an additional means of communication while remaining accessible [35]. Phone and video interventions have both been shown to be effective in reducing symptoms and maintaining patients in care compared with in-person care for mood disorders [36]. Phone calls may be more broadly accessible than video calls owing to inherent technological barriers; however, video calls allow for additional nonverbal cues to be assessed in care [36]. They can serve as check-ins, encouragement, and reminders and help monitor patient progress between treatment sessions. Through this additional support, a review of therapeutic concepts and words of encouragement are found to help decrease symptom severity and improve therapy completion [37]. This method can help with building therapist-patient rapport throughout treatment and act as a valuable tool in online therapy.

Objectives

This study was a single-blinded, randomized controlled trial exploring the efficacy of a stepped care model in an i-CBT program compared with an i-CBT–only program for adults with MDD. The i-CBT program for depression was provided through a secure, online mental health clinic called the Online Psychotherapy Tool (OPTT) [38]. The stepped care model considered the varying digital modalities of text-based, phone-based, and video-based interventions to help support individuals in their care.

Methods

Ethical Considerations

The research study was approved by the Health Sciences and Affiliated Teaching Hospitals Research Ethics Board (file number 6031992) at Queen's University in Kingston, Ontario, Canada. The research study followed the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) guidelines and can be viewed in Multimedia Appendix 1 [39].

No monetary compensation was provided to participants for participation in this study because they were provided with CBT treatment for their symptoms.

Data Privacy

To maintain participant confidentiality, each participant was assigned a randomized participant ID number. This ID was used to identify the participant for data analysis purposes. Only associated care providers had access to the participant's identity owing to the nature of the CBT intervention. If any technical issues on OPTT arose, this ID was used to resolve the issues to maintain participant anonymity.

Participant information maintained online was password protected, including consent forms, identities, and CBT content. All encrypted files were stored on a safe server run by Queen's University. Participant identities and consent forms are stored on-site at Queen's University's exclusive storage in Kingston, Ontario, Canada, for 5 years following the conclusion of the study because they are considered as medical records. Following the 5 years, the participant records will be destroyed. Participants were informed about the possibility to withdraw from the study at any time if they wished to do so. In upcoming plans for knowledge dissemination and publication of outcomes, participant identification will be secured and maintained anonymously.

The online platform used to deliver the i-CBT program, OPTT [38,40], serves as the repository for all data. OPTT complies with the Personal Information Protection and Electronic Documents Act, Health Insurance Portability and Accountability Act, and Service Organization Control–2. The cloud infrastructure of Amazon Web Service Canada is used to host all servers and databases, and Medstack manages it to ensure compliance with all the local, state, and federal privacy and security laws. For privacy reasons, OPTT does not gather any identifying personal data or IP addresses. Only anonymous metadata were gathered by OPTT to enhance the quality of its services and provide the care provider team access to participant analytics (ie, interaction with the OPTT platform). No OPTT employee had direct access to participant data owing to data encryption procedures (ie, participant ID). All encrypted backups are stored at Queen's University.

Outcome Evaluation

Patient Health Questionnaire–9

The Patient Health Questionnaire–9 (PHQ-9) is a self-assessment questionnaire designed to examine a participant's depression severity through a 9-item questionnaire, with an additional question about functional health [41]. Each of the 9 items is scored from 0 (not at all) to 3 (nearly every day) and corresponds to the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* criteria for MDD symptoms [3]. Each item is then scored for a total to indicate depression severity from none to severe, ranging from 0 to 27. High scores indicate high depression.

Quick Inventory of Depressive Symptomatology Questionnaire

The Quick Inventory of Depressive Symptomatology (QIDS) questionnaire is a self-assessment questionnaire designed to screen for depression and measure depression severity based on 16 items [42]. Each item correlates with the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* criteria for MDD symptoms and is scored using a 4-point scale (0-3), for a total score ranging between 0 and 27 [3]. To calculate the final score, the questions are summed using the highest response in the following domains: sleep patterns (questions 1-4), sad mood (questions 5), change in appetite (questions 6-9), concentration and decision-making (questions 10), self-view (questions 11), suicidal thoughts (questions 12), general interest (questions 13), energy level (questions 14), and psychomotor effects (questions 15 and 16). High scores indicate high depression severity.

Quality of Life Enjoyment and Satisfaction Questionnaire–Short Form

The Quality of Life Enjoyment and Satisfaction Questionnaire–Short Form (Q-LES-Q) is a self-assessment questionnaire designed to collect information about the level of enjoyment and satisfaction in various aspects of daily functioning through a series of 16 items [43]. Each item is rated on a scale from 1 (very poor) to 5 (very good). Of the 16 items, only the first 14 items are summed for a total raw score ranging from 14 to 70. High scores indicate great quality of life.

Measured Outcomes

The primary outcomes measured in this study included changes in depressive symptoms, as measured by the PHQ-9 and QIDS. Using the Q-LES-Q to measure changes in the participant's quality of life was also a primary outcome that was evaluated. Participants completed the PHQ-9 every 3 weeks (weeks 1, 4, 7, 10, and 13) and the QIDS and Q-LES-Q at 3 time points (weeks 1, 7, and 13) throughout the program. All the 3 questionnaires are intended to be completed at the 3-, 6-, 9-, and 12-month follow-up periods during the 1-year follow-up phase (Textbox 1). Pretreatment and posttreatment changes in depressive symptoms (ie, PHQ-9 and QIDS) and quality of life (Q-LES-Q) between the control and experimental groups were compared. The number of sessions completed by each participant was a secondary measure that was considered when evaluating compliance with the i-CBT program between the 2 groups.

Textbox 1. Schedule of the assigned clinical questionnaires sent to participants, including the 1-year follow-up period.

Session 1

- Patient Health Questionnaire–9 (PHQ-9), Quality of Life Enjoyment and Satisfaction Questionnaire–Short Form (QIDS), Quick Inventory of Depressive Symptomatology (Q-LES-Q), and demographic survey

Session 2

- No questionnaires are provided during these sessions.

Session 3

- No questionnaires are provided during these sessions.

Session 4

- PHQ-9 and initial stepped care assessment (if the participant is in the stepped care group, an intervention decision will be made at the end of every session, starting at session 4. Interventions will be implemented, beginning from session 5)

Session 5

- Interventions begin as needed (if the participant is in the stepped care group, an intervention decision will be made at the end of every session, starting at session 4; Interventions will be implemented, beginning from session 5)

Session 6

- No questionnaires are provided during these sessions.

Session 7

- PHQ-9, QIDS, and Q-LES-Q

Session 8

- No questionnaires are provided during these sessions.

Session 9

- No questionnaires are provided during these sessions.

Session 10

- PHQ-9

Session 11

- No questionnaires are provided during these sessions.

Session 12

- No questionnaires are provided during these sessions.

Session 13

- PHQ-9, QIDS, and Q-LES-Q

Follow-up 1

- PHQ-9, QIDS, and Q-LES-Q

Follow-up 2

- PHQ-9, QIDS, and Q-LES-Q

Follow-up 3

- PHQ-9, QIDS, and Q-LES-Q

Follow-up 4

- PHQ-9, QIDS, and Q-LES-Q

Participants

This study was registered at ClinicalTrial.gov (NCT04747873). Participants were recruited from the Providence Care outpatient psychiatry clinic and Kingston Health Sciences Center sites (Hotel Dieu Hospital and Kingston General Hospital), both of which are located in Kingston, Ontario, Canada. Physicians familiar with the Queen's Online Psychotherapy Lab (QUOPL) research team were also informed about the study and directed patients who may benefit from CBT toward the study when considered appropriate. Self-referrals were also accepted for this study. Recruitment was managed by the laboratory manager who was the initial point of contact for all participants.

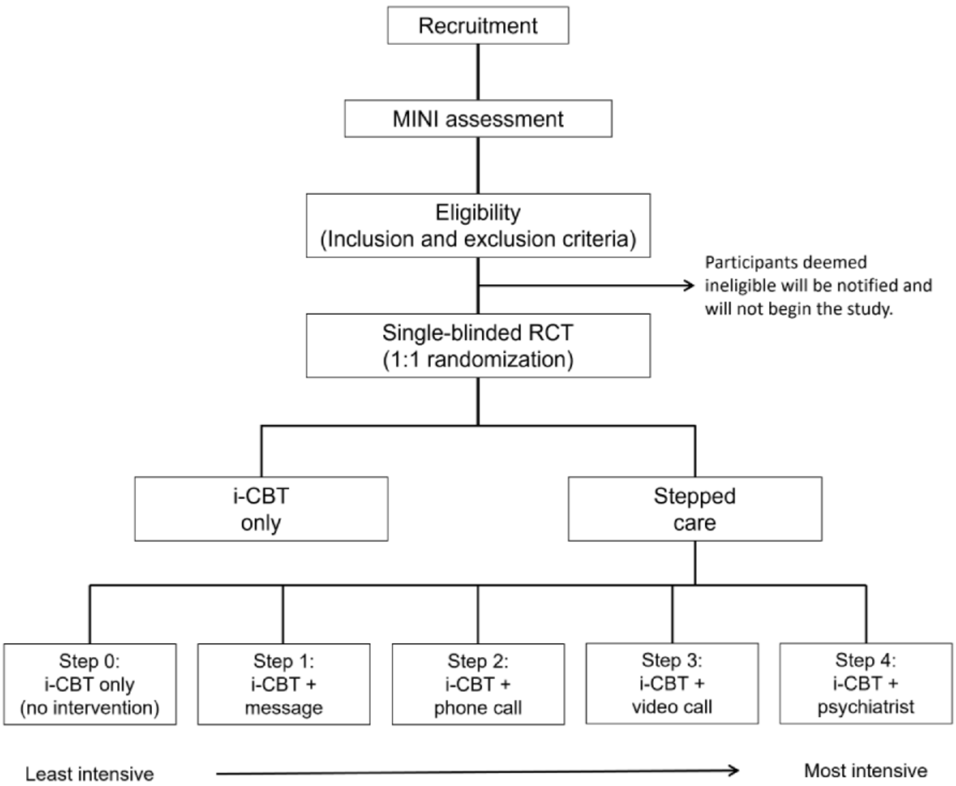
Eligibility Criteria

Individuals interested in the research study were provided with a letter of information and consent form to understand the study design before beginning the eligibility process. Following the informed consent process (written or verbal), a trained research assistant at QUOPL screened the individual based on the eligibility criteria. Individuals were eligible for the study if they met the following criteria: (1) aged ≥18 years, (2) met the criteria of MDD by a trained research assistant according to the Mini International Neuropsychiatric Interview (MINI) [44], (3) showed competence to consent to participate, (4) fluent in English because the i-CBT program was provided in English only, and (5) had consistent and reliable access to the internet. For the MINI assessment, individuals were first assessed by a trained research assistant to support a diagnosis of MDD. This MINI assessment was completed through a secure Microsoft Teams video call. Participants were ineligible for the research study if they presented with active psychosis, acute mania, severe alcohol or substance use disorder, or active suicidal or homicidal ideation. If a participant received or was receiving CBT in the past year at the time of beginning the study, they were excluded from the study to avoid confounding effects on the efficacy of this i-CBT program.

Following the consent and eligibility screening process, participants were randomized in a 1:1 allocation ratio to either the i-CBT group (control group) or the i-CBT with stepped care group (experimental group). Randomization was computer generated in Microsoft Excel using the *RANDBETWEEN* function. This method returned an integer at random that indicated either 1 (designated as the i-CBT group) or 2 (for the stepped care group). This integer was used to determine the group allocation for the participant.

Using G*Power (version 3.1.9.7 [45]), a priori power analysis was performed to determine the minimum sample size required to test the study hypothesis. The average PHQ-9 score decreased from 16.2 before i-CBT to 11.48 (combined SD 5.45) after 12 sessions in our previous clinical trials and data collection regarding i-CBT for depression [46]. These figures led to an effect size (Hedge g of 0.86). A paired sample *t* test (1-tailed) would require 14 individuals to detect a significant impact, given the effect size and a power of 0.8. For online CBT programs, compliance and care adherence are common challenges; therefore, we predict 50% dropout rate based on previous clinical trials conducted in our laboratory [47,48]. Consequently, considering the proposed sample size calculation accounted for this dropout rate, we aimed to recruit 28 participants in each group. With the 2 treatment arms, our total sample size was expected to be 56 participants. Figure 1 summarizes the study design in a flow chart.

Figure 1. A flowchart providing an overview of the study design. i-CBT: internet-based cognitive behavioral therapy; MINI: Mini International Neuropsychiatric Interview; RCT: randomized controlled trial.



Care Providers

Participants were assigned a primary care provider to help build rapport through assigning sessions, providing feedback, and providing additional interventions as needed. All care providers were trained in psychotherapy by the lead psychiatrist who is a licensed psychotherapist and expert in internet-based psychotherapy. Each participant was assigned a care team that consisted of a care provider and a psychiatrist with extensive experience in i-CBT. Care providers were research assistants recruited and trained by the team psychiatrist. Their role was to provide participants with feedback about weekly homework and act as a point of contact for any questions that the participants may have regarding the study design. The care providers included master's degree students studying in the field of neuroscience and psychology, associated with QUOPL, and research assistants who have completed a medical degree. Through training, all care providers were well experienced in internet-based psychotherapy, with a specific focus on CBT techniques. They were all taught the standard care pathway and the aim, and content of each therapeutic session. They also continued receiving specialized training through webinars, CBT workshops led by a psychiatrist trained in CBT, and exercises with feedback during the study. Moreover, they were provided predesigned feedback templates tailored to each session to be used when writing the weekly feedback. The feedback templates helped standardize the feedback by providing a basic structure that ensured all aspects of the homework were acknowledged and that the newly presented CBT concepts were reviewed. Feedback templates varied between sessions, and care providers personalized each template for each participant's homework. The use of feedback templates that were drafted by the team psychiatrist trained in online psychotherapy allowed some control over treatment consistency across sessions and participants. Before submission to the participant, feedback was always reviewed by another care provider (MSc candidate) trained in psychotherapy and overlooked by a psychiatrist to ensure the quality of the feedback. The team psychiatrist fulfilled the role to overlook the feedback; manage any crisis events if necessary; and provide support in the stepped care model, specifically the highest-intensity intervention, step 4.

Intervention

i-CBT Program

All participants (56/56, 100%) were provided with the same 13-week i-CBT program designed for MDD through OPTT—a secure, cloud-based, digital mental health platform [38]. This i-CBT program, titled "Electronic Cognitive Behavioural Therapy (e-CBT) For Depression," is a set of predesigned modules created to address depressive symptoms through various CBT techniques to work on cognitive restructuring [46,49-51]. This specific program has undergone validation through previous studies and has been shown to significantly decrease depressive symptoms [47,49]. The program was intended to imitate in-person, standard CBT for MDD in an asynchronous format. In this study, 1 module was assigned to participants on the same day of each week. A module included 20 to 30 slides of CBT content and took approximately 45 to 50 minutes to view (Multimedia Appendix 2). Each module focused on 1 CBT technique (Table 1). Each module followed a similar outline, beginning with an introduction to the topic for the week, an overview of the technique with examples, and concluding with homework related to the weekly technique [50]. The homework consisted of a few questions based on a stressful event that occurred in the participant's life in the past week and provided participants with the practice of the new CBT technique. Participants were provided with 4 days to asynchronously review the module and submit their homework to their assigned care provider on the OPTT platform. Participant answers were then reviewed by the assigned care provider over 3 days who provided participants with feedback about their answers along with their next session on the assigned day of the week.

Table 1. A brief description of each session for the internet-based cognitive behavioral therapy (i-CBT) program, titled "Electronic Cognitive Behavioural Therapy (e-CBT) For Depression," on the Online Psychotherapy Tool platform.

Module	Title	Description
1	What is depression?	• Introduced i-CBT and discussed common depressive symptoms while setting expectations for the course
2	The 5-Part Model	• Described the 5-Part Model: connections and interactions between a situation, thoughts, feelings, physical reactions, and behaviors
3	Sleep hygiene	• Focused on how to improve rest through sleep habits and offered a variety of techniques
4	Strategies for stressful situations	• Provided a general review of practical strategies that can be applied under difficult circumstances, including breathing techniques and activities
5	Thoughts, feelings, behavior, physical reactions, and environment	• Explained the 5-Part Model in detail and how modifications to 1 part can influence the remaining 4 parts
6	The thought record	• Featured the first 3 columns of the thought record: a tool for challenging ineffective thoughts • The first 3 columns included the situation, feelings, and automatic thoughts associated with the situation
7	Automatic thoughts	• Explored the function of automatic thoughts and how they affect emotions • This included learning how to recognize automatic thoughts and the "hot thought" • Common cognitive errors were also discussed
8	Activity scheduling	• Described the activity record: a tool for recording weekly activities and finding connections between activities and associated moods
9	Evidence	• Back to the thought record, the fourth and fifth columns were explained, which included looking at evidence for and against the most intense thought
10	Alternative and balanced thinking	• Concluded the 5-Part Model with the final 2 columns outlining alternative or balanced thinking and rerating feelings
11	Experiments	• Introduced a behavioral activation technique of experiments to promote belief in alternative or balanced thinking
12	Action plans	• Promoted working on the identified problems using a structured tool called the action plan
13	Review	• Reviewed the 12 modules in the program and summarized the key tools and techniques

The feedback was based on predesigned feedback templates created by the lead psychiatrist and i-CBT expert at QUOPL. The feedback for each session followed a similar structure and was delivered as a letter addressed to the participant. The feedback template mirrored the following structure: (1) addressed the participant and thanked them for their work over the week while acknowledging any adverse events that occurred; (2) commented about their mood and sleep quality in the previous week; (3) summarized the content of the module, highlighting the main concepts; (4) summarized the participant's answers; (5) empathized with the participant based on their shared experiences and encouraged them to use the learned CBT techniques; and (6) thanked them again for their time and provided the participant with a brief introduction about the next module's content (Multimedia Appendix 3—sample feedback template for session 1). Overall, the feedback itself focused on the participant's mood and sleep patterns over the week, progress on their weekly goals, and their understanding of CBT concepts. The care providers personalized the feedback template based on the participant's responses.

The OPTT platform provided participants and care providers with a modality to communicate asynchronously regarding the availability of the next module and any questions or concerns regarding the OPTT operations or the study design. These questions and concerns were limited to the study design and technical issues, rather than personal or therapy-related concerns. Care providers viewed and replied to these OPTT messages at least once a week. For concerns regarding the OPTT platform and technical issues, participants were redirected to OPTT's technical support team.

i-CBT With Stepped Care

Although all participants (56/56, 100%) had access to the i-CBT program, a designated care provider, and weekly feedback, only participants in the stepped care group received additional interventions as per the proposed stepped care model. The stepped care interventions were provided by their care provider or the team psychiatrist associated with their care as required. The structure of the stepped care model used in this study followed the 4 identified steps of a preventative stepped care model based on 10 randomized controlled trials [52]. These four steps included the following: (1) watchful waiting, (2) self-help psychotherapy, (3) face-to-face psychotherapy, and (4) referral to specialists. The intervention that participants received was dictated by the participant's care team: the care provider and the team psychiatrist. This model initiated all participants at the lowest intervention of i-CBT beginning at session 1 and provided additional interventions beginning at session 5 after monitoring their progress through the first 4 sessions and obtaining 2 PHQ-9 scores as a quantitative assessment measure, to help with the decision (Textbox 1). Subsequently, an intervention could be added or changed in each session by the care team, following session 5. Beginning stepped care at session 5 allowed for a watchful waiting period as the initial step to understand the patient's symptoms better, as seen in preventative stepped care models for anxiety and depressive disorders [52]. The first 3 interventions, steps 1 to 3, are similar in content and care provider, but the delivery modality varies. These first 3 steps included varying intensities to care and effort required by the care provider. Step 4 differed from the first 3 steps in that the team psychiatrist was involved and the content of care in this intervention was different. Details of each step in the intervention are provided in the following sections.

Step 0: i-CBT Only

This intervention was considered to have the lowest intensity because the participant was provided with an i-CBT program for depression that is effective in treating MDD [49]. This was also known as the starting point for all participants. All participants began at step 0, which included completing the i-CBT program. In this step, no intervention was added to the participants' care, which allows for patient monitoring during a watchful waiting phase as they receive a low-intensity treatment [52]. Furthermore, many studies have control groups that encompass waitlist groups; however, more studies should focus on implementing i-CBT treatments as the control group for a wholesome review of the effects of stepped care models placed in these i-CBT programs. Therefore, step 0 mimics the procedure for the control group to allow for comparable effects between the 2 groups.

Step 1: i-CBT With Messaging

This intervention was the first added intervention following step 0 and included the addition of asynchronous check-in messages to the participant from their care provider on the OPTT platform. When the care provider sent the next session, they also sent a personalized message based on the message templates for that weekly session (Multimedia Appendix 3—sample message template). The message focused on addressing the participant and checking in with them by asking them how their week was so far. This intervention was a brief exchange between the participant and their care provider and a way to add active human support to the participant's care through direct messaging [53,54]. Upon receiving the participant's response, care providers acknowledged their experiences and reminded them to complete their next session.

Step 2: i-CBT With a Telephone Call

This intervention included a brief telephone call from the care provider for live, verbal support [55-57]. The call is limited to one 15 to 20–minute call during the week [58]. After sending the weekly session, care providers called the participant before the next session. The telephone call focused on asking participants how their week was and acknowledging the participant's experience (Multimedia Appendix 3—sample call template). Care providers also asked for updates regarding module completion for the week and reminded participants to submit their weekly homework. This template was similar to that of step-1 messages, except that it included direct verbal encouragement.

Step 3: i-CBT With a Video Call

This intervention included a brief, secure, Microsoft Teams video call from the care provider for live support and visual contact between the participant and care provider, in addition to verbal cues. The video call is limited to one 20 to 30–minute call during the week [22,59]. Care providers set up a video call with the participant before the next session. The video call focused on the participant's experience over the past week and how they are doing (Multimedia Appendix 3—sample message template). Care providers also encouraged participants to submit their weekly homework if they had not submitted yet. This template was similar to that of step-2 telephone calls and included direct, live encouragement in hopes to mimic a face-to-face setting.

Step 4: i-CBT With Psychiatrist Call

This intervention was the highest-intensity intervention provided in the stepped care model. This intervention included a web-based, one-on-one psychiatrist appointment with the team psychiatrist using Microsoft Teams. The psychiatrist discussed the participant's current challenges and potential options for the participant. This may include the possibility of adding medication to the participant's care based on the severity of the case. The selection of medication was based on the Canadian Network for Mood and Anxiety Treatments guidelines, and these guidelines were referenced to decide the first or second line of treatment for each participant [60].

When deciding the appropriate intervention, a participant's care was not always increased sequentially; instead, a participant who exhibited severe symptoms by the end of session 4 was provided a high step as opposed to step 1. The decision about which intervention the participant would receive was dependent on a few factors considered by the care providers. These included (1) changes in participant's PHQ-9 scores, (2) engagement with treatment, (3) progress in weekly goals, and (4) homework submission.

Changes in PHQ-9 Scores

PHQ-9 scores for each participant were collected every 3 weeks (ie, weeks 1, 4, 7, 10, and 13). The initial step-up decision was based on the first 2 PHQ-9 scores collected at weeks 1 and 4. If the participant's PHQ-9 score at week 4 increased by >2 points compared with week 1, the participant was stepped up in their care to either step 1 (message) or step 2 (phone call). During the subsequent weeks, changes in the subsequent PHQ-9 scores were compared with the previous week's score (ie, increase of >2 points) to determine whether an additional intervention was required [61].

Treatment Engagement

A systematic review including 35 studies showed that great treatment engagement significantly improved postintervention mental health outcomes [62]. Thus, participant engagement with treatment through the OPTT platform was another factor that was considered when determining which intervention would be most suitable for the participant. If the participant exhibited limited interaction with the platform (eg, read receipts for messages indicating that messages were not seen when notifying participants about session availability) or expressed difficulties in navigating the platform, the participant was provided with the step-2 (phone call) intervention. This limited engagement indicated that step 1 (messages) would be ineffective because the participant was not viewing the intervention. Therefore, the step-2 (phone call) intervention was identified as a more effective modality than the step-1 (messages) intervention. In addition, if participants did not respond to step-1 (messages) intervention, step-2 (phone call) interventions were provided as a follow-up.

Homework Submission

Participants' answers to weekly homework were also considered as a factor when deciding their care intensity [63]. If participants indicated ongoing depressive symptoms that have not changed within the past 2 sessions, they were stepped up in care intensity. Similarly, if depressive symptoms did not improve within the past 2 weeks and if they did not find the current intervention of steps 1, 2, or 3 helpful, step 4 (psychiatrist support) was provided to the participant. Finally, if the care provider sensed any form of suicidal or homicidal ideation or any severe depressive behavior at any point, step 4 (psychiatrist support) was referred to the participant.

Goal Progress

Beginning in session 1, participants were encouraged to set a goal that they would like to achieve by the end of the program. Each week, participants were prompted to complete a small step toward the goal to help them achieve it. The weekly progress toward this goal rather than absolute achievement was used as another indication of whether a step-up in intervention was required [64]. Studies have shown some support for goal planning in positive treatment outcomes for mental health care through assisting the therapeutic relationship by building rapport and allowing for open communication [65,66]. If participants struggled with the progression in their weekly small step for >2 consecutive sessions, they were stepped up in their intervention.

Data Analysis

A combination of descriptive statistics, independent *t* tests, and ANOVA were used to determine any difference between the primary outcomes across the 2 groups. All analyses were performed at a 1-tailed significance level of α=.05. To determine any significant difference in the clinical questionnaire scores between the 2 groups, ANOVA was used. A 2×5 repeated-measures ANOVA was conducted for PHQ-9 scores, and a 2×3 repeated-measures ANOVA was performed for Q-LES-Q and QIDS scores. The time points at which meaningful differences appear were examined using the Bonferroni post hoc method. If Mauchly test of sphericity was significant, a Huynh-Feldt correction was applied. A chi-square test was conducted to compare treatment compliance across the 2 groups based on the number of participants completing all 13 sessions of the program. An independent samples *t* test was also performed to compare the number of sessions completed by participants in the 2 groups. In addition, an intention-to-treat (ITT) analysis was completed to assess the clinical outcomes of treatment on participants who withdrew prematurely. No intermediate analyses were performed, and all statistical analyses were completed after the trial. IBM SPSS Statistics (version 28) was used to conduct all analyses [67].

Results

Participants

Recruitment was initiated in May 2021 after receiving approval from the Queen's University Health Sciences and Affiliated Teaching Hospitals Research Ethics Board. Recruitment was conducted between May 2021 and June 2022, and 69 individuals were found to be eligible for this study. Table 2 displays the participant demographics. Of the 69 eligible individuals, 34 (49%) were randomized to the control group and only 28 (41%) initiated treatment. Of the 28 participants in the control group, 9 (32%) completed the full round of therapy (ie, 13 sessions of i-CBT). Of the 35 individuals randomized to the stepped care group, 28 (80%) initiated treatment. In the stepped care group, only 54% (15/28) of the participants completed all 13 sessions of the i-CBT program. Unfortunately, owing to some data collection errors, some questionnaire scores were not collected during the program, leading to some gaps in pretreatment, midtreatment, or posttreatment scores. These scores were treated as missing and were not imputed during data analysis. Figure 2 shows the participant flow.

Participants in the stepped care group were allocated to the stepped care intervention as follows: step 0 (8/28, 29%), step 1 (message; 2/28, 7%), step 2 (phone call; 15/28, 54%), step 3 (video call; 1/28, 4%), step 4 (psychiatrist consultation; 2/28, 7%). Overall, of the 28 participants, a total of 20 (71%) participants were stepped up in their care. This includes participants who dropped out prematurely and those who completed all 13 sessions. Of the 15 participants who completed all 13 sessions, 12 (80%) participants were stepped up in their care at some point during the program. Table 3 provides information regarding the number of occurrences of each step, and Table 4 provides a summary of each participant's progression in the stepped care program throughout the treatment, beginning at session 5. Currently, the 1-year follow-up period is ongoing, in which PHQ-9, QIDS, and Q-LES-Q scores are collected at the 3-, 6-, 9-, and 12-month follow-up periods after the treatment. The follow-up period is expected to be completed in November 2023.

Table 2. Demographics and characteristics of the participants who began treatment, categorized based on treatment group: internet-based cognitive behavioral therapy (i-CBT) and i-CBT with stepped care.

Characteristics	i-CBT (n=28)	Stepped care (n=28)
Age (y), mean (SD)	37.88 (13.08)	40.57 (14.28)
Baseline PHQ-9[a] score, mean (SD)	16.63 (4.40)	17.75 (5.33)
Sex, n (%)		
Female	17 (61)	17 (61)
Male	7 (25)	11 (39)
Other	2 (7)	0 (0)
Missing	2 (7)	0 (0)
Ethnicity, n (%)		
Asian	1 (4)	1 (4)
Hispanic	1 (4)	0 (0)
White	7 (25)	7 (25)
Other	17 (61)	20 (71)
Missing	2 (7)	0 (0)
First language, n (%)		
English	24 (86)	26 (93)
Hebrew	0 (0)	1 (4)
Hindi	0 (0)	1 (4)
Cantonese	1 (4)	0 (0)
Spanish	1 (4)	0 (0)
Missing	2 (7)	0 (0)
Immigration status, n (%)		
Born in Canada	12 (43)	13 (46)
Immigrated to Canada	3 (11)	2 (7)
Missing	13 (46)	13 (46)
Employment, n (%)		
Full time	13 (46)	15 (54)
Part time	3 (11)	4 (14)
Unemployed	6 (21)	8 (29)
Student	4 (14)	1 (4)
Missing	2 (7)	0 (0)
Marital status, n (%)		
Married	7 (25)	10 (36)
Never married	14 (50)	13 (36)
Divorced	1 (4)	2 (46)
Widowed	2 (7)	1 (7)
Other	2 (7)	2 (4)
Missing	2 (7)	0 (0)
Children, n (%)		
Yes	5 (18)	13 (46)
No	21 (75)	15 (54)
Missing	2 (7)	0 (0)

Characteristics	i-CBT (n=28)	Stepped care (n=28)
Income (CAD $; CAD $1=US 0.74), n (%)		
<20,000	7 (25)	6 (21)
20,000-34,999	5 (18)	3 (11)
35,000-49,999	5 (18)	8 (29)
50,000-74,999	4 (14)	6 (21)
75,000-99,999	0 (0)	2 (7)
>100,000	3 (11)	3 (11)
Missing	4 (14)	0 (0)

[a]PHQ-9: Patient Health Questionnaire–9.

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) flow diagram of participant flow through the study. i-CBT: internet-based cognitive behavioral therapy; MINI: Mini International Neuropsychiatric Interview; PHQ-9: Patient Health Questionnaire–9; QIDS: Quick Inventory of Depressive Symptomatology; Q-LES-Q: Quality of Life Enjoyment and Satisfaction Questionnaire–Short Form.

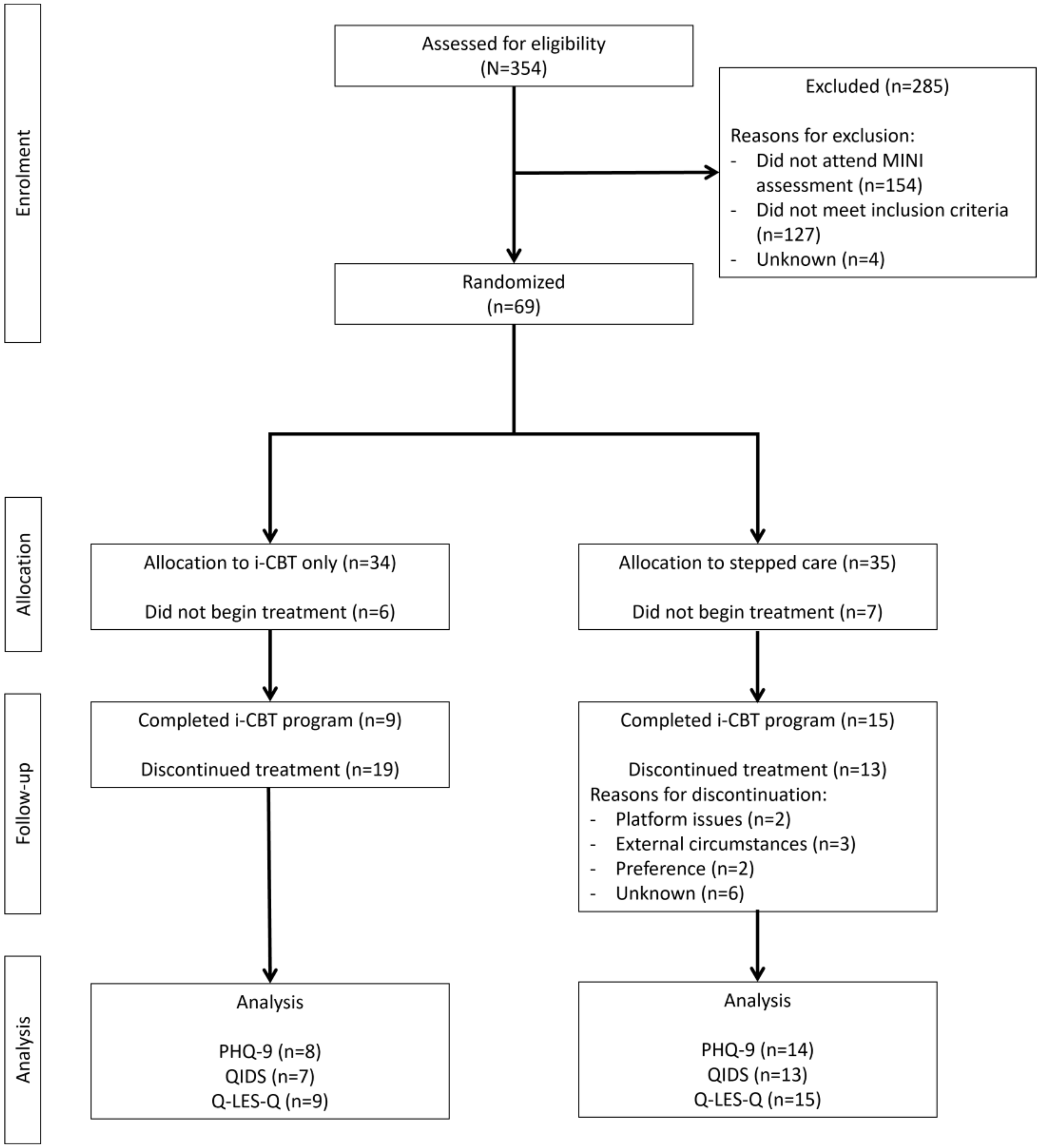


Table 3. Sample size of the highest intervention provided for participants in the stepped care group, sorted based on completion and dropout from the internet-based cognitive behavioral therapy (i-CBT) program.

Stepped care interventions	Step 0: i-CBT only, n (%)	Step 1: message, n (%)	Step 2: phone call, n (%)	Step 3: video call, n (%)	Step 4: psychiatrist, n (%)
Dropout (n=13)	5 (38)	1 (8)	6 (46)	0 (0)	1 (8)
Completed (n=15)	3 (20)	1 (7)	9 (60)	1 (7)	1 (7)
Total (n=28)	8 (29)	2 (7)	15 (54)	1 (4)	2 (7)

Table 4. The number of participants provided with each stepped care intervention for each session, from session 5 to session 13 (n=28).

Step	Session 5, n (%)	Session 6, n (%)	Session 7, n (%)	Session 8, n (%)	Session 9, n (%)	Session 10, n (%)	Session 11, n (%)	Session 12, n (%)	Session 13, n (%)
N/A[a]	3 (11)	5 (18)	6 (21)	6 (21)	8 (29)	9 (32)	11 (39)	13 (46)	13 (46)
0	16 (57)	17 (61)	16 (57)	15 (54)	12 (43)	15 (54)	10 (36)	12 (43)	8 (29)
1	4 (14)	1 (4)	2 (7)	1 (4)	4 (14)	0 (0)	3 (11)	1 (4)	2 (7)
2	5 (18)	5 (18)	4 (14)	6 (21)	4 (14)	2 (7)	2 (7)	1 (4)	3 (11)
3	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	2 (7)	1 (4)	1 (4)	1 (4)
4	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (4)	0 (0)	0 (0)

[a]N/A: not applicable.

Measured Outcomes

PHQ-9 Score

At baseline, for the participants who initiated the i-CBT program, the mean PHQ-9 score was 16.63 (SD 4.40; 27/28, 96%) for the control group and 17.75 (SD 5.33; 28/28, 100%) for the stepped care group, showing no statistically significant difference in pretreatment scores between the 2 groups (t_{53}=−0.85; *P*=.40; 95% CI −3.77 to –1.53). A 2×5 repeated-measures ANOVA determined that the mean PHQ-9 scores differed significantly between time points ($F_{4,80}$=9.95; *P*<.001; Figure 3), but there was no significant difference at different time points between the 2 groups ($F_{4,80}$=0.43; *P*=.78; Figure 4). Post hoc analysis with Bonferroni adjustment revealed that PHQ-9 scores significantly decreased from pretreatment period (week 1) to the second time point (week 4; 4.554, 95% CI 1.264-7.843; *P*=.003), from pretreatment period (week 1) to the third time point (week 7; 4.973, 95% CI 1.904-8.042; *P*<.001), from pretreatment period (week 1) to the fourth time point (week 10; 4.384, 95% CI 0.862-7.906; *P*=.008), and from pretreatment period (week 1) to posttreatment period (week 13; 5.357, 95% CI 2.169-8.546; *P*<.001). All between–time point scores were statistically significant (Figure 3). A detailed breakdown of the results are summarized in Multimedia Appendix 4.

Figure 3. Estimated marginal means of Patient Health Questionnaire–9 (PHQ-9) scores at 5 treatment time intervals corresponding to sessions 1, 4, 7, 10, and 13, including both groups—internet-based cognitive behavioral therapy (i-CBT) only and i-CBT with stepped care. Error bars depict –2 or +2 SE.

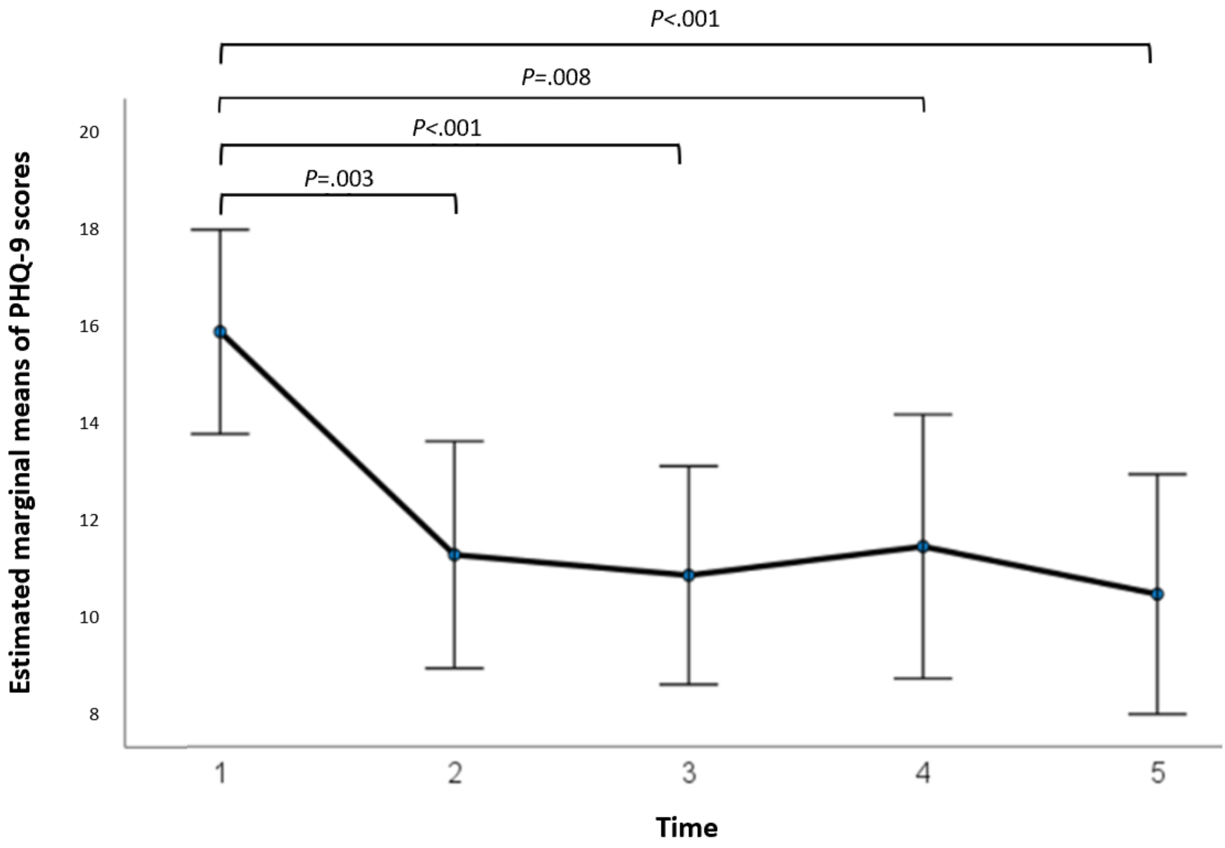
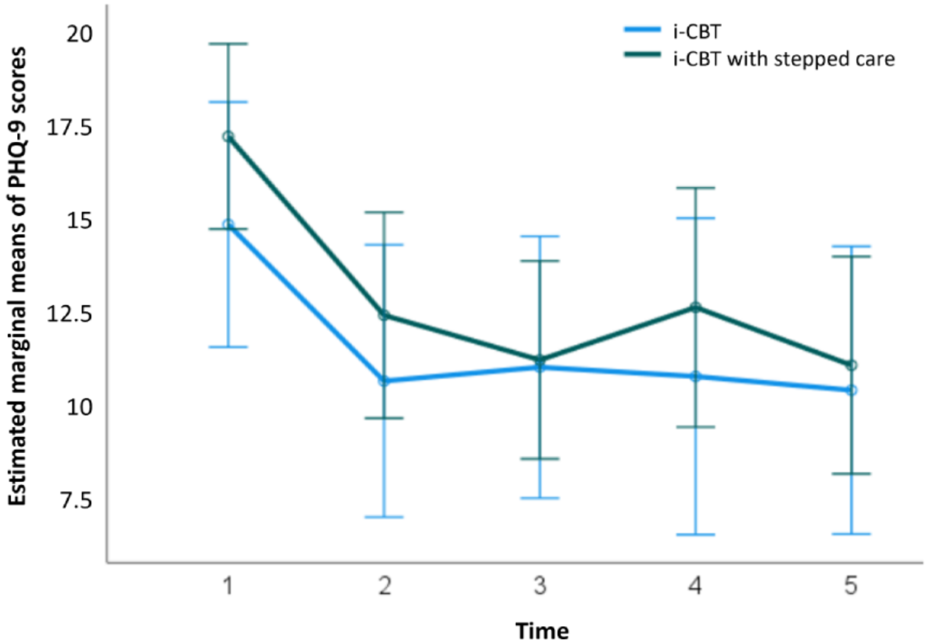


Figure 4. Estimated marginal means of Patient Health Questionnaire–9 (PHQ-9) scores at 5 intervals of treatment (sessions 0, 4, 7, 10, and 13), in both internet-based cognitive behavioral therapy (i-CBT; blue) and i-CBT with stepped care (green) treatment conditions. Error bars depict –2 or +2 SE.



QIDS Score

At baseline, for the participants who initiated the i-CBT program, the mean QIDS score was 14.62 (SD 4.64; 21/28, 75%) for the control group and 16.83 (SD 4.43; 23/28, 82%) for the stepped care group, showing no statistical significance in pretreatment scores between the 2 groups (t_{42}=−1.61; *P*=.11; 95% CI −4.97 to 0.55; independent samples *t* test). A 2×3 repeated-measures ANOVA determined that the mean QIDS scores differed significantly between time points ($F_{2,28}$=5.73; *P*=.008; Figure 5), but there was no significant difference at different time points between the 2 groups ($F_{2,28}$=3.05; *P*=.06; Figure 6). Post hoc analysis with Bonferroni adjustment revealed that QIDS scores significantly decreased from pretreatment period (week 1) to posttreatment period (week 13; 4.12, 95% CI 0.76-7.47; *P*=.02). There was no significant difference between pretreatment period (week 1) and midtreatment period (week 7; 2.08, 95% CI −1.10 to 5.26; *P*=.29) or between midtreatment period (week 7) and posttreatment period (week 13; 2.03, 95% CI −1.34 to 5.41; *P*=.37; Figure 5). A detailed breakdown of the results are summarized in Multimedia Appendix 4.

Figure 5. Estimated marginal means of Quick Inventory of Depressive Symptomatology (QIDS) scores at 3 treatment time intervals corresponding to sessions 1, 7, and 13, including both groups—internet-based cognitive behavioral therapy (i-CBT) only and i-CBT with stepped care. Error bars depict –2 or +2 SE.

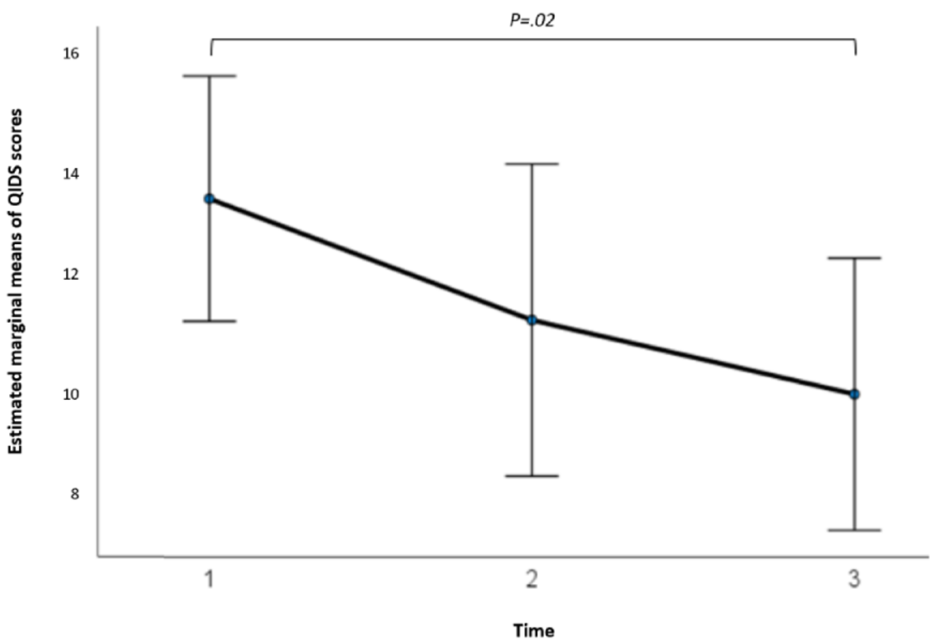
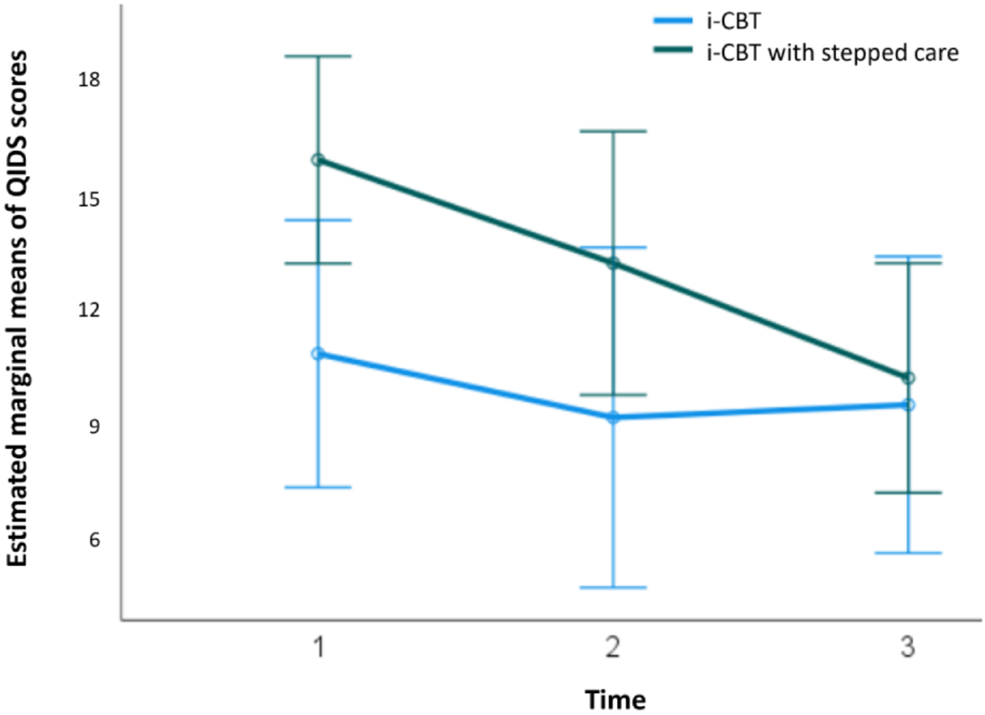


Figure 6. Estimated marginal means of Quick Inventory of Depressive Symptomatology (QIDS) scores at 3 intervals of treatment (sessions 0, 7, and 13), in both internet-based cognitive behavioral therapy (i-CBT; blue) and i-CBT with stepped care (green) treatment conditions. Error bars depict –2 or +2 SE.



Q-LES-Q Score

At baseline, for the participants who initiated the i-CBT program, the mean Q-LES-Q score was 36.88 (SD 7.80; 26/28, 93%) for the control group and 36.25 (SD 7.81; 28/28, 100%) for the stepped care group, showing no statistical significance in pretreatment scores between the 2 groups (t_{52}=0.30; *P*=.77; 95% CI −3.63 to 4.90; independent samples *t* test). A 2×3 repeated-measures ANOVA determined that the mean Q-LES-Q scores differed significantly between time points ($F_{2,38}$=4.18; *P*=.02; Figure 7), but there was no significant difference at different time points between the 2 groups ($F_{2,38}$=0.19; *P*=.83; Figure 8). Post hoc analysis with Bonferroni adjustment revealed that Q-LES-Q scores significantly increased from pretreatment period (week 1) to midtreatment period (week 7; −5.50, 95% CI −10.91 to –0.94; *P*=.045). There were no significant differences between pretreatment period (week 1) and posttreatment period (week 13; −4.77, 95% CI −11.03 to 1.49; *P*=.18) and between midtreatment period (week 7) and posttreatment period (week 13; 0.73, 95% CI −5.18 to 3.72; *P*=.99; Figure 7). A detailed breakdown of the results are summarized in Multimedia Appendix 4.

Figure 7. Estimated marginal means of Quality of Life Enjoyment and Satisfaction Questionnaire–Short Form (Q-LES-Q) scores at 3 treatment time intervals corresponding to sessions 1, 7, and 13, including both groups—internet-based cognitive behavioral therapy (i-CBT) only and i-CBT with stepped care. Error bars depict –2 or +2 SE.

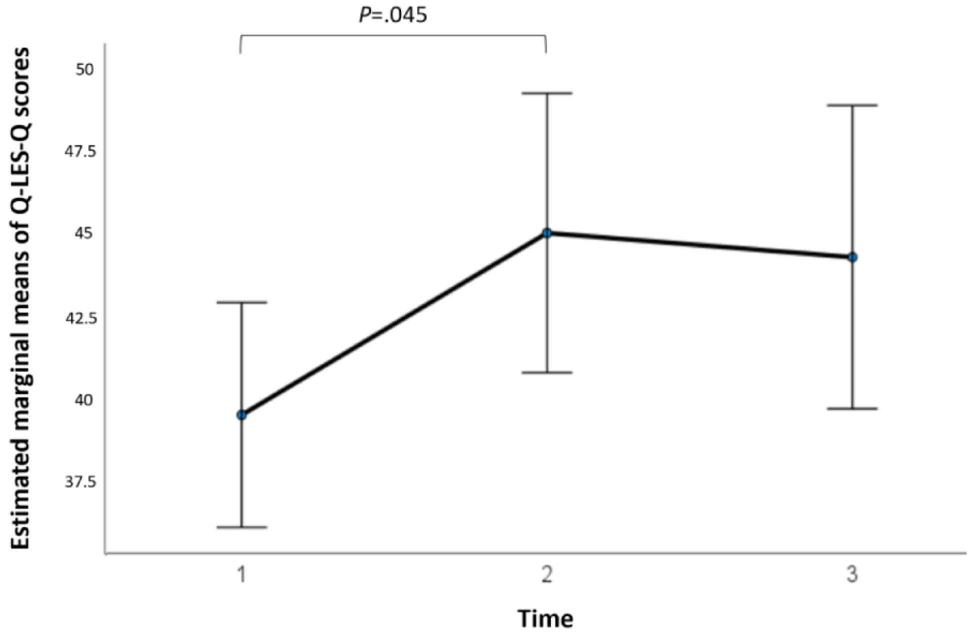
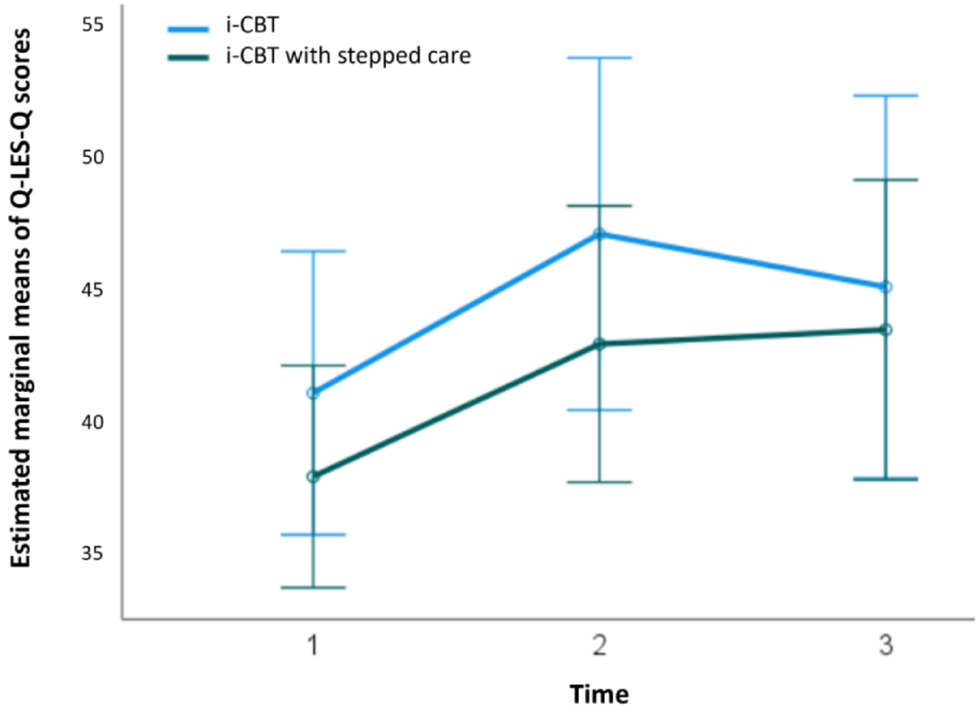


Figure 8. Estimated marginal means of Quality of Life Enjoyment and Satisfaction Questionnaire–Short Form (Q-LES-Q) scores at 3 intervals of treatment (sessions 0, 7, and 13), in both internet-based cognitive behavioral therapy (i-CBT; blue) and i-CBT with stepped care (green) treatment conditions. Error bars depict –2 or +2 SE.



Treatment Compliance

The proportion of participants completing the full round of therapy (ie, 13 sessions) was 32% (9/28) in the i-CBT–only group and 54% (15/28) in the stepped care group. There was no significant difference between the 2 groups regarding completion of the i-CBT program for depression (χ^2_1=2.6; *P*=.10; chi-square test). An independent samples *t* test was conducted to determine whether there was a difference in the number of completed sessions between the i-CBT group and the stepped care group for the individuals who had begun the program. The results indicate significant difference between the average number of sessions completed by participants in the i-CBT–only group (mean 7, SD 4.64; 28/28, 100%) and stepped care group (mean 9.41, SD 4.44; 28/28, 100%; t_{55}=−2; *P*=.03; 95% CI −4.83 to −0.002; independent samples *t* test). Multimedia Appendix 5 shows the distribution of the sessions completed in each group.

ITT Analysis

A linear mixed model analysis was used to conduct the intent-to-treat analysis for each participant across the program. The group (i-CBT or stepped care) and evaluation time points were indicated as fixed factors. This analysis included participants who did not complete the full program and dropped out of the study before completion. As seen previously, there was significant change in PHQ-9 scores across the time points ($F_{4,48.79}$=10.98; P<.001), but there was no significant difference between the 2 groups ($F_{1,107.07}$=2.79; P=.10). In addition, for the QIDS scores, there was significant change in scores across the time points ($F_{2,41.08}$=8.53; P<.001), but no significant difference between the 2 groups ($F_{1,54.44}$=2.72; P=.10). Q-LES-Q scores imitated a similar pattern, with significant difference in scores across the time points ($F_{2,44.99}$=9.36; P<.001), but no significant difference between the 2 groups ($F_{1,60.75}$=3.28; P=.08).

Discussion

Principal Findings

This study evaluated the effectiveness of an i-CBT program with and without stepped care for depression among adults. The results of this study indicate that the proposed stepped care model was not significantly better in reducing depressive symptoms, as measured by PHQ-9 and QIDS, than the i-CBT program alone. Some previous studies have also found no significant differences between stepped care treatments and care as usual in improving depressive symptoms [52,68]. Ho et al [52], who reviewed stepped care for both depressive and anxiety disorders, found that stepped care treatment was significantly better in improving anxiety symptoms but did not find any significant difference in reducing depressive symptoms. This suggests some differences in the populations and the sensitivity of the stepped care structure in addressing different mental health issues. When viewing the results of both groups together, this study indicated that the i-CBT program itself was effective in significantly reducing depressive symptoms from pretreatment period to posttreatment period, as measured by the PHQ-9 and QIDS; however, there were no significant differences in the reduction of depressive symptoms between the 2 groups. The stepped care group did not show to be significantly better in reducing depressive symptoms than the i-CBT group. This finding is consistent with other studies of i-CBT programs, which show improvements in depressive symptoms for mild to moderate depression [69,70]. This also reinforces the efficacy of the i-CBT program for depression on OPTT used in this study, based on similar results from previous clinical trials using this program [47,49].

Furthermore, the results suggest that the i-CBT program indicated significant difference in the quality of life, as measured by Q-LES-Q; however, post hoc analysis revealed that overall, there were no significant differences before and after treatment in improving the quality of life. Significant difference was observed between pretreatment period and midtreatment period, which may suggest that the i-CBT program had the greatest impact on improving quality of life in the early stages of the program. No significant differences were observed between the i-CBT and stepped care groups. A systematic review and meta-analysis reviewed 3 studies of the effects of i-CBT on the quality of life and found inconclusive evidence of i-CBT compared with in-person outcomes [71]. We found no significant difference between the effects of i-CBT with stepped care and i-CBT without stepped care on the quality of life.

Although the completion rates between the 2 groups were not significantly different, participants in the stepped care group significantly completed, on average, 2 more sessions of therapy than those in the i-CBT group. The increase in the number of completed sessions may have been caused by the added support offered in the stepped care group. This is consistent with previous studies that report improvement in treatment adherence in mental health treatment with additional interventions in care [21,28,63]. Specifically, phone and video calls assist with keeping patients in care compared with in-person treatment [36]. Compared with in-person treatment, receiving care for depression over the phone [56,72] or via video [22] is associated with high completion rates and few dropouts. Further studies are required to decipher whether the stepped care group's high completion rate of sessions is associated with better treatment outcomes than the i-CBT group.

Furthermore, Multimedia Appendix 5 shows that although participant dropout occurs uniformly across the first 10 sessions in the stepped care group, most participants (19/28, 68%) in the i-CBT group dropped out in the first half of the program (ie, first 7 sessions): 25% (7/28) of participants in the stepped care group dropped out of the program before the midpoint of treatment (ie, session 7) and 68% (19/28) of participants in the i-CBT group. Previous studies have noted that most patients drop out of treatment programs after 2 to 4 sessions [16,31,73]. During this time, participants were provided with i-CBT care only in both groups (additional interventions for the stepped care group were introduced at session 5 after monitoring their PHQ-9 scores and interaction with the i-CBT program); however, introducing a possibility of intervention soon in the stepped care group may have helped participants who were indicated as an early dropout (ie, before session 5) complete more sessions. Concurrently, it is noteworthy that the i-CBT program's 13 sessions were not completed by several participants (32/56, 57%) in both groups. This is consistent with previous studies that reveal high dropout rates in depression i-CBT programs. Compliance is generally a challenge for i-CBT programs, with dropout rates averaging approximately 32% and ranging between 0% and 75% [26,37], whereas traditional CBT dropout rates are approximately 25% on average, ranging between 0% and 68% [74]. However, another study found that additional human support results in a large effect size regarding the efficacy of i-CBT (g=0.673) than providing no additional support (g=0.239) [34]. Thus, it was expected that the stepped care group would exhibit significant improvements following treatment, but this was not observed in this study. This may be owing to the nature of the study design, as the i-CBT group received some support during the treatment through the homework feedback and did not encompass a truly no-support intervention. Upon reviewing participant progression in the stepped care model (Table 4), it was seen that on average, most participants (13/28, 46%) spent their time in step 0 of the stepped care model, which mimics the control group (i-CBT only). This may be a factor in the observed nonsignificant differences between the 2 groups because not many participants were stepped up in their care across treatment. Moreover, the small sample size may have influenced the outcomes. A previous study has shown that patients who terminate CBT prematurely show high symptom severity compared with patients who complete therapy, but the 2 groups did not vary in the rate of symptom change [75]. This suggests that the increase in symptom severity may arise from completing few sessions of treatment in the dropout group compared with the completer group and explains the differing results observed between the 2 groups as dropout was high. Given the effectiveness of i-CBT treatment and the high dropout rates, it is important to assist patients in completing more sessions by providing sufficient resources and care. Our proposed stepped care model assisted participants in completing a great number of sessions. Stepped care has been associated with high treatment satisfaction, which may have assisted in participants completing more sessions in this group [68,76]. Future studies should examine the variables that may be responsible for treatment attrition in i-CBT programs and devise methods to raise the rates of treatment engagement and completion.

Limitations

It is important to consider further limitations of this study. This includes the relatively small sample size; observed sex imbalance, with 61% (34/56) women in the study (approximately 15% more women in the i-CBT group compared with the stepped care group); and lack of long-term follow-up, which is currently ongoing. It is important to note that the sample size for analysis of the i-CBT group was smaller than that of the stepped care group. In addition, the participants in this study may not be representative of the general population, as they were predominantly women, English speaking (owing to the limitations of the i-CBT program), employed full time, and recruited from a specific clinical setting of self-referrals and clinics limited to Kingston, Ontario (Table 2). Missing data were also a challenge that affected the total sample size. A large portion of the data was unavailable owing to collection errors and dropouts (Figure 2). We attempted to address this issue by using ITT analysis and observed similar results with no significant differences between groups but observed significant differences across various time points for PHQ-9, QIDS, and Q-LES-Q scores.

This study was not specifically designed to investigate the effects of the different treatment interventions in the stepped care group and how they influenced the results. In hindsight, using 4 different factors to decide about the stepped intervention results in increased variability in the stepped care decision and limits our ability to decipher the effectiveness of the approach. Most stepped care models use clinical questionnaire scores, such as the PHQ-9 score, to decide when to step up or step down care [68,76-78]; however, our study did not include strict cutoff guidelines and instead adopted a subjective monitoring approach of PHQ-9 scores as one of the evaluating factors. Changes in PHQ-9 scores (scale ranging from 0-27) and homework submission (submitted or not submitted) can be quantized predictors for stepped care; however, monitoring engagement with the OPTT platform and goal progression are variables that are multifactorial and are subject to interpretation. This design prevented us from making conclusions about a best-fit intervention model. Furthermore, beginning the stepped care interventions in session 5 allowed for a watchful waiting period that allowed reflection about patient status to determine the correct intensity of care if they did not improve with a low-resource intervention (ie, i-CBT). However, this may be a limitation of the proposed stepped care model, as the waiting period may be harmful because it may delay optimal treatment [68]. With a small sample size, it is difficult to analyze the effects of these deciding factors. Future studies need to be conducted to explore the effects of the individual stepped interventions provided in this study and any potential relationships among the modalities of care.

Therapeutic alliance in psychotherapy is another significant factor in predicting treatment outcomes [79-81]. The i-CBT–only group had limited interaction with their assigned care provider, and the ability to build rapport with participants was limited. In the stepped care group, participants were potentially able to gain a deep connection with their care provider, beginning in session 5, based on the intervention provided to them. Therapeutic alliance follows two important phases: (1) initial alliance development, usually occurring within the first 5 sessions, and (2) challenging the patient more actively, which can cause strain to the therapeutic relationship [79]. With stepped care interventions being provided in session 5, the first phase seems to be neglected and may affect the potential relationship, thus the structure of the program hinders participants from achieving the desired treatment outcome and reducing their depressive symptoms. Some studies demonstrate that a positive outcome is more predictive by the quality of the alliance rather than the type of the intervention [79,82-84]. In this case, the quality of the alliance was affected by the limited interaction with the care provider and may reflect one of the limitations of online therapy. Some studies found that building therapeutic alliances face to face is significantly more effective than online psychological treatments [85-89]. In contrast, other studies have found the opposite and report no correlation between the therapeutic alliance and treatment outcomes in digital interventions [90-93]. Further studies are required to better understand the characteristics of therapeutic alliance in digital contexts [25,94]. It would also be interesting to assess participant and care provider opinions about their therapeutic alliance during the study to better understand this limitation.

Conclusions

This study provides further evidence that i-CBT is effective for treating depressive symptoms. However, the study showed no significant evidence that the proposed stepped care model was more effective than i-CBT alone. The stepped care model allowed participants to complete 2 more sessions on average than the i-CBT–only group, indicating that stepped care is an effective method for guiding patients to treatment completion. Future studies should examine the long-term effects of such interventions and the efficacy of specific stepped care interventions in large and more diverse groups. To improve the design and implementation of such a model, studies might also investigate the processes through which stepped care interventions reduce depression symptoms and enhance the quality of life.

Acknowledgments

The authors would like to thank all the care providers involved in this program for ensuring that participants were supported throughout the study. The authors are also grateful to the participants of this study who provided their time and trust in the process to help the authors complete and further investigate this online cognitive behavioral therapy approach.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

NA and MO have cofounded Online Psychotherapy Tool (OPTT), the platform used to provide care to participants. They have ownership stakes in OPTT Inc.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 7464 KB - mental_v11i1e51704_app1.pdf]

Multimedia Appendix 2
Sample of the e–cognitive behavioral therapy program for depression used in the study, showcasing session 5.
[PDF File (Adobe PDF File), 3875 KB - mental_v11i1e51704_app2.pdf]

Multimedia Appendix 3
Feedback templates for e–cognitive behavioral therapy used throughout the study by care providers including feedback for session homework and live interactions between the care provider and participants.
[DOCX File , 20 KB - mental_v11i1e51704_app3.docx]

Multimedia Appendix 4
Statistical analysis results.
[DOCX File , 31 KB - mental_v11i1e51704_app4.docx]

Multimedia Appendix 5
The frequency of participants' last completed sessions organized by 2 groups—internet-based cognitive behavioral therapy (i-CBT) only and i-CBT with stepped care.
[PNG File , 62 KB - mental_v11i1e51704_app5.png]

References

1. GBD results. Institute for Health Metrics and Evaluation. URL: https://vizhub.healthdata.org/gbd-results [accessed 2023-09-25]
2. Vos T, Lim SS, Abbafati C, Abbas KM, Abbasi M, Abbasifard M. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet 2020 Oct;396(10258):1204-1222 [FREE Full text] [doi: 10.1016/s2215-0366(21)00395-3]
3. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders: DSM-5. 5th edition. Washington, DC: American Psychiatric Association; 2013.
4. Cho Y, Lee JK, Kim DH, Park JH, Choi M, Kim HJ, et al. Factors associated with quality of life in patients with depression: a nationwide population-based study. PLoS One 2019 Jul;14(7):e0219455 [FREE Full text] [doi: 10.1371/journal.pone.0219455] [Medline: 31295291]
5. Engel L, Chen G, Richardson J, Mihalopoulos C. The impact of depression on health-related quality of life and wellbeing: identifying important dimensions and assessing their inclusion in multi-attribute utility instruments. Qual Life Res 2018 Nov;27(11):2873-2884 [FREE Full text] [doi: 10.1007/s11136-018-1936-y] [Medline: 30006664]
6. Heu U, Bogren M, Wang AG, Brådvik L. Aspects of additional psychiatric disorders in severe depression/melancholia: a comparison between suicides and controls and general pattern. Int J Environ Res Public Health 2018 Jun 21;15(7):1299 [FREE Full text] [doi: 10.3390/ijerph15071299] [Medline: 29933593]

7. Lake J, Turner MS. Urgent need for improved mental health care and a more collaborative model of care. Perm J 2017;21:17-024 [FREE Full text] [doi: 10.7812/TPP/17-024] [Medline: 28898197]
8. Mekonen T, Chan GC, Connor JP, Hides L, Leung J. Estimating the global treatment rates for depression: a systematic review and meta-analysis. J Affect Disord 2021 Dec 01;295:1234-1242 [FREE Full text] [doi: 10.1016/j.jad.2021.09.038] [Medline: 34665135]
9. Karyotaki E, Riper H, Twisk J, Hoogendoorn A, Kleiboer A, Mira A, et al. Efficacy of self-guided internet-based cognitive behavioral therapy in the treatment of depressive symptoms: a meta-analysis of individual participant data. JAMA Psychiatry 2017 Apr 01;74(4):351-359 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.0044] [Medline: 28241179]
10. Myhr G, Payne K. Cost-effectiveness of cognitive-behavioural therapy for mental disorders: implications for public health care funding policy in Canada. Can J Psychiatry 2006 Sep;51(10):662-670 [FREE Full text] [doi: 10.1177/070674370605101006] [Medline: 17052034]
11. Webb CA, Rosso IM, Rauch SL. Internet-based cognitive-behavioral therapy for depression: current progress and future directions. Harv Rev Psychiatry 2017 May;25(3):114-122 [FREE Full text] [doi: 10.1097/HRP.0000000000000139] [Medline: 28475503]
12. Coombs NC, Meriwether WE, Caringi J, Newcomer SR. Barriers to healthcare access among U.S. adults with mental health challenges: a population-based study. SSM Popul Health 2021 Sep;15:100847 [FREE Full text] [doi: 10.1016/j.ssmph.2021.100847] [Medline: 34179332]
13. Mental health care needs, 2018. Government of Canada. 2019. URL: https://www150.statcan.gc.ca/n1/pub/82-625-x/2019001/article/00011-eng.htm [accessed 2022-02-06]
14. Levesque JF, Harris MF, Russell G. Patient-centred access to health care: conceptualising access at the interface of health systems and populations. Int J Equity Health 2013 Mar 11;12:18 [FREE Full text] [doi: 10.1186/1475-9276-12-18] [Medline: 23496984]
15. Gratzer D, Khalid-Khan F. Internet-delivered cognitive behavioural therapy in the treatment of psychiatric illness. CMAJ 2016 Mar 01;188(4):263-272 [FREE Full text] [doi: 10.1503/cmaj.150007] [Medline: 26527829]
16. Richards D, Richardson T. Computer-based psychological treatments for depression: a systematic review and meta-analysis. Clin Psychol Rev 2012 Jun;32(4):329-342 [FREE Full text] [doi: 10.1016/j.cpr.2012.02.004] [Medline: 22466510]
17. Andersson G, Cuijpers P. Internet-based and other computerized psychological treatments for adult depression: a meta-analysis. Cogn Behav Ther 2009;38(4):196-205 [FREE Full text] [doi: 10.1080/16506070903318960] [Medline: 20183695]
18. Musiat P, Tarrier N. Collateral outcomes in e-mental health: a systematic review of the evidence for added benefits of computerized cognitive behavior therapy interventions for mental health. Psychol Med 2014 Nov;44(15):3137-3150. [doi: 10.1017/S0033291714000245] [Medline: 25065947]
19. Saddichha S, Al-Desouki M, Lamia A, Linden IA, Krausz M. Online interventions for depression and anxiety - a systematic review. Health Psychol Behav Med 2014 Jan 01;2(1):841-881 [FREE Full text] [doi: 10.1080/21642850.2014.945934] [Medline: 25750823]
20. Sztein DM, Koransky CE, Fegan L, Himelhoch S. Efficacy of cognitive behavioural therapy delivered over the Internet for depressive symptoms: a systematic review and meta-analysis. J Telemed Telecare 2018 Sep;24(8):527-539. [doi: 10.1177/1357633X17717402] [Medline: 28696153]
21. Mohr DC, Lattie EG, Tomasino KN, Kwasny MJ, Kaiser SM, Gray EL, et al. A randomized noninferiority trial evaluating remotely-delivered stepped care for depression using internet cognitive behavioral therapy (CBT) and telephone CBT. Behav Res Ther 2019 Dec;123:103485 [FREE Full text] [doi: 10.1016/j.brat.2019.103485] [Medline: 31634738]
22. Shabahang R, Aruguete MS, McCutcheon L. Video-based cognitive-behavioral intervention for COVID-19 anxiety: a randomized controlled trial. Trends Psychiatry Psychother 2021 Apr;43(2):141-150 [FREE Full text] [doi: 10.47626/2237-6089-2020-0056] [Medline: 34043898]
23. Kelders SM, Kok RN, Ossebaard HC, Van Gemert-Pijnen JE. Persuasive system design does matter: a systematic review of adherence to web-based interventions. J Med Internet Res 2012 Nov 14;14(6):e152 [FREE Full text] [doi: 10.2196/jmir.2104] [Medline: 23151820]
24. Sigurðardóttir S, Helgadóttir FD, Menzies RE, Sighvatsson MB, Menzies RG. Improving adherence to a web-based cognitive-behavioural therapy program for social anxiety with group sessions: a randomised control trial. Internet Interv 2022 Apr;28:100535 [FREE Full text] [doi: 10.1016/j.invent.2022.100535] [Medline: 35433276]
25. Tremain H, McEnery C, Fletcher K, Murray G. The therapeutic alliance in digital mental health interventions for serious mental illnesses: narrative review. JMIR Ment Health 2020 Aug 07;7(8):e17204 [FREE Full text] [doi: 10.2196/17204] [Medline: 32763881]
26. Fernandez E, Salem D, Swift JK, Ramtahal N. Meta-analysis of dropout from cognitive behavioral therapy: magnitude, timing, and moderators. J Consult Clin Psychol 2015 Dec;83(6):1108-1122. [doi: 10.1037/ccp0000044] [Medline: 26302248]
27. Christensen H, Griffiths KM, Farrer L. Adherence in internet interventions for anxiety and depression: systematic review. J Med Internet Res 2009 Apr 24;11(2):e13 [FREE Full text] [doi: 10.2196/jmir.1194] [Medline: 19403466]
28. Bower P, Gilbody S. Stepped care in psychological therapies: access, effectiveness and efficiency. narrative literature review. Br J Psychiatry 2005 Jan;186:11-17 [FREE Full text] [doi: 10.1192/bjp.186.1.11] [Medline: 15630118]

29. Cross SP, Hickie I. Transdiagnostic stepped care in mental health. Public Health Res Pract 2017 Apr 27;27(2):2721712 [FREE Full text] [doi: 10.17061/phrp2721712] [Medline: 28474049]
30. Delgadillo J, Ali S, Fleck K, Agnew C, Southgate A, Parkhouse L, et al. Stratified care vs stepped care for depression: a cluster randomized clinical trial. JAMA Psychiatry 2022 Feb 01;79(2):101-108 [FREE Full text] [doi: 10.1001/jamapsychiatry.2021.3539] [Medline: 34878526]
31. Nicholas J, Ringland KE, Graham AK, Knapp AA, Lattie EG, Kwasny MJ, et al. Stepping up: predictors of 'stepping' within an iCBT stepped-care intervention for depression. Int J Environ Res Public Health 2019 Nov 25;16(23):4689 [FREE Full text] [doi: 10.3390/ijerph16234689] [Medline: 31775297]
32. Kobak KA, Wolitzky-Taylor K, Craske MG, Rose RD. Therapist training on cognitive behavior therapy for anxiety disorders using internet-based technologies. Cognit Ther Res 2017 Apr;41(2):252-265 [FREE Full text] [doi: 10.1007/s10608-016-9819-4] [Medline: 28435174]
33. Applebaum AJ, DuHamel KN, Winkel G, Rini C, Greene PB, Mosher CE, et al. Therapeutic alliance in telephone-administered cognitive-behavioral therapy for hematopoietic stem cell transplant survivors. J Consult Clin Psychol 2012 Oct;80(5):811-816 [FREE Full text] [doi: 10.1037/a0027956] [Medline: 22468908]
34. Wright JH, Owen JJ, Richards D, Eells TD, Richardson T, Brown GK, et al. Computer-assisted cognitive-behavior therapy for depression: a systematic review and meta-analysis. J. Clin. Psychiatry 2019 Mar 19;80(2):18r12188 [FREE Full text] [doi: 10.4088/jcp.18r12188]
35. Dedert E, McDuffie JR, Swinkels C, Shaw R, Fulton J, Allen KD, et al. Computerized cognitive behavioral therapy for adults with depressive or anxiety disorders. Department of Veterans Affairs, Government of US. 2013. URL: https://www.ncbi.nlm.nih.gov/books/NBK269001/pdf/Bookshelf_NBK269001.pdf [accessed 2024-01-18]
36. Chen PV, Helm A, Caloudas SG, Ecker A, Day G, Hogan J, et al. Evidence of phone vs video-conferencing for mental health treatments: a review of the literature. Curr Psychiatry Rep 2022 Oct;24(10):529-539 [FREE Full text] [doi: 10.1007/s11920-022-01359-8] [Medline: 36053400]
37. Schmidt ID, Forand NR, Strunk DR. Predictors of dropout in internet-based cognitive behavioral therapy for depression. Cognit Ther Res 2019 Jun;43(3):620-630 [FREE Full text] [doi: 10.1007/s10608-018-9979-5] [Medline: 32879540]
38. Stephenson C, Moghimi E, Gutierrez G, Jagayat J, Layzell G, Patel C, et al. User experiences of an online therapist-guided psychotherapy platform, OPTT: a cross-sectional study. Internet Interv 2023 Apr;32:100623 [FREE Full text] [doi: 10.1016/j.invent.2023.100623] [Medline: 37273941]
39. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
40. Omrani M, Alavi N, Rivera M, Khalid-Khan S. Online clinic, a new method of delivering psychotherapy. Eur Psychiatry 2017;41(S1):S148 [FREE Full text] [doi: 10.1016/j.eurpsy.2017.01.1997]
41. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatric Annals 2002 Sep;32(9):509-515 [FREE Full text] [doi: 10.3928/0048-5713-20020901-06]
42. Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, et al. The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry 2003 Sep 01;54(5):573-583 [FREE Full text] [doi: 10.1016/s0006-3223(02)01866-8] [Medline: 12946886]
43. Endicott J, Nee J, Harrison W, Blumenthal R. Quality of life enjoyment and satisfaction questionnaire: a new measure. Psychopharmacol Bull 1993;29(2):321-326. [Medline: 8290681]
44. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 1998;59 Suppl 20:22-57. [Medline: 9881538]
45. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007 May;39(2):175-191 [FREE Full text] [doi: 10.3758/bf03193146] [Medline: 17695343]
46. Alavi N, Stephenson C, Yang M, Kumar A, Shao Y, Miller S, et al. Feasibility and efficacy of delivering cognitive behavioral therapy through an online psychotherapy tool for depression: protocol for a randomized controlled trial. JMIR Res Protoc 2021 Jun 16;10(6):e27489 [FREE Full text] [doi: 10.2196/27489] [Medline: 33990076]
47. Alavi N, Hirji A, Sutton C, Naeem F. Online CBT is effective in overcoming cultural and language barriers in patients with depression. J Psychiatr Pract 2016 Jan;22(1):2-8. [doi: 10.1097/PRA.0000000000000119] [Medline: 26813483]
48. Alavi N, Hirji A. The efficacy of powerpoint-based CBT delivered through email: breaking the barriers to treatment for generalized anxiety disorder. J Psychiatr Pract 2020 Mar;26(2):89-100. [doi: 10.1097/PRA.0000000000000455] [Medline: 32134882]
49. Alavi N, Moghimi E, Stephenson C, Gutierrez G, Jagayat J, Kumar A, et al. Comparison of online and in-person cognitive behavioral therapy in individuals diagnosed with major depressive disorder: a non-randomized controlled trial. Front Psychiatry 2023 Apr 28;14:1113956 [FREE Full text] [doi: 10.3389/fpsyt.2023.1113956] [Medline: 37187863]

50. Alavi N, Omrani M. Online Cognitive Behavioral Therapy: An e-Mental Health Approach to Depression and Anxiety. Cham, Switzerland: Springer; 2019.
51. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014 Mar 07;348(mar07 3):g1687 [FREE Full text] [doi: 10.1136/bmj.g1687] [Medline: 24609605]
52. Ho FY, Yeung WF, Ng TH, Chan CS. The efficacy and cost-effectiveness of stepped care prevention and treatment for depressive and/or anxiety disorders: a systematic review and meta-analysis. Sci Rep 2016 Jul 05;6:29281 [FREE Full text] [doi: 10.1038/srep29281] [Medline: 27377429]
53. Aguilera A, Muñoz RF. Text messaging as an adjunct to CBT in low-income populations: a usability and feasibility pilot study. Prof Psychol Res Pr 2011 Dec 01;42(6):472-478 [FREE Full text] [doi: 10.1037/a0025499] [Medline: 25525292]
54. Alfonsson S, Englund J, Parling T. Tailored text message prompts to increase therapy homework adherence: a single-case randomised controlled study. Behav Change 2019 Apr 11;36(3):180-191 [FREE Full text] [doi: 10.1017/bec.2019.10]
55. Corruble E, Swartz HA, Bottai T, Vaiva G, Bayle F, Llorca PM, et al. Telephone-administered psychotherapy in combination with antidepressant medication for the acute treatment of major depressive disorder. J Affect Disord 2016 Jan 15;190:6-11 [FREE Full text] [doi: 10.1016/j.jad.2015.07.052] [Medline: 26480205]
56. Fann JR, Bombardier CH, Vannoy S, Dyer J, Ludman E, Dikmen S, et al. Telephone and in-person cognitive behavioral therapy for major depression after traumatic brain injury: a randomized controlled trial. J Neurotrauma 2015 Jan 01;32(1):45-57 [FREE Full text] [doi: 10.1089/neu.2014.3423] [Medline: 25072405]
57. Simon GE, Ludman EJ, Tutty S, Operskalski B, Von Korff MV. Telephone psychotherapy and telephone care management for primary care patients starting antidepressant treatment: a randomized controlled trial. JAMA 2004 Aug 25;292(8):935-942 [FREE Full text] [doi: 10.1001/jama.292.8.935] [Medline: 15328325]
58. Ho FY, Chung KF, Yeung WF, Ng TH, Cheng SK. Weekly brief phone support in self-help cognitive behavioral therapy for insomnia disorder: relevance to adherence and efficacy. Behav Res Ther 2014 Dec;63:147-156 [FREE Full text] [doi: 10.1016/j.brat.2014.10.002] [Medline: 25461790]
59. Ciuca AM, Berger T, Crişan LG, Miclea M. Internet-based treatment for panic disorder: a three-arm randomized controlled trial comparing guided (via real-time video sessions) with unguided self-help treatment and a waitlist control. PAXPD study results. J Anxiety Disord 2018 May;56:43-55 [FREE Full text] [doi: 10.1016/j.janxdis.2018.03.009] [Medline: 29625834]
60. Kennedy SH, Lam RW, McIntyre RS, Tourjman SV, Bhat V, Blier P, CANMAT Depression Work Group. Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 clinical guidelines for the management of adults with major depressive disorder: section 3. Pharmacological treatments. Can J Psychiatry 2016 Sep;61(9):540-560 [FREE Full text] [doi: 10.1177/0706743716659417] [Medline: 27486148]
61. Simon GE, Rutter CM, Peterson D, Oliver M, Whiteside U, Operskalski B, et al. Does response on the PHQ-9 depression questionnaire predict subsequent suicide attempt or suicide death? Psychiatr Serv 2013 Dec 01;64(12):1195-1202 [FREE Full text] [doi: 10.1176/appi.ps.201200587] [Medline: 24036589]
62. Gan DZ, McGillivray L, Han J, Christensen H, Torok M. Effect of engagement with digital interventions on mental health outcomes: a systematic review and meta-analysis. Front Digit Health 2021 Nov 4;3:764079 [FREE Full text] [doi: 10.3389/fdgth.2021.764079] [Medline: 34806079]
63. Haller E, Watzke B. The role of homework engagement, homework-related therapist behaviors, and their association with depressive symptoms in telephone-based CBT for depression. Cogn Ther Res 2020 Jul 22;45(2):224-235 [FREE Full text] [doi: 10.1007/s10608-020-10136-x]
64. Knittle K, Gellert P, Moore C, Bourke N, Hull V. Goal achievement and goal-related cognitions in behavioral activation treatment for depression. Behav Ther 2019 Sep;50(5):898-909 [FREE Full text] [doi: 10.1016/j.beth.2019.01.005] [Medline: 31422846]
65. Stewart V, McMillan SS, Hu J, Ng R, El-Den S, O'Reilly C, et al. Goal planning in mental health service delivery: a systematic integrative review. Front Psychiatry 2022 Dec 19;13:1057915 [FREE Full text] [doi: 10.3389/fpsyt.2022.1057915] [Medline: 36601527]
66. Jacob J, Stankovic M, Spuerck I, Shokraneh F. Goal setting with young people for anxiety and depression: what works for whom in therapeutic relationships? A literature review and insight analysis. BMC Psychol 2022 Jul 13;10(1):171 [FREE Full text] [doi: 10.1186/s40359-022-00879-5] [Medline: 35831897]
67. IBM SPSS Statistics for Windows. IBM Corp. 2021. URL: https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-21 [accessed 2024-01-18]
68. Apil SR, Hoencamp E, Judith Haffmans PM, Spinhoven P. A stepped care relapse prevention program for depression in older people: a randomized controlled trial. Int J Geriatr Psychiatry 2012 Jun;27(6):583-591 [FREE Full text] [doi: 10.1002/gps.2756] [Medline: 21766336]
69. DiNapoli EA, Pierpaoli CM, Shah A, Yang X, Scogin F. Effects of home-delivered cognitive behavioral therapy (CBT) for depression on anxiety symptoms among rural, ethnically diverse older adults. Clin Gerontol 2017;40(3):181-190 [FREE Full text] [doi: 10.1080/07317115.2017.1288670] [Medline: 28452665]

70. López-López JA, Davies SR, Caldwell DM, Churchill R, Peters TJ, Tallon D, et al. The process and delivery of CBT for depression in adults: a systematic review and network meta-analysis. Psychol Med 2019 Sep;49(12):1937-1947 [FREE Full text] [doi: 10.1017/S003329171900120X] [Medline: 31179960]
71. Luo C, Sanger N, Singhal N, Pattrick K, Shams I, Shahid H, et al. A comparison of electronically-delivered and face to face cognitive behavioural therapies in depressive disorders: a systematic review and meta-analysis. EClinicalMedicine 2020 Jul;24:100442 [FREE Full text] [doi: 10.1016/j.eclinm.2020.100442] [Medline: 32775969]
72. Mohr DC, Ho J, Duffecy J, Reifler D, Sokol L, Burns MN, et al. Effect of telephone-administered vs face-to-face cognitive behavioral therapy on adherence to therapy and depression outcomes among primary care patients: a randomized trial. JAMA 2012 Jun 06;307(21):2278-2285 [FREE Full text] [doi: 10.1001/jama.2012.5588] [Medline: 22706833]
73. Meyer B, Berger T, Caspar F, Beevers CG, Andersson G, Weiss M. Effectiveness of a novel integrative online treatment for depression (Deprexis): randomized controlled trial. J Med Internet Res 2009 May 11;11(2):e15 [FREE Full text] [doi: 10.2196/jmir.1151] [Medline: 19632969]
74. Hans E, Hiller W. Effectiveness of and dropout from outpatient cognitive behavioral therapy for adult unipolar depression: a meta-analysis of nonrandomized effectiveness studies. J Consult Clin Psychol 2013 Feb;81(1):75-88 [FREE Full text] [doi: 10.1037/a0031080] [Medline: 23379264]
75. Zieve GG, Persons JB, Yu LA. The relationship between dropout and outcome in naturalistic cognitive behavior therapy. Behav Ther 2019 Jan;50(1):189-199 [FREE Full text] [doi: 10.1016/j.beth.2018.05.004] [Medline: 30661559]
76. Davidson KW, Rieckmann N, Clemow L, Schwartz JE, Shimbo D, Medina V, et al. Enhanced depression care for patients with acute coronary syndrome and persistent depressive symptoms: coronary psychosocial evaluation studies randomized controlled trial. Arch Intern Med 2010 Apr 12;170(7):600-608 [FREE Full text] [doi: 10.1001/archinternmed.2010.29] [Medline: 20386003]
77. Schueller SM, Kwasny MJ, Dear BF, Titov N, Mohr DC. Cut points on the Patient Health Questionnaire (PHQ-9) that predict response to cognitive-behavioral treatments for depression. Gen Hosp Psychiatry 2015;37(5):470-475 [FREE Full text] [doi: 10.1016/j.genhosppsych.2015.05.009] [Medline: 26077754]
78. van Straten A, Hill JJ, Richards DA, Cuijpers P. Stepped care treatment delivery for depression: a systematic review and meta-analysis. Psychol Med 2015 Jan;45(2):231-246. [doi: 10.1017/S0033291714000701] [Medline: 25065653]
79. Ardito RB, Rabellino D. Therapeutic alliance and outcome of psychotherapy: historical excursus, measurements, and prospects for research. Front Psychol 2011 Oct;2:270 [FREE Full text] [doi: 10.3389/fpsyg.2011.00270] [Medline: 22028698]
80. Arnow BA, Steidtmann D, Blasey C, Manber R, Constantino MJ, Klein DN, et al. The relationship between the therapeutic alliance and treatment outcome in two distinct psychotherapies for chronic depression. J Consult Clin Psychol 2013 Aug;81(4):627-638 [FREE Full text] [doi: 10.1037/a0031530] [Medline: 23339536]
81. Baier AL, Kline AC, Feeny NC. Therapeutic alliance as a mediator of change: a systematic review and evaluation of research. Clin Psychol Rev 2020 Dec;82:101921 [FREE Full text] [doi: 10.1016/j.cpr.2020.101921] [Medline: 33069096]
82. Karver MS, Handelsman JB, Fields S, Bickman L. Meta-analysis of therapeutic relationship variables in youth and family therapy: the evidence for different relationship variables in the child and adolescent treatment outcome literature. Clin Psychol Rev 2006 Jan;26(1):50-65 [FREE Full text] [doi: 10.1016/j.cpr.2005.09.001] [Medline: 16271815]
83. Martin DJ, Garske JP, Davis MK. Relation of the therapeutic alliance with outcome and other variables: A meta-analytic review. J Consult Clin Psychol 2000;68(3):438-450 [FREE Full text] [doi: 10.1037//0022-006x.68.3.438]
84. Shirk SR, Karver M. Prediction of treatment outcome from relationship variables in child and adolescent therapy: a meta-analytic review. J Consult Clin Psychol 2003 Jun;71(3):452-464 [FREE Full text] [doi: 10.1037/0022-006x.71.3.452] [Medline: 12795570]
85. Andersson E, Ljótsson B, Hedman E, Enander J, Kaldo V, Andersson G, et al. Predictors and moderators of internet-based cognitive behavior therapy for obsessive–compulsive disorder: results from a randomized trial. J Obsessive Compuls Relat Disord 2015 Jan;4:1-7 [FREE Full text] [doi: 10.1016/j.jocrd.2014.10.003]
86. Herbst N, Franzen G, Voderholzer U, Thiel N, Knaevelsrud C, Hertenstein E, et al. Working alliance in internet-based cognitive-behavioral therapy for obsessive-compulsive disorder. Psychother Psychosom 2016;85(2):117-118 [FREE Full text] [doi: 10.1159/000441282] [Medline: 26807976]
87. Mercadal Rotger J, Cabré V. Therapeutic alliance in online and face-to-face psychological treatment: comparative study. JMIR Ment Health 2022 May 02;9(5):e36775 [FREE Full text] [doi: 10.2196/36775] [Medline: 35499910]
88. Nordgren LB, Hedman E, Etienne J, Bodin J, Kadowaki A, Eriksson S, et al. Effectiveness and cost-effectiveness of individually tailored internet-delivered cognitive behavior therapy for anxiety disorders in a primary care population: a randomized controlled trial. Behav Res Ther 2014 Aug;59:1-11 [FREE Full text] [doi: 10.1016/j.brat.2014.05.007] [Medline: 24933451]
89. Wagner B, Brand J, Schulz W, Knaevelsrud C. Online working alliance predicts treatment outcome for posttraumatic stress symptoms in Arab war-traumatized patients. Depress Anxiety 2012 Jul;29(7):646-651 [FREE Full text] [doi: 10.1002/da.21962] [Medline: 22678971]
90. Andersson E, Enander J, Andrén P, Hedman E, Ljótsson B, Hursti T, et al. Internet-based cognitive behaviour therapy for obsessive–compulsive disorder: a randomized controlled trial. Psychol. Med 2012 Feb 21;42(10):2193-2203 [FREE Full text] [doi: 10.1017/s0033291712000244]

91. Don FJ, Driessen E, Peen J, Spijker J, DeRubeis RJ, Blankers M, et al. The temporal associations of therapeutic alliance and manual adherence with depressive symptom change in cognitive behavioral therapy for adult outpatient major depression. Front Psychiatry 2020 Jan 13;11:602294 [FREE Full text] [doi: 10.3389/fpsyt.2020.602294] [Medline: 33519551]
92. Hadjistavropoulos HD, Pugh NE, Hesser H, Andersson G. Therapeutic alliance in internet-delivered cognitive behaviour therapy for depression or generalized anxiety. Clin Psychol Psychother 2017 Mar;24(2):451-461 [FREE Full text] [doi: 10.1002/cpp.2014] [Medline: 27060617]
93. Preschl B, Maercker A, Wagner B. The working alliance in a randomized controlled trial comparing online with face-to-face cognitive-behavioral therapy for depression. BMC Psychiatry 2011 Dec 06;11:189 [FREE Full text] [doi: 10.1186/1471-244X-11-189] [Medline: 22145768]
94. Wehmann E, Köhnen M, Härter M, Liebherz S. Therapeutic alliance in technology-based interventions for the treatment of depression: systematic review. J Med Internet Res 2020 Jun 11;22(6):e17195 [FREE Full text] [doi: 10.2196/17195] [Medline: 32525484]

Abbreviations

CBT: cognitive behavioral therapy
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
i-CBT: internet-based cognitive behavioral therapy
ITT: intention-to-treat
MDD: major depressive disorder
MINI: Mini International Neuropsychiatric Interview
OPTT: Online Psychotherapy Tool
PHQ-9: Patient Health Questionnaire–9
Q-LES-Q: Quality of Life Enjoyment and Satisfaction Questionnaire–Short Form
QIDS: Quick Inventory of Depressive Symptomatology
QUOPL: Queen's Online Psychotherapy Lab

Edited by J Torous; submitted 21.08.23; peer-reviewed by D Adler, A AL-Asadi; comments to author 23.10.23; revised version received 28.11.23; accepted 13.12.23; published 09.02.24.

Please cite as:
Jagayat JK, Kumar A, Shao Y, Pannu A, Patel C, Shirazi A, Omrani M, Alavi N
Incorporating a Stepped Care Approach Into Internet-Based Cognitive Behavioral Therapy for Depression: Randomized Controlled Trial
JMIR Ment Health 2024;11:e51704
URL: https://mental.jmir.org/2024/1/e51704
doi:10.2196/51704
PMID:38173167

Original Paper

HealthySMS Text Messaging System Adjunct to Adolescent Group Cognitive Behavioral Therapy in the Context of COVID-19 (Let's Text!): Pilot Feasibility and Acceptability Study

Lauren M Haack[1], PhD; Courtney C Armstrong[1], BSc; Kate Travis[1], MD; Adrian Aguilera[1,2], PhD; Sabrina M Darrow[1], PhD

[1]Department of Psychiatry and Behavioral Sciences, University of California San Francisco, San Francisco, CA, United States

[2]School of Social Welfare, University of California Berkeley, Berkeley, CA, United States

Corresponding Author:
Lauren M Haack, PhD
Department of Psychiatry and Behavioral Sciences
University of California San Francisco
675 18th Street
San Francisco, CA, 94107
United States
Phone: 1 415 502 8060
Email: lauren.haack@ucsf.edu

Abstract

Background: The widespread occurrence and devastating impact of adolescent depression warrant health service research focused on feasible and acceptable digital health tools to supplement evidence-based intervention (EBI) efforts, particularly in the context of shelter-in-place guidelines disrupting youth socialization and service use in the wake of the COVID-19 pandemic. Given the promise of SMS text message interventions to enhance EBI engagement, our team developed the HealthySMS system as an adjunct to one of the most empirically supported interventions for adolescent depression: cognitive behavioral therapy (CBT) group services. The system sends daily SMS text messages requesting responses assessing mood, thoughts, and activities; weekly attendance reminder messages; daily tips about adherence (eg, a prompt for activity completion); and personalized responses based on participants' texts.

Objective: This study aims to evaluate the feasibility and acceptability of HealthySMS in a real-world setting and explore potential mechanisms of change in EBI engagement, before evaluating the system's impact on adolescents' group CBT engagement and, ultimately, depression outcomes.

Methods: Over the course of 2020, we invited all 20 adolescents receiving CBT group services for depression at an outpatient psychiatry clinic to enroll in our HealthySMS study; ultimately, 17 (85%) adolescents agreed to participate. We tracked participant initiation and engagement with the HealthySMS system as well as the content of SMS text message responses to HealthySMS. We also invited each participant to engage in a semistructured interview to gather additional qualitative inputs on the system.

Results: All (n=17, 100%) research participants invited agreed to receive HealthySMS messages, and 94% (16/17) of the participants maintained use during the first month without opting out. We uncovered meaningful qualitative themes regarding the feasibility and acceptability of HealthySMS, as well as its potential impact on EBI engagement.

Conclusions: Taken together, the results of this pilot study suggest that HealthySMS adjunct to adolescent CBT group depression services is feasible and acceptable, as evidenced by high rates of HealthySMS initiation and low rates of dropout, as well as meaningful themes uncovered from participants' qualitative feedback. In addition, the findings provide evidence regarding iterative improvements to the HealthySMS system and research protocol, as well as potential mechanisms of change for enhanced EBI engagement and, ultimately, adolescent depression outcomes, which can be used in future effectiveness research.

(JMIR Ment Health 2024;11:e49317) doi:10.2196/49317

KEYWORDS

depression; adolescents; evidence-based intervention; texting; SMS text message; cognitive behavioral therapy; CBT; group CBT; shelter-in-place; COVID-19; mobile health; mHealth; therapy; cognitive; behavior; web-based therapy; e-therapy; youth; young adults; mobile phone

Introduction

Background

Depression is a major public health concern for adolescents [1-3]; it is a risk factor for mental health problems that occur in adults, medical illnesses, disability, substance abuse, and suicide [2-6]. Evidence-based efforts to improve adolescent depression and related risks became more imperative during the COVID-19 pandemic owing to concerns about adolescents' mental health deterioration in the context of shelter-in-place (SIP) guidelines causing substantial educational and social disruptions [7-10]. Further, most in-person services were halted during the implementation of SIP guidelines, rapidly necessitating a transition to telehealth and digital health (dHealth) supports [7,11-14]. In response, the need for pilot feasibility and acceptability testing of dHealth tools targeting adolescent depression in real-world settings became glaringly apparent. Such efforts would provide an essential first step toward testing the effectiveness of dHealth tools for adolescent depression. The impact of this work would extend beyond the initial COVID-19 SIP context, given that research suggests that many systems will and should continue using telehealth and dHealth supports in the foreseeable future [13,15,16], which may be particularly beneficial in reducing disparities for historically marginalized populations [15].

Evidence-Based Intervention for Adolescents With Depression

There are several effective evidence-based interventions (EBIs) for adolescent depression [17-19]. The most well-supported psychosocial approach is cognitive behavioral therapy (CBT [18,19]), which targets unhelpful thinking patterns and the avoidance of goal-directed and social activities, which are characteristic of depression [20-22]. CBT can be effectively delivered via a group format [23], demonstrating efficacy similar to that of individual CBT in decreasing depression in a more cost-effective modality [24]. CBT can also be delivered electronically (electronic CBT [eCBT]), enhancing accessibility by eliminating the requirement for patients to travel for in-person services [25-29]. The potential benefits of electronically delivered EBIs became increasingly evident in the context of the implementation SIP guidelines during the COVID-19 pandemic, which prohibited many individuals from receiving in-person mental health treatment. Importantly, our society may experience future SIP guidelines during COVID-19 surges, safety lockdowns, or natural disasters in response to climate change. Fortunately, recent research supports the effectiveness and efficacy of eCBT in decreasing depression [25-29].

Despite the promising efficacy research supporting CBT as an EBI for adolescent depression, the effect sizes are heterogeneous and low, demonstrating room for improvement [23,30]. Optimal treatment of adolescent depression requires patient engagement, including the initiation of treatment after referral, the attendance of sessions, the completion of homework between sessions, and the continued engagement in treatment (ie, getting the whole "dose" of treatment [31]). Thus, poor initiation, poor adherence, and treatment dropout are barriers to effective services in real-world settings and are known to mediate treatment outcome [31]. A meta-analysis of EBIs concluded that the average treatment dropout rate is approximately 29% and that patients with depression are at a higher risk for dropout [32]. Treatment dropout appears particularly pronounced among youths and in web-based mental health interventions; therefore, efforts to enhance engagement with eCBT among youth populations are especially called for [33].

SMS Text Message Interventions as Automated dHealth Supports Aiming to Enhance EBI Engagement

There is a strong scientific premise and public health call to develop and evaluate dHealth supports adjunct to EBIs that are easy to implement, efficient, and sustainable. Using SMS text messages is a promising approach aiming to enhance adolescent EBI engagement by making use of a tool that is readily available and widely used by adolescents. The use of mobile phones is ubiquitous among adolescents, and SMS text messages are used at high rates. As of 2015, more than 85% of adolescents across races and ethnicities had access to a cell phone, and 90% of them used cell phones to send SMS text messages [34]. In fact, two-thirds of adolescents reported that they are more likely to use their cell phones to text rather than talk to friends, and a typical adolescent in the United States sends ≥30 SMS text messages each day [34]. Importantly, in mobile health (mHealth) research, youths report high satisfaction and readily engage with technology [35-37]. In addition, digital technology allows for the automation of SMS text messages; thus, an SMS text message intervention leveraging automation should not add to providers' burden, making it a more sustainable and efficient services intervention.

SMS text messaging interventions have several advantages over other forms of dHealth interventions, such as app-based and website-based interventions, and are particularly well suited to improve engagement. SMS text messages allow for a more equitable delivery of care, given that they can reach anyone with a phone and do not rely on smartphone ownership or internet access, which are affected by socioeconomic, racial, and ethnic disparities [38]. In addition, although dHealth interventions have demonstrated promise, web-based and mobile app–based interventions are subject to more difficulties with engagement, such as problems with adherence and dropout [26,39].

Some dHealth SMS text message interventions were designed as stand-alone supports (eg, those in the works of Aguilera et al [40] and Bendsten et al [41]; MacDougall et al [42] conducted a scoping review of SMS text message–delivered adolescent mental health interventions); however, there may be unique benefits to SMS text message interventions designed to supplement EBIs and enhance treatment engagement. There are many effective examples of adjunctive SMS text message interventions for disease management and health behavior change (eg, smoking cessation and diabetes management [43]), including SMS text message interventions for adolescents [44]. In the emerging literature on SMS text message interventions, adolescents generally react favorably and show good compliance [35,42,45,46]. Pilot research on integrating SMS text messages into individual CBT has also demonstrated encouraging results [47].

Given the promise of SMS text message interventions to enhance EBI engagement, our team developed the HealthySMS system as an adjunct to group CBT depression services. HealthySMS sends customized SMS text messages to participants, inquiring about mood as well as reminding them to attend CBT and practice strategies learned during group treatment. For safety, adolescent SMS text message responses are monitored for keywords or phrases that could indicate whether someone is expressing suicide risk; providers are sent immediate alerts if the system is triggered by an adolescent SMS text message response with these keywords or phrases. Safety keyword triggers are important features of any automated dHealth system implemented in real-world settings to keep participants safe and providers informed; in addition, this feature may be particularly helpful during the implementation of SIP guidelines, given the probable decrease in adolescent safety monitoring from other settings, such as schools. The addition of HealthySMS to CBT for adults with depression in a public sector treatment setting was associated with an increased number of sessions attended and a longer duration of treatment [48]. The mood ratings sent by adults through HealthySMS also predicted attendance [49]. Thus, adding HealthySMS to EBIs for adolescent depression may be a promising change to existing services, targeting increased engagement and, ultimately, improving outcomes.

This Study

As a first step in the process of implementing and evaluating the HealthySMS system adjunct to the most evidence-based treatment for adolescent depression (ie, CBT), we conducted a feasibility pilot study in an outpatient clinic embedded within an academic medical center's department of psychiatry and behavioral sciences. Our primary objectives were to investigate the feasibility and acceptability of HealthySMS in a real-world setting and explore potential mechanisms of change in EBI engagement, before evaluating the system's impact on adolescents' group CBT engagement and, ultimately, depression outcomes. We also aimed to monitor the HealthySMS safety keyword alert triggers and provider responses to inform system and research protocol adjustments before future HealthySMS research. This multiphase design featuring a preliminary feasibility pilot to inform decisions about future effectiveness testing is aligned with the Medical Research Council framework [50,51], which has been prolifically used in mental health and dHealth intervention research. We predicted the following hypotheses:

1. Hypothesis 1: most adolescents invited to use HealthySMS (ie, ≥75%) would initiate and maintain use without opting out during their group CBT experience.
2. Hypothesis 2: adolescents enrolled in HealthySMS would display high rates of engagement with the SMS text message system (ie, ≥50% response rate).
3. Hypothesis 3: our team would uncover meaningful qualitative themes from participants' SMS text message responses and semistructured interviews about the feasibility and acceptability of HealthySMS to inform iterative system and research protocol improvements supporting future HealthySMS effectiveness research.
4. Hypothesis 4: our team would uncover meaningful qualitative themes from participants' SMS text message responses and semistructured interviews about the potential impact of HealthySMS on EBI engagement to inform decisions on which mechanisms of change to evaluate in future HealthySMS effectiveness research.

Methods

Participants

Participants were adolescents aged 13 to 18 years. They were recruited from the University of California San Francisco (UCSF) Department of Psychiatry and Behavioral Sciences Child and Adolescent Services clinic between December 2019 and September 2020. Adolescents were eligible to participate if they were enrolled in the clinic's group CBT for depression running from January 2020 to September 2020.

Of the 20 eligible adolescents who started the group CBT for depression, 17 (85%) agreed to participate in the study. Participant characteristics are listed in Table 1. Briefly, the average age of the participants was 15.4 (SD 1.5) years, and most participants (15/17, 88%) were diagnosed with major depressive disorder. Most participants (12/17, 71%) had engaged in prior mental health care. Moreover, 18% (3/17) of the participants had a history of at least 1 suicide attempt, with most participants (11/17, 65%) reporting suicidal ideation in the past year. A little more than half (9/17, 53%) of the participants engaged in concurrent individual therapy, family therapy, or a combination.

Table 1. Participants' demographic and clinical information (N=17).

	Values
Demographic information	
Age (years), mean (SD)	15.4 (1.5)
Gender, n (%)	
Woman	8 (47)
Man	8 (47)
Transwoman	1 (6)
Ethnicity, n (%)	
Latinx	7 (41)
Non-Latinx	10 (59)
Race, n (%)	
Asian	1 (6)
Black	1 (6)
White	10 (59)
Biracial or multiracial	4 (24)
Something else (Hispanic or Mexican)	1 (6)
Sexual orientation, n (%)	
Heterosexual	8 (47)
Lesbian	1 (6)
Bisexual	3 (18)
Gay	1 (6)
Pansexual	1 (6)
Not sure or do not care	3 (18)
Clinical history, n (%)	
Prior therapy experience[a]	
Participated in prior outpatient therapy	9 (53)
Participated in IOP[b] or PHP[c]	2 (12)
Psychiatric hospitalization	1 (6)
Residential treatment facility	2 (12)
None	5 (29)
Prior suicide attempt	3 (18)
SI[d] at intake	
Denied	6 (35)
SI in the past week	3 (18)
SI in the past month	2 (12)
SI in the past year	6 (35)
NSSI[e] at intake	
Denied	10 (59)
NSSI in the past week	1 (6)
NSSI in the past month	2 (12)
NSSI in the past year	4 (24)
School problems[a]	

	Values
None reported	9 (53)
IEP[f] or accommodations	6 (35)
School refusal	3 (18)
Major depressive disorder	
Met criteria in the past	1 (6)
Reports some symptoms	1 (6)
Currently meets criteria	15 (88)
Symptoms from comorbid disorders	
Generalized anxiety disorder	5 (29)
Panic disorder	2 (12)
Social anxiety disorder	1 (6)
Eating disorder	1 (6)
ADHD[g]	1 (6)
None	7 (41)
Concurrent therapy[a]	
Individual CBT[h]	6 (35)
Family therapy	4 (24)
None	8 (47)
Concurrent medication management	15 (88)

[a]Response options are not mutually exclusive.
[b]IOP: intensive outpatient treatment.
[c]PHP: partial hospitalization program.
[d]SI: suicidal ideation.
[e]NSSI: nonsuicidal self-injury.
[f]IEP: individualized education plan.
[g]ADHD: attention-deficit/hyperactivity disorder.
[h]CBT: cognitive behavioral therapy.

Of note, our research protocol did not make any changes to how the real-world clinic provided services. Adolescents were not incentivized to start treatment, attend sessions, or complete their homework; they only were incentivized for completing additional research tasks (see below). Adolescents in the clinic could be engaged in individual therapy, family therapy, or both while attending group treatment.

Ethical Considerations

All study procedures were approved by the UCSF Institutional Review Board (reference number 255820). When participants were minors (ie, age <18 years), their parent or legal guardian completed informed consent and they completed assent procedures; 18-year-old participants completed informed consent procedures. As part of these procedures, participants (and their parents or legal guardians, if applicable) were informed that their information would be kept private and housed on a secure UCSF server only accessible to the study team; they were informed that participants would be compensated with a US $30 gift card for attending the interview).

CBT Group Intervention

The Cognitive Behavioral Therapy for Depression Group for Adolescents (CBT-D) consists of three 4-week modules on thoughts, activities, and people (ie, 12 group sessions in total). It is based on the Building Recovery by Improving Goals, Habits, and Thoughts (BRIGHT) group CBT manual for depression for adults developed by Miranda et al [52], which was subsequently adapted for adolescents with a diagnosis of major depression or persistent depressive disorder. The thought module involves cognitive interventions with a focus on awareness of helpful and harmful thoughts, the activity module focuses on pleasant activities and behavioral activation, and the people module encourages group members to improve relationships and evaluate the impact of positive and negative social relationships on mood. The sessions are structured with an initial homework check-in followed by a didactic discussion on the covered topic, an interactive discussion led by 2 group providers, and the setting of homework goals related to skill use. In this study, the first 7 group sessions were held in person; the remainder were held via the Health Insurance Portability and Accountability Act–compliant Zoom (Zoom Video

Communications, Inc) platform after SIP orders were implemented in March 2020.

The HealthySMS System

We used the Health Insurance Portability and Accountability Act–compliant, web-based texting platform called HealthySMS, developed by our team member (AA), to send and receive SMS text messages, administer weekly surveys, and track attendance. HealthySMS sent four types of automated SMS text messages to participants: (1) daily mood prompts asking participants to rate their mood and reflect on their mood, thoughts, and behavior ("[First name], what is your mood right now on a scale of 1 to 9 (9 being best)? Please respond with a number and a message about what you are doing or thinking"); (2) personalized and reinforcing responses to 20% of participants' mood ratings (ie, encouraging the participants to engage in behavioral activation in response to a low mood rating and reinforcing the participants in response to a high mood rating); (3) daily reminders about the concepts and skills learned during the corresponding CBT module for that month; and (4) weekly reminders to attend and come prepared to CBT sent the day before the group session. The participants could opt out of receiving SMS text messages at any point by texting "stop."

Our clinical research team oriented adolescent participants to the HealthySMS system. We also oriented providers to the SMS text messages and HealthySMS web-based provider dashboard, which visualizes client responses to the SMS text messages, including graphs of mood ratings. HealthySMS monitors participants' SMS text message responses and alerts providers to words and phrases that may correspond to suicidal behaviors (eg, "die," "kill," and "cut"), and providers were trained on how to respond to these alerts in alignment with existing clinic policies and procedures.

Measures

Participants' Demographics and Clinical History

The participants were asked to complete a demographic survey before beginning group CBT. CBT-D providers recorded psychiatric diagnoses and treatment history information for each participant, including other therapy or medication management services, based on their evaluation and review of the medical record.

HealthySMS Engagement

We tracked the number of participants who were invited to receive HealthySMS messages, as well as those who agreed to initiate receipt of HealthySMS messages. We also tracked the number of participants who texted "stop" to opt out of SMS text messages before ending participation in the group, as well as how long each participant received the SMS text messages before opting out. We measured engagement with the HealthySMS messages by tracking the number of mood ratings that participants texted in response to mood rating request messages.

HealthySMS Safety Keyword Triggers

We tracked the number and content of participant SMS text message responses that were flagged for the risk of suicide (Multimedia Appendix 1 provides the trigger words). We also kept observation notes about the provider responses.

Qualitative Feedback

We collected qualitative feedback in several ways. We tracked responses to our monthly SMS text message prompts to participants asking, "What is the most positive part of receiving these text messages?" and "What do you not like about receiving the text messages?" We also invited all adolescent participants to share feedback in a semistructured interview after their CBT-D completion with an incentive of US $30 Amazon gift card. The interviews were moderated by a member of our clinical research team who followed a semistructured guide containing the study objectives to explain, questions to pose, and prompts to use when needed. Specifically, we explained that researchers hoped to obtain information about mHealth interventions such as HealthySMS and group CBT services for depression. Next, we asked for general feedback about the HealthySMS system and then specifically asked about different aspects of HealthySMS, such as the mood prompts, responses to mood ratings, and group reminders. We also showed participants a list of the HealthySMS texts and asked for feedback on their impact, content, and phrasing. Textbox 1 lists the interview questions and prompts. Each interview lasted between 19 and 45 (mean 32, SD 8.8) minutes, with the length depending on the amount of details provided by the respondents.

Textbox 1. Qualitative semistructured interview questions and prompts.

How did you like the LET’S TEXT! message program overall?

- Was there anything that made it difficult for you to receive the messages? For example, how did you like: the timing of the messages, the amount of messages, the phrasing of messages?
- Was there anything that made it difficult for you to respond to the messages? For example, how did you like: the timing of the messages, the amount of messages, the phrasing of messages?
- Is there anything about it you would suggest we change?

How did you like the LET’S TEXT! Mood Prompts and Responses? (after general feedback was given, participants were shown the list of printed texts for specific feedback on this question)

- What did you think worked well?
- What was difficult or did not go well?
- Is there anything you would change?
- Did these messages change your mood, thoughts, or behavior?

How did you like the LET’S TEXT! Skill Practice Reminders? (after general feedback was given, participants were shown the list of printed texts for specific feedback on this question)

- What did you think worked well?
- What was difficult or did not go well?
- Is there anything you would change?
- Did these messages change your mood, thoughts, or behavior?
- For example, did you need more/less help with any of the skills; was the purpose of the reminders clear; -were the messages too few/many; were the reminders relevant to your goals; were they phrased appropriately?

How did you like the LET’S TEXT! Group Reminders?

- Did these messages change your mood, thoughts, or behavior?

Some teens find...

- ...it easier to participate if they feel: comfortable, respected, and understood by the group members and group leader attached to the messages. How was your relationship with the group and group leader and did that affect your experience with the text messages?
- ...that the messages help them become more active. Do you think your activity completion was impacted by the text messages?
- ...the messages help them go to group and participate more often and/or effectively. How was your group engagement and did that change with the text messages?
- ...the messages are reinforcing and help them practice the group skills more often and/or effectively. How was your practice of the group skills and did that change with the text messages?

Is there ANYTHING ELSE you would like to share that we haven’t asked you?

Data Analytic Plan

We analyzed quantitative data on participant characteristics, HealthySMS engagement, and safety keyword triggers using descriptive statistics in SPSS (IBM Corp). We calculated the SMS text message response rates by dividing the number of responses by the number of SMS text messages received for mood ratings.

We analyzed the qualitative data in a multistep process using thematic analysis principles [53]. First, we developed a hierarchical coding system based on recurrent concepts that we uncovered when conducting the qualitative interviews and in consideration of the related theoretical literature. Next, members of our research team reviewed the qualitative message content and interview transcriptions to collaboratively refine recurrent themes while iteratively updating the coding system. Our team selected exemplary quotes for each theme.

Results

Hypothesis 1

Our first hypothesis was that most of the adolescents (ie, >75%) invited to HealthySMS would initiate and maintain use. All participants who agreed to the research study (n=17, 100%) opted to enroll in the HealthySMS system and initiate the receipt of messages upon starting CBT-D; 94% (16/17) of the participants maintained use during the CBT-D group experience. Only 1 (6%) participant opted out of the SMS text messages by texting “stop” 30 days after initiation.

Hypothesis 2

Our second hypothesis was that adolescents enrolled in HealthySMS would display high rates of engagement with the SMS text message system (ie, ≥75% response rate). As shown in Table 2, the HealthySMS response rate varied among participants. The average response rate to daily mood ratings was 61%, with a range of 0.00 to 1.77 responses per message. Of the 17 participants, only 1 (6%) participant did not respond to any of the SMS text messages prompting mood ratings, and 2 (12%) participants responded multiple times to several prompts (indicated by a response proportion >1). Most participants (10/17, 59%) responded to >50% of the daily mood rating prompts.

Table 2. HealthySMS engagement.

Participant	Proportion of responses to mood ratings	Weeks until opting out
1	0.04	N/A[a]
2	0.16	N/A
3	0.30	N/A
4	0.64	N/A
5	0.01	N/A
6	1.54	N/A
7	0.00	N/A
8	0.87	N/A
9	0.05	N/A
10	0.82	N/A
11	1.77	N/A
12	0.12	N/A
13	0.87	N/A
14	0.77	N/A
15	0.79	N/A
16	0.75	N/A
17	0.84	4

[a]N/A: not applicable.

Hypothesis 3

Our third hypothesis was that we would uncover meaningful qualitative themes from participants' SMS text message responses and semistructured interviews about the feasibility and acceptability of HealthySMS. We posited that meaningful themes would be beneficial in informing iterative system and research protocol improvements supporting future HealthySMS effectiveness research. When examining the context of participants' SMS text message responses to our monthly message prompts and semistructured interviews asking for feedback, we identified themes and exemplary quotes regarding the feasibility and acceptability of HealthySMS (Tables 3-5).

Table 3. Qualitative themes uncovered and example quotes regarding HealthySMS's feasibility

Theme	Example quote supporting feasibility	Example quote about limited feasibility
Text modality	"I am not really great with the emails. I think just going through texts is almost always better."	"For a while, I just didn't even open the texts."
Consistency of delivery	"I always got the text and sometimes if I didn't reply, like with my mood, I would get like a reminder, maybe 10 minutes later, respond your mood. So I would say technically or technologically, it was all good. I didn't run into any issues."	"Sometimes my mood ratings just wouldn't come in. Like some days, like towards the ending of it, they just came in periodically like not every day."
Amount of effort to use	"[It was fine to] like, give a number, how was your day? But if it was like kind of describe your day, I don't think anyone would want to do that because it would take too long."	"If you didn't like type in your mood right away, it would like send you a reminder a lot. So I got those a lot if I wasn't doing it."
Timing of texts	"I definitely didn't get anything like super early or late. I would say they did a pretty good job of, like, changing up time, so it wasn't like the same time every like 8AM and 10PM, like it was pretty good at switching up times so you could get like different times of the day."	"I guess sometimes they came at weird times, like really early in the morning or really late at late night...It kind of felt less helpful if they came later in the day because either I had already figured it out or like got past it or it just didn't help anymore."
Amount of texts	"I don't remember getting, like, bombarded with them. So I would say the amount is probably pretty good and reasonable."	"It felt sometimes like it was getting too many. But that was mainly because that was like at a time when I didn't need them and so it just felt like a waste of message if that makes sense."
Text length	"They're all quick...It's short and simple and sweet..."	"I liked it, but it was also like kind of a lot sometimes."

Table 4. Qualitative themes uncovered and example quotes regarding HealthySMS's acceptability

Theme	Example quote supporting acceptability	Example quote about limited acceptability
Group reminders	"I liked it, because sometimes when things got a little hectic, I would forget about group so it's a useful reminder."	"They weren't the most helpful, especially because I'm pretty sure I only ever got one text and it was just 'there is group tomorrow. Don't forget your binder,' which I realize is probably works a lot better when you actually have to go somewhere to do it. But like I already said, I have nothing else going on [in quarantine], so I remember yep, that's tomorrow. I felt like there definitely could be some variation in things like maybe it could bring up, like, remember the skills you learned or don't be scared to share or something like that, that could be a little more emotional based."
Personalized response texts	"I did like how if your mood was like a bad mood, it would give you like, it wouldn't it just be like that sucks feel better, but it would give you like advice and strategies of how you can get better. I think that was nice. And I liked when if you were in a good mood, it would kind of still give you like a different kind of advice to keep you in that space and like be like, like you could use that experience to feel better later when you remember."	"...having a little bit of background about like who it is I'm responding to, like, is it a computer or is it a person like this data is being used for what, kind of thing might be helpful."
Mood ratings	"The rating of the mood...it kind of helped me figure out how I was feeling, like in a number form."	"I would say it was kind of easy when I was like doing super well or super bad..., but little bit harder, I guess, like to be: 'OK, well, I don't really know. I'm average.'"
Skill reminder texts overall	"I thought they really were helpful...a really nice boost, and it was and it would remind me of the other things I had learned that I could also use to feel better."	"Um if I was able to do something like kind of just the reminder was helpful. But with COVID a lot of the times, it wasn't applicable."
Activity skill reminders	"I always think it's good to like set goals for yourself so you have something to work towards. So it's almost like a little bit of motivation."	"...like: 'do a new activity'...a lot of activities have like kind of gotten harder to do with everything [in the context of the COVID-19 pandemic]."
Social skill reminders	"I liked the people [texts], because I struggle with my relationships with people."	"Sometimes I couldn't hang out with friends. So then sometimes that even made me, like, a little frustrated."
Cognition skill reminders	"I think sometimes like I get so stuck in like the past or just like in the moment that it's good to just like think about your future and like the good things that are to come."	"...it was just odd getting text telling me to, like, change the way I'm thinking because I know like the way my brain works, it's not going to just happen...So it was just annoying because I would like to make [the thoughts] go away, but they're not going to."
Statements vs questions	"I felt that they were good, I like them more when they were more statement based...I felt that just a clear like statement or like advice boost helped more for me personally."	"I got confused a bit with the questions they'd ask because I wasn't sure if I was supposed to, like, respond to them and it would respond back or it was just kind of a moment to reflect."

Table 5. Qualitative themes uncovered and example quotes regarding HealthySMS's potential impact on evidence-based intervention (EBI) engagement.

Theme	Example quote supporting impact on EBI engagement	Example quote about limited impact on EBI engagement
Participation in group sessions	"It was just like a reminder to like start thinking about group. So like before group, I'd just like start thinking about it so I'd like have more to say."	"I wouldn't really say that the text messages changed my experience in the actual group. It felt more like kind of a recap throughout the week, and less than, sort of a second part of group, if that makes sense. Yeah like we talked about the text messages a bit in group, but the text messages were always about the thing that happened last week, and so we'd want to move on. It could be helpful if the text messages like cover, like some newer things to introduce you, but um yeah."
Group homework completion	"That was kind of like a good check in where I was like, OK, did I have homework? What was it? Could that help me? So I would keep that in there."	"I would definitely do techniques more if I got a text message reminding me to do it [the specific homework rather than the general skill reminder] in the week."
Connection with others in the group	"It felt like we were closer because we were all getting the same messages and we were all in the same boat together."	"My experience with my group and group leader was pretty good, but I wouldn't really say my experience in group and my experience with the text messages were linked."
Validation and support	"And it felt like someone was like caring about me."	"But it also could backfire on if you don't do it [the suggestion in the text], feeling guilty."
Impact on mood	"It kind of gave me some positivity boost, and it was nice to have someone or something to talk to me while I was feeling that way."	"I wasn't doing anything because I couldn't go out and it just made me think about it more...So it just made me a little bit more sad."
Behavior change and skill use	"I got something like the, call or spend time with people who make you feel happy...And I ended up calling one of my closest friends and she did bring up my mood a little bit..."	"I didn't usually act on it, probably just because it was like, OK, well, right now I'm doing something. So then I wouldn't really remember to do it later, but I can see how, what the idea was."

Hypothesis 4

Our final hypothesis was that we would uncover meaningful qualitative themes from participants' SMS text message responses and semistructured interviews about the potential impact of HealthySMS on EBI engagement. We posited that meaningful themes would be beneficial in informing decisions on which mechanisms of change to evaluate in future HealthySMS effectiveness research. When examining the context of participants' SMS text message responses to our monthly message prompts and semistructured interviews asking for feedback, we identified themes and exemplary quotes regarding the potential impact of HealthySMS on EBI engagement (Table 5).

Safety During the HealthySMS Intervention

A secondary aim of our pilot study was to monitor the HealthySMS safety keyword alert triggers and provider responses to inform system and research protocol adjustments before future HealthySMS research. During the study period, 76 (7.58%) of the 1002 total SMS text messages sent by participants alerted providers to potential suicide risk throughout our flagged keyword system. When examining the content of participants' SMS text message responses that were flagged for risk of suicide, only 2 (3%) of these 76 messages were determined to contain true risk-related content (eg, texts about wanting to hurt oneself or die by suicide). In one case, the notified provider determined that the text may be an indication that the participant was about to self-harm, and they followed the clinic's safety protocol; no indication that harm occurred was received, and the participant continued attending the group sessions and engaging in the study. In the second case, the notified provider determined that the text did not represent an increase in risk and addressed the client's worry about the future in their following session.

Discussion

Principal Findings

The results of this feasibility pilot study demonstrate that the use of HealthySMS adjunct to adolescent group CBT depression services (CBT-D) appears feasible and acceptable, as evidenced by high rates of HealthySMS initiation and low rates of dropout, as well as meaningful themes uncovered from participants' qualitative feedback. Importantly, the findings also provide evidence regarding iterative improvements to the HealthySMS system and research protocol, as well as potential mechanisms of change for enhanced EBI engagement and, ultimately, adolescent depression outcomes, which can be used in future effectiveness research. It is compelling that the results of this study were obtained in the context of ongoing clinical services at a real-world outpatient clinic experiencing a transition to telehealth services amidst the onset of the COVID-19 pandemic and that the research protocol did not alter the clinical service procedures in any way, such as by incentivizing adolescents to attend the group sessions. Furthermore, it should be noted that we were able to implement this intervention in a safe manner during at a time when an increasing number of youths were at a risk for suicide; the 2 instances of HealthySMS alerts indicating risk were managed via clinical procedures, and no adverse outcomes occurred to any participant during the study.

HealthySMS Feasibility and Acceptability

As predicted, adolescents enrolled in and maintained the use of HealthySMS at high rates; in fact, no adolescent who agreed to the research study declined initiation of HealthySMS, and only 1 (6%) of the 17 enrolled adolescents opted out after a month of use. HealthySMS response rate was slightly lower than predicted but, of note, varied among participants. Some adolescents had very low response rates (ie, n=1, 6% never responded to any messages), and some adolescents had high response rates (ie, n=1, 6% responded multiple times to most prompts). Overall, most adolescents responded more than half of the time to daily mood rating prompts (ie, response rates averaging >60%). Qualitative feedback suggested that quick and short messages, as well as midday versus early or late message timing, may be the most feasible to respond to. In addition, some adolescents shared that they were motivated to respond by the "interactive" nature of the mood rating texts, which triggered responses 20% of the time. We received feedback that the number of HealthySMS messages and reminders felt appropriate for some participants; however, we received other feedback indicating a preference for less frequent messages and reminders. A future direction to explore is whether personalizing the message and reminder timing and frequency (eg, by requesting participants to share their preferences in the initial survey and monthly feedback prompts to adapt the system to allow for differences by preference) may increase response rates. Interestingly, some adolescents also sent SMS text messages back to the skill reminder messages (ie, response rates averaging 12%), although there were no explicit requests for responses. Adding explicit response requests to skill reminder messages may be another way to explore the potential benefits of HealthySMS.

Regarding the HealthySMS safety triggers, the system appeared to appropriately monitor risk. Throughout the 10 months of the study, 76 (7.58%) of the 1002 total messages sent by participants during the study period triggered provider alerts via flagged keywords (eg, cut, kill, and die; listed in Multimedia Appendix 1). We added words to the system for the current trial after consulting with data from the adolescent crisis text line. Importantly, only 2 of the alerts in this study indicated a potential suicide risk and required clinical follow-up. Providers complied with the established clinical policies and procedures, and no known adverse outcomes occurred among participants throughout the study. In fact, the vast majority of alerts were false positives (eg, "I just got my haircut," "I went for a bikeride over the bridge," and "my throat hurts"). One of the most sensitive keywords appeared to be "end," given that it is a frequently used word and is included in many words (eg, "almost the end of the day" and "FaceTime with friend"). This information can be used to iteratively improve the HealthySMS system to balance the need to detect safety concerns with high sensitivity while trying to avoid false positives and, therefore, reducing providers' burden in reviewing the alerted SMS text messages, such as by programming keyword alerts to trigger only when the word "end" by itself is sent and not when it is included in other words, such as "friend." It will be important for future HealthySMS efforts to consider how future HealthySMS research designed to evaluate the appropriateness of the safety keyword triggers and subsequent provider responses to prevent self-harming behaviors would be beneficial.

Potential Impact on EBI Engagement to Be Evaluated in Future Effectiveness Research

When asked about the utility of HealthySMS, we received input providing initial evidence that the system may indeed enhance EBI engagement. Our identified themes may be used to inform decisions on which mechanisms of change to evaluate in future HealthySMS effectiveness research designed to measure the impact of the system on EBI engagement and, ultimately, adolescent depression outcomes. To begin, most adolescents gave feedback that group reminder texts may have increased their likelihood to attend and meaningfully participate in the group sessions, although one of the participants felt that the group reminder texts were intrusive and did not impact engagement in the group sessions. Variation in responses again highlighted that a personalized SMS text messaging system would be ideal for accommodating individual needs and desires.

Qualitative responses also suggest that the HealthySMS skill messages provided beneficial reminders and motivation for some adolescents to engage in CBT strategies, including behavioral activation and helpful thinking. Some participants did suggest that we sync skill messages with the content being delivered each week (rather than with each module, as is the current setup) to optimize relevance. Although weekly agendas are generally known from the beginning, given the CBT-D format in this study, group providers may adjust agendas to meet the needs of the current group members. Thus, personalizing the timing and content of messages and syncing messages more precisely to weekly content would likely be perceived as helpful by participants but would need to be weighed against the feasibility of such measures, which could increase providers' burden.

We were also interested in the impact of HealthySMS on the interpersonal connection of the adolescents in the current trial, given the reports of adults in our team's prior trial of HealthySMS that participants felt cared for and supported by the messages, as well as closer to the CBT group (sources removed for masked review). We received feedback suggesting that HealthySMS may help some adolescents feel more connected to the group leader and other members, as well as validated and supported even with the knowledge that it is a "bot" responding. One of the providers shared that a participant felt that HealthySMS was like a "friend." Feedback from participant interviews indicated that this sense of connection was especially true when they received the responses to their mood rating texts, which were aligned with how they rated their mood (ie, different responses were sent for high, medium, or low mood ratings).

Limitations and Future Directions

Our study has several limitations that should be acknowledged and addressed in future work. First, our pilot study investigating the feasibility and acceptability of HealthySMS in a real-world treatment setting was not designed to evaluate the system's role in improving the ultimate target of adolescent depression. To accomplish this, future effectiveness research building on the lessons learned and iteratively updated deliverables (ie, the HealthySMS system and research protocol) from this study is warranted. Specifically, although we were able to explore the potential impact of HealthySMS on EBI engagement in participants' qualitative feedback, we did not have the data or sample size required to quantitatively evaluate this construct. In addition, our decision to conduct this pilot study in a real-world setting without making changes to existing clinical procedures created several potentially confounding variables among participants, such as the number of other EBI services engaged in adjunct to CBT-D. Subsequent studies fully powered to detect mechanisms of change as well as control for potential covariates to EBI engagement and outcomes using a control group are called for. Finally, this pilot study took place in the initial months of SIP enforcement owing to the COVID-19 pandemic; thus, it is not known how well our findings will generalize beyond this context.

Conclusions

Our pilot study suggests that HealthySMS adjunct to the most evidence-based treatment for adolescent depression (ie, CBT) is feasible and acceptable, warranting future effectiveness research evaluating the system's impact on adolescent EBI engagement and subsequent depression outcomes. As SMS text messaging is cheap and uses technology already in the hands of most adolescents, it is well suited to be added to existing clinical services. Feedback from the participants in our studies suggested that mHealth may be particularly helpful during times of SIP enforcement, given the limited ability for adolescents to engage in activities, social interactions, in-person mental health treatment, and safety monitoring by adults in their lives (eg, teachers and providers). However, it is possible that SMS text message systems are beneficial adjuncts to EBIs in all contexts, given their potential to increase the likelihood and effectiveness of service participation, enhance feelings of connectedness and validation, and encourage skill use in-between sessions. Continued service research on the implementation and effectiveness of mHealth tools has the potential to improve mental health services for a population experiencing drastic increases in depression and suicide risk: adolescents.

Conflicts of Interest

AA is the owner of the HealthySMS program license and has licensed it to other researchers for use in their studies. He was not paid for the license for this study.

Multimedia Appendix 1
Risk alert words.
[DOCX File , 15 KB - mental_v11i1e49317_app1.docx]

References

1. Costello EJ, Copeland W, Angold A. Trends in psychopathology across the adolescent years: what changes when children become adolescents, and when adolescents become adults? J Child Psychol Psychiatry 2011 Oct;52(10):1015-1025. [doi: 10.1111/j.1469-7610.2011.02446.x] [Medline: 21815892]
2. Lu W. Adolescent depression: national trends, risk factors, and healthcare disparities. Am J Health Behav 2019 Jan 01;43(1):181-194. [doi: 10.5993/AJHB.43.1.15] [Medline: 30522576]
3. Maughan B, Collishaw S, Stringaris A. Depression in childhood and adolescence. J Can Acad Child Adolesc Psychiatry 2013 Feb;22(1):35-40 [FREE Full text] [Medline: 23390431]
4. Clayborne ZM, Varin M, Colman I. Systematic review and meta-analysis: adolescent depression and long-term psychosocial outcomes. J Am Acad Child Adolesc Psychiatry 2019 Jan;58(1):72-79. [doi: 10.1016/j.jaac.2018.07.896] [Medline: 30577941]
5. Copeland WE, Alaie I, Jonsson U, Shanahan L. Associations of childhood and adolescent depression with adult psychiatric and functional outcomes. J Am Acad Child Adolesc Psychiatry 2021 May;60(5):604-611 [FREE Full text] [doi: 10.1016/j.jaac.2020.07.895] [Medline: 32758528]
6. Weavers B, Heron J, Thapar AK, Stephens A, Lennon J, Bevan Jones R, et al. The antecedents and outcomes of persistent and remitting adolescent depressive symptom trajectories: a longitudinal, population-based English study. Lancet Psychiatry 2021 Dec;8(12):1053-1061. [doi: 10.1016/S2215-0366(21)00281-9] [Medline: 34672991]
7. Danese A, Smith P. Debate: recognising and responding to the mental health needs of young people in the era of COVID-19. Child Adolesc Ment Health 2020 Sep 18;25(3):169-170 [FREE Full text] [doi: 10.1111/camh.12414] [Medline: 32812356]
8. Hill RM, Rufino K, Kurian S, Saxena J, Saxena K, Williams L. Suicide ideation and attempts in a pediatric emergency department before and during COVID-19. Pediatrics 2021 Mar 16;147(3):e2020029280. [doi: 10.1542/peds.2020-029280] [Medline: 33328339]
9. Liang L, Ren H, Cao R, Hu Y, Qin Z, Li C, et al. The effect of COVID-19 on youth mental health. Psychiatr Q 2020 Sep 21;91(3):841-852 [FREE Full text] [doi: 10.1007/s11126-020-09744-3] [Medline: 32319041]
10. Power E, Hughes S, Cotter D, Cannon M. Youth mental health in the time of COVID-19. Ir J Psychol Med 2020 Dec 02;37(4):301-305 [FREE Full text] [doi: 10.1017/ipm.2020.84] [Medline: 32611470]

11. Davenport TA, Cheng VW, Iorfino F, Hamilton B, Castaldi E, Burton A, et al. Flip the clinic: a digital health approach to youth mental health service delivery during the COVID-19 pandemic and beyond. JMIR Ment Health 2020 Dec 15;7(12):e24578 [FREE Full text] [doi: 10.2196/24578] [Medline: 33206051]
12. Kannarkat JT, Smith NN, McLeod-Bryant SA. Mobilization of telepsychiatry in response to COVID-19-moving toward 21 century access to care. Adm Policy Ment Health 2020 Jul;47(4):489-491 [FREE Full text] [doi: 10.1007/s10488-020-01044-z] [Medline: 32333227]
13. Oblath R, Twohy E, Higdon C, Duncan A, Folk JB, Schiel MA, et al. The provision and utilization of telehealth within academic mental health clinics in North America during the COVID-19 pandemic. JAACAP Open 2023 Nov;1(3):218-229. [doi: 10.1016/j.jaacop.2023.08.003]
14. Tolou-Shams M, Folk J, Stuart B, Mangurian C, Fortuna L. Rapid creation of child telemental health services during COVID-19 to promote continued care for underserved children and families. Psychol Serv 2022;19(Suppl 2):39-45 [FREE Full text] [doi: 10.1037/ser0000550] [Medline: 34110866]
15. Fortuna LR. Editorial: disparities in access to child psychiatric services: can we shift the landscape? J Am Acad Child Adolesc Psychiatry 2022 Jul;61(7):862-863. [doi: 10.1016/j.jaac.2022.01.006] [Medline: 35065189]
16. Islam S, Sanchez AL, McDermott CL, Clapp D, Worley J, Becker-Haimes EM. To proceed via telehealth or not? Considerations for pediatric anxiety and related disorders beyond COVID-19. Cogn Behav Pract (Forthcoming) 2023 Mar 21 [FREE Full text] [doi: 10.1016/j.cbpra.2023.01.004] [Medline: 37363367]
17. Curry JF, Meyer AE. Evidence-based psychosocial treatments of depression in adolescents. In: Steele RG, Roberts MC, editors. Handbook of Evidence-Based Therapies for Children and Adolescents: Bridging Science and Practice. Cham, Switzerland: Springer; 2020:121-136.
18. Klein JB, Jacobs RH, Reinecke MA. Cognitive-behavioral therapy for adolescent depression: a meta-analytic investigation of changes in effect-size estimates. J Am Acad Child Adolesc Psychiatry 2007 Nov;46(11):1403-1413 [FREE Full text] [doi: 10.1097/chi.0b013e3180592aaa] [Medline: 18049290]
19. Maalouf FT, Brent DA. Child and adolescent depression intervention overview: what works, for whom and how well? Child Adolesc Psychiatr Clin N Am 2012 Apr;21(2):299-312, viii. [doi: 10.1016/j.chc.2012.01.001] [Medline: 22537728]
20. Martin F, Oliver T. Behavioral activation for children and adolescents: a systematic review of progress and promise. Eur Child Adolesc Psychiatry 2019 Apr 23;28(4):427-441 [FREE Full text] [doi: 10.1007/s00787-018-1126-z] [Medline: 29476253]
21. Hu T, Zhang D, Yang Z. The relationship between attributional style for negative outcomes and depression: a meta-analysis. J Soc Clin Psychol 2015 Apr;34(4):304-321. [doi: 10.1521/jscp.2015.34.4.304]
22. Kanter JW, Manos RC, Bowe WM, Baruch DE, Busch AM, Rusch LC. What is behavioral activation? A review of the empirical literature. Clin Psychol Rev 2010 Aug;30(6):608-620. [doi: 10.1016/j.cpr.2010.04.001] [Medline: 20677369]
23. Keles S, Idsoe T. A meta-analysis of group Cognitive Behavioral Therapy (CBT) interventions for adolescents with depression. J Adolesc 2018 Aug 26;67(1):129-139. [doi: 10.1016/j.adolescence.2018.05.011] [Medline: 29957492]
24. Rosselló J, Bernal G, Rivera-Medina C. Individual and group CBT and IPT for Puerto Rican adolescents with depressive symptoms. Cultur Divers Ethnic Minor Psychol 2008 Jul;14(3):234-245. [doi: 10.1037/1099-9809.14.3.234] [Medline: 18624588]
25. Luo C, Sanger N, Singhal N, Pattrick K, Shams I, Shahid H, et al. Addendum to ‘a comparison of electronically-delivered and face to face cognitive behavioural therapies in depressive disorders: a systematic review and meta-analysis’ [EClinicalMedicine 24 (2020) 100442]. EClinicalMedicine 2020 Oct;27:100580. [doi: 10.1016/j.eclinm.2020.100580]
26. Grist R, Croker A, Denne M, Stallard P. Technology delivered interventions for depression and anxiety in children and adolescents: a systematic review and meta-analysis. Clin Child Fam Psychol Rev 2019 Jun 18;22(2):147-171 [FREE Full text] [doi: 10.1007/s10567-018-0271-8] [Medline: 30229343]
27. Lamb T, Pachana NA, Dissanayaka N. Update of recent literature on remotely delivered psychotherapy interventions for anxiety and depression. Telemed J E Health 2019 Aug;25(8):671-677. [doi: 10.1089/tmj.2018.0079] [Medline: 30300082]
28. Stasiak K, Fleming T, Lucassen MF, Shepherd MJ, Whittaker R, Merry SN. Computer-based and online therapy for depression and anxiety in children and adolescents. J Child Adolesc Psychopharmacol 2016 Apr;26(3):235-245. [doi: 10.1089/cap.2015.0029] [Medline: 26465266]
29. Ebert DD, Zarski A, Christensen H, Stikkelbroek Y, Cuijpers P, Berking M, et al. Internet and computer-based cognitive behavioral therapy for anxiety and depression in youth: a meta-analysis of randomized controlled outcome trials. PLoS One 2015;10(3):e0119895 [FREE Full text] [doi: 10.1371/journal.pone.0119895] [Medline: 25786025]
30. Eckshtain D, Kuppens S, Ugueto A, Ng MY, Vaughn-Coaxum R, Corteselli K, et al. Meta-analysis: 13-year follow-up of psychotherapy effects on youth depression. J Am Acad Child Adolesc Psychiatry 2020 Jan;59(1):45-63. [doi: 10.1016/j.jaac.2019.04.002] [Medline: 31004739]
31. Kennard BD, Clarke GN, Weersing VR, Asarnow JR, Shamseddeen W, Porta G, et al. Effective components of TORDIA cognitive-behavioral therapy for adolescent depression: preliminary findings. J Consult Clin Psychol 2009 Dec;77(6):1033-1041 [FREE Full text] [doi: 10.1037/a0017411] [Medline: 19968380]
32. Fernandez E, Salem D, Swift JK, Ramtahal N. Meta-analysis of dropout from cognitive behavioral therapy: magnitude, timing, and moderators. J Consult Clin Psychol 2015 Dec;83(6):1108-1122. [doi: 10.1037/ccp0000044] [Medline: 26302248]

33. Clarke AM, Kuosmanen T, Barry MM. A systematic review of online youth mental health promotion and prevention interventions. J Youth Adolesc 2015 Jan;44(1):90-113. [doi: 10.1007/s10964-014-0165-0] [Medline: 25115460]
34. Lenhart A. Teens, social media and technology overview 2015. Pew Research Center. 2015. URL: https://policycommons.net/artifacts/619187/teens-social-media-technology-overview-2015/1600266/ [accessed 2024-01-29]
35. Kauer SD, Reid SC, Sanci L, Patton GC. Investigating the utility of mobile phones for collecting data about adolescent alcohol use and related mood, stress and coping behaviours: lessons and recommendations. Drug Alcohol Rev 2009 Jan;28(1):25-30. [doi: 10.1111/j.1465-3362.2008.00002.x] [Medline: 19320672]
36. Reid SC, Kauer SD, Dudgeon P, Sanci LA, Shrier LA, Patton GC. A mobile phone program to track young people's experiences of mood, stress and coping. Development and testing of the mobiletype program. Soc Psychiatry Psychiatr Epidemiol 2009 Jun;44(6):501-507. [doi: 10.1007/s00127-008-0455-5] [Medline: 19011723]
37. Silk JS, Steinberg L, Morris AS. Adolescents' emotion regulation in daily life: links to depressive symptoms and problem behavior. Child Dev 2003 Nov 17;74(6):1869-1880. [doi: 10.1046/j.1467-8624.2003.00643.x] [Medline: 14669901]
38. Anderson M, Kumar M. Digital divide persists even as lower-income Americans make gains in tech adoption. Pew Research Center. URL: https://policycommons.net/artifacts/616828/digital-divide-persists-even-as-lower-income-americans-make-gains-in-tech-adoption/1597530/ [accessed 2024-04-30]
39. Bohleber L, Crameri A, Eich-Stierli B, Telesko R, von Wyl A. Can we foster a culture of peer support and promote mental health in adolescence using a web-based app? A control group study. JMIR Ment Health 2016 Sep 23;3(3):e45 [FREE Full text] [doi: 10.2196/mental.5597] [Medline: 27663691]
40. Aguilera A, Hernandez-Ramos R, Haro-Ramos AY, Boone CE, Luo TC, Xu J, et al. A text messaging intervention (staywell at home) to counteract depression and anxiety during COVID-19 social distancing: pre-post study. JMIR Ment Health 2021 Nov 01;8(11):e25298 [FREE Full text] [doi: 10.2196/25298] [Medline: 34543230]
41. Bendtsen M, Müssener U, Linderoth C, Thomas K. A mobile health intervention for mental health promotion among university students: randomized controlled trial. JMIR Mhealth Uhealth 2020 Mar 20;8(3):e17208 [FREE Full text] [doi: 10.2196/17208] [Medline: 32196462]
42. MacDougall S, Jerrott S, Clark S, Campbell LA, Murphy A, Wozney L. Text message interventions in adolescent mental health and addiction services: scoping review. JMIR Ment Health 2021 Jan 08;8(1):e16508 [FREE Full text] [doi: 10.2196/16508] [Medline: 33416504]
43. Cole-Lewis H, Kershaw T. Text messaging as a tool for behavior change in disease prevention and management. Epidemiol Rev 2010;32(1):56-69 [FREE Full text] [doi: 10.1093/epirev/mxq004] [Medline: 20354039]
44. Rempel GR, Ballantyne RT, Magill-Evans J, Nicholas DB, Mackie AS. Texting teens in transition: the use of text messages in clinical intervention research. JMIR Mhealth Uhealth 2014 Nov 06;2(4):e45 [FREE Full text] [doi: 10.2196/mhealth.3232] [Medline: 25379624]
45. Whittaker R. Issues in mHealth: findings from key informant interviews. J Med Internet Res 2012 Oct 02;14(5):e129 [FREE Full text] [doi: 10.2196/jmir.1989] [Medline: 23032424]
46. Whittaker R, Stasiak K, McDowell H, Doherty I, Shepherd M, Chua S, et al. MEMO: an mHealth intervention to prevent the onset of depression in adolescents: a double-blind, randomised, placebo-controlled trial. J Child Psychol Psychiatry 2017 Sep 02;58(9):1014-1022. [doi: 10.1111/jcpp.12753] [Medline: 28573672]
47. Kobak KA, Mundt JC, Kennard B. Erratum to: integrating technology into cognitive behavior therapy for adolescent depression: a pilot study. Ann Gen Psychiatry 2016 Jan 15;15(1):2 [FREE Full text] [doi: 10.1186/s12991-015-0089-4] [Medline: 26779277]
48. Aguilera A, Bruehlman-Senecal E, Demasi O, Avila P. Automated text messaging as an adjunct to cognitive behavioral therapy for depression: a clinical trial. J Med Internet Res 2017 May 08;19(5):e148 [FREE Full text] [doi: 10.2196/jmir.6914] [Medline: 28483742]
49. Bruehlman-Senecal E, Aguilera A, Schueller SM. Mobile phone-based mood ratings prospectively predict psychotherapy attendance. Behav Ther 2017 Sep;48(5):614-623 [FREE Full text] [doi: 10.1016/j.beth.2017.01.002] [Medline: 28711112]
50. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M, Medical Research Council Guidance. Developing and evaluating complex interventions: the new Medical Research Council guidance. BMJ 2008 Sep 29;337(sep29 1):a1655 [FREE Full text] [doi: 10.1136/bmj.a1655] [Medline: 18824488]
51. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ 2021 Sep 30;374:n2061 [FREE Full text] [doi: 10.1136/bmj.n2061] [Medline: 34593508]
52. Miranda J, Woo S, Lagomasino I, Hepner K, Wiseman S, Muñoz R. BRIGHT CBT depression group manuals. i4Health, Palo Alto University. 2006. URL: https://i4health.paloaltou.edu/ [accessed 2017-04-05]
53. Creswell JW, Clark VL. Designing and Conducting Mixed Methods Research. Thousand Oaks, CA: Sage Publications; 2007.

Abbreviations

BRIGHT: Building Recovery by Improving Goals, Habits, and Thoughts

CBT: cognitive behavioral therapy
CBT-D: Cognitive Behavioral Therapy for Depression Group for Adolescents
dHealth: digital health
EBI: evidence-based intervention
eCBT: electronic cognitive behavioral therapy
mHealth: mobile health
SIP: shelter-in-place
UCSF: University of California San Francisco

Edited by J Torous; submitted 24.05.23; peer-reviewed by N Dippel, P Watson; comments to author 27.07.23; revised version received 19.12.23; accepted 11.01.24; published 19.02.24.

Please cite as:
Haack LM, Armstrong CC, Travis K, Aguilera A, Darrow SM
HealthySMS Text Messaging System Adjunct to Adolescent Group Cognitive Behavioral Therapy in the Context of COVID-19 (Let's Text!): Pilot Feasibility and Acceptability Study
JMIR Ment Health 2024;11:e49317
URL: https://mental.jmir.org/2024/1/e49317
doi:10.2196/49317
PMID:38373030

Review

Translating Suicide Safety Planning Components Into the Design of mHealth App Features: Systematic Review

Kim Gryglewicz[1*], MSW, PhD; Victoria L Orr[2*], BS; Marissa J McNeil[2*], MA; Lindsay A Taliaferro[3], MPH, PhD, CHES; Serenea Hines[2], MSW; Taylor L Duffy[2], BS; Pamela J Wisniewski[4*], PhD

[1]School of Social Work, University of Central Florida, Orlando, FL, United States

[2]Center for Behavioral Health Research & Training, University of Central Florida, Orlando, FL, United States

[3]Department of Population Health Sciences, University of Central Florida, Orlando, FL, United States

[4]Department of Computer Science, Vanderbilt University, Nashville, TN, United States

[*]these authors contributed equally

Corresponding Author:
Kim Gryglewicz, MSW, PhD
School of Social Work
University of Central Florida
12805 Pegasus Drive HS I
Orlando, FL, 32816
United States
Phone: 1 14078232954
Email: kgryglew@ucf.edu

Abstract

Background: Suicide safety planning is an evidence-based approach used to help individuals identify strategies to keep themselves safe during a mental health crisis. This study systematically reviewed the literature focused on mobile health (mHealth) suicide safety planning apps.

Objective: This study aims to evaluate the extent to which apps integrated components of the safety planning intervention (SPI), and if so, how these safety planning components were integrated into the design-based features of the apps.

Methods: Following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, we systematically analyzed 14 peer-reviewed studies specific to mHealth apps for suicide safety planning. We conducted an analysis of the literature to evaluate how the apps incorporated SPI components and examined similarities and differences among the apps by conducting a comparative analysis of app features. An independent review of SPI components and app features was conducted by downloading the available apps.

Results: Most of the mHealth apps (5/7, 71%) integrated SPI components and provided customizable features that expanded upon traditional paper-based safety planning processes. App design features were categorized into 5 themes, including interactive features, individualized user experiences, interface design, guidance and training, and privacy and sharing. All apps included access to community supports and revisable safety plans. Fewer mHealth apps (3/7, 43%) included interactive features, such as associating coping strategies with specific stressors. Most studies (10/14, 71%) examined the usability, feasibility, and acceptability of the safety planning mHealth apps. Usability findings were generally positive, as users often found these apps easy to use and visually appealing. In terms of feasibility, users preferred using mHealth apps during times of crisis, but the continuous use of the apps outside of crisis situations received less support. Few studies (4/14, 29%) examined the effectiveness of mHealth apps for suicide-related outcomes. Positive shifts in attitudes and desire to live, improved coping strategies, enhanced emotional stability, and a decrease in suicidal thoughts or self-harm behaviors were examined in these studies.

Conclusions: Our study highlights the need for researchers, clinicians, and app designers to continue to work together to align evidence-based research on mHealth suicide safety planning apps with lessons learned for how to best deliver these technologies to end users. Our review brings to light mHealth suicide safety planning strategies needing further development and testing, such as lethal means guidance, collaborative safety planning, and the opportunity to embed more interactive features that leverage the advanced capabilities of technology to improve client outcomes as well as foster sustained user engagement beyond a crisis. Although preliminary evidence shows that these apps may help to mitigate suicide risk, clinical trials with larger sample sizes and more robust research designs are needed to validate their efficacy before the widespread adoption and use.

(JMIR Ment Health 2024;11:e52763) doi:10.2196/52763

KEYWORDS

suicide prevention; suicide safety planning; mobile health; mHealth apps; eHealth; digital health; systematic review; Preferred Reporting Items for Systematic Reviews and Meta-Analyses; PRISMA

Introduction

Background

Suicide is one of the leading causes of death in the United States, accounting for >45,000 deaths annually [1]. Over the last decade, suicide rates have doubled for youth aged 10 to 24 years [2] and have steadily increased for racial and ethnic minority youth [1,3,4]. Suicide ideation and attempt rates have also risen [5,6], especially among youth and minoritized populations [5,7-11]. Numerous studies have shown that untreated mental illness, limited or lack of available care, and low perceived need for mental health treatment are common, yet preventable, suicide risk antecedents [12-19]. Moreover, stigma, difficulties recognizing suicide warning signs, preferences for self-reliance and autonomy, fear of burdening others, and negative treatment experiences can negatively affect help-seeking intentions and engagement in mental health services [20-24].

Researchers have identified various suicide prevention strategies to reduce the public health problem of suicide [25,26]. Safety planning is an integral component of suicide care [27] and has been empirically validated for reducing suicidality [28,29]. The process of safety planning involves collaboration between a clinical and client, as well as with the at-risk individual and their support network. This means that the support network could also be part of the safety planning process [30]. Safety planning involves jointly identifying, problem-solving, and communicating strategies to keep an individual safe if a crisis arises. Core strategies focus on uncovering warning signs or triggers that precede an emotional event, identifying and reinforcing the use of healthful self-management strategies to cope with distress, encouraging the use of positive socialization strategies for distraction and support, creating a network of external support and professional contacts to solicit assistance and support, and reducing access to lethal means [31]. The individualized nature of creating a safety plan (ie, a written document detailing the plan to keep an individual safe during a crisis) allows the person at risk of suicide the ability to incorporate culturally relevant and meaningful strategies, thereby making these plans useful and relevant for diverse populations [30,32].

Suicide safety planning is a brief intervention that has been used in both acute and clinical settings [31,33,34] and as a self-help tool [35]. Overall, researchers have found this intervention to be feasible, acceptable, and useful to facilitate support and reduce suicide risk [32,33,35-37]. Researchers have found safety plans and related interventions, such as crisis response planning [38], to be effective in reducing the risk of hospitalization, increasing engagement in mental health treatment, and promoting the use of healthful coping strategies when used alongside other therapeutic approaches [33,34,36,39,40]. Although safety planning has shown initial success in reducing suicidal urges and offering a sense of hope to individuals in crisis [41], some clinicians and researchers have criticized this process [42,43]. For example, safety planning encourages clinicians to revisit and update safety plans with their clients over time [44], which can prove challenging if service use barriers prevent clients from reaccessing care or if clients misplace or throw away their paper-based safety plan.

Considering these challenges, mobile health (mHealth) technologies could offer a timely and effective solution to address some of the criticisms directed at traditional safety planning methods. mHealth, particularly the use of apps, represents a common tool used by consumers with access to mobile phones [45,46]. In addition, mHealth has garnered attention as a practical and convenient method for implementing mental health interventions [47], with increase in the quantity and functionality of applications and tools resulting in increased use [48]. In general, mHealth apps have been used to effectively help individuals identify and manage symptoms of various mental health problems and conditions such as depression, anxiety, substance abuse, posttraumatic stress, and eating disorders [49,50]. Thus, incorporating mHealth apps into mental health treatment and adjunctive interventions may prove beneficial.

Furthermore, incorporating mHealth apps into established evidence-based interventions may also serve as a culturally inclusive way of disseminating treatment to younger, more technologically savvy generations who also happen to demonstrate higher rates of suicidal thoughts and behaviors than adults [6]. mHealth apps may also help address service use barriers and risk factors (eg, stigma) that hinder individuals from seeking help and participating in treatment for suicidality. Combining suicide safety planning practices with mHealth apps may combat accessibility concerns as well, including a commonly reported flaw of the traditional intervention—the reliance on a paper format [35]. Given the widespread proliferation of mHealth apps for suicide prevention, there is a need to examine the components and features that have been incorporated into the design of suicide safety planning apps.

Objectives

The purpose of this systematic literature review was to first assess the extent to which suicide safety planning mHealth apps integrated the 6 steps or components of a widely used safety planning intervention (SPI) developed by Stanley and Brown [31] (research question [RQ] 1). Next, we independently reviewed available mHealth suicide safety planning apps via download from iOS and Android app stores to assess the integration of SPI components and to categorize different app design features used to personalize the end users' experience (RQ2). We also examined the evidence on the effectiveness of these apps in terms of usability, acceptability, app engagement, and suicide-related outcomes (RQ3). This review aims to synthesize the extant research to inform suicide prevention efforts, clinical practice, and future development of suicide safety planning mHealth apps.

Methods

Overview

In accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement guidelines (Multimedia Appendix 1 [51]), a comprehensive systematic review of existing literature on suicide safety planning via mHealth apps was conducted. The process is described in the following sections.

Systematic Literature Review

Eligibility Criteria

The inclusion criteria for the reviewed research studies were as follows: (1) a primary focus on suicide safety planning involving the use of a mHealth app, (2) publication in a peer-reviewed article written in English, and (3) availability of the full text of the article. Studies were excluded if (1) the word suicide, safety plan, or app was not included in the title; (2) they included other forms of mHealth technologies as the primary focus (eg, web-based applications); (3) the apps were designed with safety planning as a secondary focus (ie, not exclusively for suicide safety planning, not intended as a crisis intervention, or use of safety planning as a secondary tool to other treatment modalities); and (4) they were part of other systematic reviews or meta-analyses. We included studies across the entire system development life cycle (eg, formative evaluations and 1 group pre-posttest designs) owing to limited research on the topic and the relatively recent emergence of such research.

Information Sources

The following 5 bibliographic databases were used to systematically review the literature: PsycINFO, PubMed, ACM Digital Libraries, Academic Search Premier, and ERIC. We limited our results to articles published between January 2000 and May 2023. All databases were last searched on July 2, 2023.

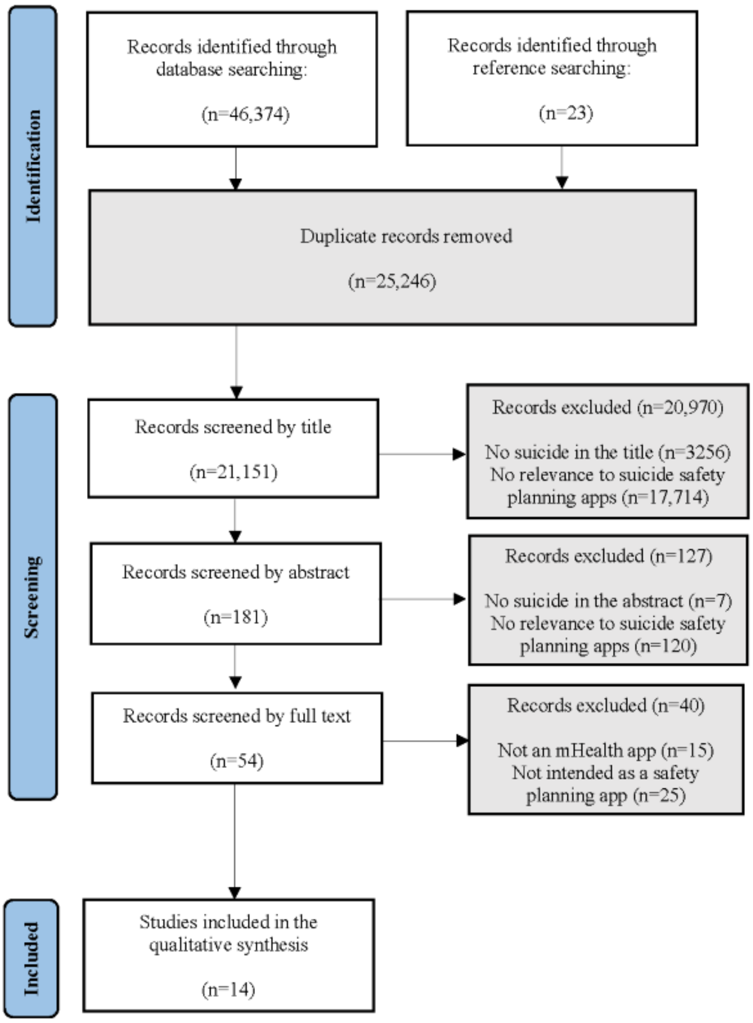
Search Strategy

We used the following keywords to search for the topic of interest in each scientific database: “Safety Plan*” AND (“Applications” OR “Apps”); (“Suicide” OR “Safety Plan*”) AND (“Applications” OR “Apps”); “Suicide Interven*” AND (“Applications” OR “Apps”); “Suicide Prevent*” AND (“Applications” OR “Apps”); “Suicide Contract” AND (“Applications” OR “Apps”); “mHealth” AND “Suicide”; “Crisis Response” AND “Plan*.” Asterisks were added to search for words that began with the preceding letters (eg, prevent*: prevent, prevention, and preventing). An example of the search strategy outlined above is provided in Multimedia Appendix 2.

Selection Process

Citations obtained from electronic databases were imported into Zotero (version 6.0.16). Two reviewers (KG and VLO) independently screened the articles to remove duplicates and assessed inclusion and exclusion criteria by title and abstract. For articles about which the reviewers were uncertain after the title and abstract review, 4 reviewers independently analyzed the full-text articles to determine whether they met the inclusion criteria. The reviewers discussed discrepancies until they reached a consensus. The references of all articles that met the inclusion criteria were reviewed and cross-referenced for additional relevant articles. We included all eligible studies (N=14) in this systematic review (Figure 1 [51]).

Figure 1. Flowchart of the studies in line with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. mHealth: mobile health.



Data Collection Process

Data from eligible studies were analyzed using the Cochrane Collaboration's data extraction template for included studies (version 1.8) [52]. We added study-specific items to the template to answer RQ1 and RQ2. Specifically, to answer RQ1, we reviewed articles describing each mHealth app and coded, using a dichotomous (yes or no) coding scheme, for the following SPI components: (1) personal warning signs, (2) coping strategies, (3) ways to distract oneself through social activities, (4) identification of and ways to access trusted individuals (eg, family and friends) for support, (5) identification of and ways to access community supports (eg, mental health professionals, nonmental health adult supports, crisis, or emergency services), and (6) information about keeping the environment safe (eg, restricting access to lethal means). To answer RQ2, we downloaded available mHealth apps via the Apple App Store or Google Play Store or contacted app developers to conduct an independent review of SPI components and app features described in the articles. Next, we created codes to describe app features, organized and categorized codes based on similarities, and generated 5 themes to capture the core aspects of features. To answer RQ3, we extracted both qualitative and quantitative findings reported on primary and secondary outcomes. We categorized the study outcomes into 3 main research themes.

Two reviewers coded 2 research articles to assess interrater reliability based on the coding template and made refinements as necessary (eg, added operational definitions to describe SPI components and provided examples of app features). Once finalized, the reviewers used the template to extract the data from the remaining studies. Data items included (1) general article information (eg, author, publication year, and country); (2) study methods (eg, aims and research design); (3) study characteristics (eg, sample size, sample demographics, and setting); (4) SPI intervention characteristics (RQ1); (5) mHealth app design features (RQ2) and primary and secondary outcomes (RQ3); and (6) study implications and future directions (Multimedia Appendix 3 [42,43,53-64]). A similar process was used to independently code the SPI components and app features of the mHealth apps available for download.

Risk of Bias Assessment

The risk of bias for each study was assessed by 2 independent reviewers (KG and VLO) using Joanna Briggs Institute (JBI) appraisal tools for quasi-experimental [65] and qualitative research study designs [66]. For studies that included mixed methods designs, we used both tools as recommended by the

JBI. Each appraisal tool used a rating scale with yes, no, unclear, and nonapplicable responses. The overall appraisal rating was based on the following categories: include, exclude, and seek further information. Disagreements between the reviewers were discussed until they reached a consensus (Multimedia Appendix 4 [65,66]).

Synthesis of Results

Owing to the heterogeneity of the study designs, participants, and outcomes collected, we could not perform a meta-analysis of the identified studies in this review. Therefore, we present a narrative synthesis of the study findings.

Results

Study Selection

The initial search of electronic databases and hand-searched references resulted in a total of 46,397 peer-reviewed articles. After duplicate records were removed, 21,151 studies remained. Titles were screened for relevancy (eg, relating to suicide, suicide safety planning, and mHealth apps), and 20,970 articles were excluded. A total of 181 abstracts were reviewed. Following full-text reviews of 54 articles, 40 articles were removed (15 studies did not include an mHealth app and 25 were not intended as a suicide safety planning app). A total of 14 articles met the inclusion criteria (refer to Figure 1 for breakdown).

Study Characteristics

Overview

The detailed study characteristics of the selected articles (N=14) are presented in Multimedia Appendix 3. Most studies (12/14, 86%) were conducted outside the United States [42,53-63]. The year range of the selected articles was between 2015 and 2023.

Study Design

As shown in Tables 1 and 2, a total of 7 mHealth suicide safety planning apps were studied across the 14 articles in our data set (Multimedia Appendix 3). We classified the articles based on the research design (ie, formative feedback, usability assessment, single cohort pre-posttest, and random control trial protocol). Formative designs assessed SPI components and features to guide app development [43,56,61,64], whereas usability designs assessed interface design issues and functionality (eg, task difficulty and time to complete tasks) [55,60,61,64]. Other studies evaluated the acceptability or feasibility of a fully developed mHealth app [54,58-60,62,63]. Across these studies, participants rated the frequency and duration of app use; ease of navigation; and level of satisfaction, comfort, confidence, or engagement in using the app.

Table 1. Formative, usability, and acceptability assessments of mobile health suicide safety planning apps (n=8 articles).

Safety planning app	Year span	Formative feedback only	Usability assessment	Random control trial	Use period	Key findings
Unnamed [43]	2015	Clinicians (n=9), at-risk youth, and guardians (n=20)	N/A[a]	N/A	None	Qualitative feedback to inform app design; no app developed
SERO [56]	2022	Clinicians, at-risk individuals (n=11), and relatives (n not reported)	N/A	N/A	None	Summary of 6 suicide prevention strategies aligned with app design
ED-SAFE [64]	2023	Clinicians and subject matter experts (n=7), and at-risk adults (n=6)	Emergency department patients after discharge (N=14)	N/A	Unclear	High usability scores and low uptake (use); no significant clinical outcomes reported
MYPLAN [53,54,62]	2016, 2017, and 2020	At-risk youth, adults, relatives, and clinicians (n=26)	N/A	Protocol only (N=546)	Familiarity with app	Qualitative feedback only; no clinical outcomes assessed
SafePlan [57,61]	2020 and 2023	Clinicians and experts (N=15)	Students (N=18)	Protocol only (N=80)	Single use	High usability scores; no clinical outcomes assessed

[a]N/A: not applicable.

Table 2. Pre-posttest assessments of mobile health suicide safety planning apps (n=6 articles).

Safety planning app name	Year span	Formative feedback only	Usability assessment	Single cohort pre-posttest	Use period	Key findings
Brake of My Mind [55]	2020	Expert heuristic evaluations (N=5)	Clinician interviews (N=6)	At-risk youth (N=3)	1 week	High usability scores; sample size too low to assess clinical efficacy
BackUp [42,60,63]	2017, 2018, and 2022	Designed informed by the expert panel (n=8) and at-risk adults (n=21)	Usability assessment combined with pre-posttests	Protocol only, single cohort design (N=80); at-risk adults (n=21) and at-risk adults (n=12)	1 week and 3 months	High usability scores, no significant decrease in suicidal thoughts; low clinical and patient uptake (use)
BeyondNow [58,59]	2019 and 2020	N/A[a]	N/A	At-risk youth and adults (n=22) as well as at-risk youth (n=17)	2 months and 6 weeks	High usability scores. Significant reduction in severity and intensity of suicidal ideation; significant increase in coping; and no significant change in suicide resilience for mixed samples. For youth only sample, no significant decrease in suicidal thoughts and significant increase in suicidal resilience. No conclusions regarding clinical efficacy.

[a]N/A: not applicable.

Sample Characteristics

Across studies, the study sample varied in age, type of participant (eg, youth or adults at risk of suicide and clinicians collaborating with suicidal clients), and setting (eg, suicide prevention clinic and pediatric inpatient facility). Among studies that recruited participants to inform or evaluate mHealth suicide safety planning apps [43,54-56,58-64], the sample size ranged from 11 to 36 participants. However, after reporting dropout rates, sample sizes dropped to as low as 2 participants and as high as 22 participants.

Integration of SPI Components Within mHealth Apps

Most articles (5/7, 71%) describing the mHealth apps incorporated SPI components into the design of their apps [54,58,61,63,64] (Table 3). Creating a safe environment from lethal means was the missing component in 29% (2/7) of the apps [55,56].

Table 3. Safety planning intervention (SPI) components and app features.

	MYPLAN (Min-Plan) [54,62]	BoMM[a] [55]	BeyondNow [58,59]	SafePlan [61]	BackUp [63]	ED-SAFE [64]	SERO [56]
SPC[b]							
SPI 1: warning signs	X[c]	X	X	X	X	X	X
SPI 2: coping strategies	X	X	X	X	X	X	X
SPI 3: distractions social activities	X	X	X	X	X	X	X
SPI 4: trusted supports (family and friends)	X	X	X	X	X	X	X
SPI 5: community supports (MHP[d])	X	X	X	X	X	X	X
SPI 6: safe environment (lethal means)	X	—[e]	X	X	X	X	—
Interactive features[f]							
Links stressors to cope strategies; SPI 1 and 2	X	—	—	X	—	—	X
Inclusion of media (distraction); SPI 3	X	X	✓[g]	X	X	X	—
Access to trusted supports; SPI 4	X	X	X	X	X	X	X
Access to community supports; SPI 5	X	X	X	X	X	X	X
GPS tracking; SPI 5 and 6	X	X	—	✓	—	—	—
Individualized user experience							
Revisable safety plan	X	X	X	X	X	X	X
Personality and mood exercises or tracking	—	X	—	X	—	—	X
Visual customization	✓	X	✓	X	X	—	X
Enabling notifications	X	—	—	X	✓	—	✓
Interface design							
Easy to navigate	X	X	X	X	X	X	✓
Guidance and training							
In-app tutorial	X	X	✓	—	X	X	✓
Privacy and sharing							
Secure username and password	✓	X	—	X	✓	X	✓
Shareable data and safety plan	X	—	X	X	—	X	✓

[a]BoMM: Brake of My Mind.
[b]SPC: safety planning component.
[c]SPC or app feature included in the app.
[d]MHP: mental health professional.
[e]SPC or app feature missing in the app.
[f]Denotes innovative app features aligned with SPI components.
[g]Feature included in the app that was not mentioned in the article.

We used the JBI quasi-experimental appraisal tool [65] to assess the risk of bias across 5 studies [55,58-60,63]. These studies did not include a control or comparison group, increasing the threat to internal validity. Pre- and posttest measures were used to assess the immediate effects of the mHealth apps. However, the lack of repeated outcome measures over time, selection bias (nonrandom samples), and small sample sizes pose a risk of bias within and across these studies.

The qualitative appraisal checklist tool [66] was used to assess the risk of bias in 4 studies [43,54,56,62]. Across 2 studies [43,54], the cultural or theoretical orientation of the researchers and their influence on the research process was unclear. These issues were noted in the other 2 studies [56,62] as well. In these studies [56,62], it was also difficult to identify the philosophical perspective and congruity between the research methods, data analysis, and interpretation. The studies included more of a description of the design of the apps and included general perceptions from stakeholders.

The remaining studies [61,64] were assessed using both the quasi-experimental and qualitative appraisal tools owing to their mixed methods designs. In both studies, it was unclear whether the researchers' cultural or theoretical orientation, their influence on the research, and the adequate representation of the participants and their voices were addressed. Other key issues included the lack of a control or comparison group, nonrandom and small sample sizes, and the use of posttest measures to assess usability at only 1 time point. JBI appraisal results are included in Multimedia Appendix 4.

On the basis of our independent review of available mHealth suicide safety planning apps, SPI components described in each article were verified in 71% (5/7) of the apps [54,56,58,61,63]. The app features described in the articles were also confirmed in these apps. App features not highlighted in the articles but found within the apps are listed in Table 3. We were unable to verify SPI components and app features in 2 of the reviewed apps in the literature [55,64].

Comparative Analysis of SPI Components and App Features

In our analysis of the literature and available mHealth apps for download, we synthesized the commonalities of app features and categorized them into 5 broad themes: interactive features, individualized user experience, interface design, guidance and training, and privacy and sharing. These features are described in the following sections.

Interactive Features

Three of the suicide safety planning mHealth apps [54,56,61] allowed users to associate suicide warning signs or precipitating stressors with their personalized coping strategies (aligns with SPI 1 and 2 in Table 3). O'Grady et al [61] stressed the importance of including this feature in apps, as this functionality can serve to preemptively address an impending crisis before it fully manifests. Most of the suicide safety planning mHealth apps (6/7, 86%) also included social distractor features in which users had access to their phone's camera with the ability to upload or view media content (eg, pictures, quotes, music, activities, videos, and inspirational stories; SPI 3) [54,55,58,61,63,64]. In the *BackUp* app [63], loved ones, trusted supports, and suicidal users were able to upload media and share content to inspire hope and distract users from negative thinking.

Each mHealth app also included a built-in feature for users to save and contact trusted individuals within their social support networks (SPI 4). Typically, users entered contact information into the mHealth app directly or linked to their contact directories. A unique feature of the *MYPLAN* app [54] allowed users to create prewritten messages that they could send to their social supports during times of distress. Although this feature was created to inform loved ones of the app user's emotional state during a crisis, participants (ie, app users) noted concerns about messages being misunderstood, whereas relatives felt that messages could minimize emotional states or provide inaccurate information about the app user's safety. All apps included the ability to access community supports such as mental health professionals (SPI 5). Three apps [54,55,61] included GPS capabilities, which enabled users to search for nearby counseling agencies or emergency services, and, after selecting a search result, users received directions for quick access (SPI 5 and 6). The *ED-SAFE* app [64] included a referral search engine that allowed users to find behavioral health care by specialty and zip code. Emergency service numbers, mostly displayed via a phone icon or brief words (eg, "Crisis"), were clearly visible (listed on all pages) in 57% (4/7) of the mHealth apps [56,58,61,63], which is the suggested ethical guideline from prior work [67]. Three apps did not include access to emergency service numbers on all pages but provided them somewhere else within the app [54,55,64].

Individualized User Experience

All apps (7/7, 100%) allowed users to continually add to or revise their safety plans. Examples included the addition of new warning signs, reasons for living, and identifying coping strategies. None of the apps maintained a historical record of the previous safety plans or provided a visual mechanism to track daily, weekly, or monthly patterns based on stressors encountered or coping strategies used. Other personalization aspects included the ability to enable or disable therapeutic modalities [61], the inclusion of web-based resources to take an aptitude and personality test [55], exercises to express moods [55], and mood tracking [55,56,61]. In addition, all apps had built-in features to make esthetic customizations, such as personalizing the home screen, changing the color palate, and adding background pictures [54-56,58,61,63]. In 57% (4/7) of the apps, notifications were enabled to remind users about using their safety plan or skills to practice [54,56,61,63].

Interface Design

Several studies used iterative feedback from content and app design experts to create easy-to-navigate interfaces [58,61,63]. To enhance the navigation experience, a simple layout, clear or user-friendly language, and accessibility features were important design considerations included in some mHealth apps [54,58,61,64]. For example, *SafePlan*'s layout mimicked the paper version of the safety plan to better transition users from using the paper version to the app [61].

Guidance and Training

In-app tutorials or instructional videos were included in 86% (6/7) of the suicide safety planning mHealth apps [54-56,58,63,64]. Some of these tutorials focused on how to use the app, whereas others explained the safety planning process. For example, the *BeyondNow* app [58] included a video outlining the process of safety planning and links to other helpful information. The most extensive tutorials were seen in the companion app to *ED-SAFE* [64], where tutorials could be received from a female provider, a male community member, or an avatar. The mHealth suite of apps also included self-care education materials about suicidality, safety plans, and life plans. In addition, the *BackUp* app [63] provided supportive contacts with web-based information on ways to identify warning signs and strategies to talk with suicidal individuals. The *Brake of My Mind* app [55] included an introduction from the developer with additional web-based resources to increase app usability.

Privacy and Sharing

Researchers also highlighted app privacy and sharing capabilities as important features to consider when designing mHealth suicide safety planning apps. Given the personal nature of the information saved, most mHealth apps required a username and password to log in [54-56,61,63,64]. For example, *ED-SAFE* [64] used the username and password feature to verify user identity and connect information collected in the emergency department setting to the mHealth app. Other apps disabled GPS for location tracking or did not use external servers to store users' information for privacy and security concerns [61,63]. Several apps (5/7, 71%) included features allowing users to share self-monitoring data or share safety plans with clinicians or trusted individuals [54,56,58,61,64]. For instance, *ED-SAFE* [64] allowed users to share safety plans as well as appointment information, self-care education, helplines, referrals, and distractions through password-protected privileges given to authorized family members.

mHealth App Evidence of Effectiveness

The qualitative and quantitative findings were categorized into 3 main research themes: app usability and acceptability, app use and engagement, and suicide-related outcomes.

App Usability and Acceptability Findings

Across 71% (10/14) of the studies [54-56,58-64] that assessed the initial usability or acceptability of mHealth suicide safety planning apps, stakeholders' experiences testing the mHealth apps were generally positive. Four studies [55,60,61,64] included standard rating scales (ie, System Usability Scale [68]) to assess the perceived usability of their apps, and scores exceeded the minimum usability standards (ie, >70). The remaining studies used qualitative feedback from focus groups, case reports, and open-ended questionnaires. For example, in the study by Buus et al [54], participants found the *MYPLAN* safety planning app useful in recognizing patterns of impending crises and for reinforcing personalized strategies to cope with distress. In describing the benefits of the *BeyondNow* safety planning app, participants in the study by Melvin et al [58] reported developing a sense of hope and connection from using the app. Researchers have attributed these findings to the accessibility of the app and its customizable features. According to the authors, stakeholders regarded apps as highly intuitive, easy to use, and visually appealing interface in terms of the design [59,61,62,64].

App Use and App Engagement

Five studies examined app use over time [58-60,63,64]. Overall app engagement and use were minimal. Across 3 studies, >70% of the participants used the apps at least once during the testing period, which ranged from 1 to 10 weeks [58,59,63]. In the study by Melvin et al [58], 77% (17/22) of the participants reported using the mHealth app "occasionally" or "a lot," including to make changes to safety plans. Most participants also reported using the mHealth app during a suicidal crisis (15/22, 68%) or when experiencing suicidal ideation (18/22, 82%). Increased frequency of app use during a crisis or among participants with high levels of suicide ideation was reported in studies by Pauwels et al [63] and Muscara et al [59]. Larkin et al [64] reported that 2 (40%) out of 5 participants reported downloading the *ED-SAFE* patient mHealth app after discharge. Low uptake rates were mostly attributed to the participants' forgetfulness to download the app. Although most participants acknowledged the benefits of using mHealth suicide safety planning apps during times of crisis [58,63], participant feedback from the study by Muscara et al [59] suggested that participants did not believe or were unsure whether the use of the *BeyondNow* safety planning app could help them manage their symptoms or keep individuals safe during a crisis. Only 35% (6/17) of the participants favored using the app in the future. Conversely, participants in the study by Nuij et al [60] noted that easy access to the *Backup* mHealth app provided a sense of reassurance and helped to deter suicidal thoughts.

Suicide-Related Outcomes

Suicide-related outcomes were examined across 29% (4/14) of the small-scale pilot studies (with sample sizes ranging between 3 and 22) [55,58,59,63]. The study by Jeong et al [55] assessed the Theory of Planned Behavior constructs, including attitudes, subjective norms, perceived behavioral control, and intentions toward engaging in suicide attempts, using a pre-posttest design with a small (N=3) sample of adolescent survivors of suicide attempts. The results showed statistically significant changes in attitudes, perceived behavioral control, and intentions, suggesting that the suicide safety planning app helped to positively shift attitudes toward life and reduce beliefs and intentions to engage in self-harm behavior.

Suicide coping or resilience was evaluated in 2 studies using pre-posttest designs [58,59]. Both studies used the same safety planning app (ie, *BeyondNow*) to examine the changes in protective factors. Melvin et al [58] found a statistically significant increase in suicide-related coping among youth and adult participants (n=22). This finding suggests an increase in knowledge and confidence to use internal coping strategies and external resources to manage suicide ideation. However, the researchers did not observe statistically significant changes in suicide resilience (ie, the perceived ability to manage suicidal thoughts and feelings). In contrast, Muscara et al [59] found a significant increase in 1 subscale of suicide resilience, emotional stability (ie, the ability to regulate emotions), among youth participants (N=17) in their study.

Suicidal ideation or self-harm behavior were measured in 3 studies [58,59,63]. In an open-label, single-group design, Melvin et al [58] found statistically significant reductions in both the severity and intensity of suicide ideation following exposure to an 8-week trial that evaluated the clinical effectiveness of using the *BeyondNow* suicide safety planning app as an adjunct to treatment as usual (ie, existing mental health services). In an evaluation of the same mHealth app, but with the addition of a personalized toolbox app (ie, *BlueIce*), instead of treatment as usual, Muscara et al [59] also found a reduction in suicide ideation and self-harm behaviors (ie, attempts to harm oneself with and without suicidal intent). However, these findings were not conclusive or statistically significant owing to the small sample size and lack of a control group. Pauwels et al [63] found a similar, nonsignificant decrease in suicide ideation scores in a study examining pre-posttest changes following exposure to the *BackUp* suicide safety planning app. Although these studies provide some evidence of clinical utility, these researchers noted study limitations and the need for further evaluation using randomized controlled trials (RCTs).

Discussion

Principal Findings

The primary aim of this study was to conduct a comprehensive analysis of the integration and inclusion of the SPI components developed by Stanley and Brown [31] in the design of mHealth suicide safety planning apps. The secondary aim was to synthesize and assess the research methods of studies that reported on the effectiveness of these apps. Implications of these findings and practical recommendations for future directions in mHealth suicide safety planning research are described in the following sections.

Integrating Components of Suicide Safety Planning Into mHealth Apps

Overall, most apps included the core components of the SPI developed by Stanley and Brown [31], such as the identification of suicide warning signs, coping strategies, and supportive persons. Therefore, the results from this review provide evidence of some level of successful integration of SPI components into mHealth suicide safety planning apps (RQ1). Lethal means safety was 1 component that was not incorporated in 2 of the apps reviewed. Reducing access to lethal means is a critical part of suicide safety planning [31] and warrants inclusion in mHealth apps as it brings attention to methods that could be used to attempt or die by suicide if not removed from a user's environment.

An important aspect of suicide safety planning is access to one's safety plan. In this review, having access to safety plans at any time [54,55,58,60,61] and being able to continually revise the plan were considered benefits over traditional paper-based safety planning. In some apps, users could create associations between different suicide safety planning components (SPCs; eg, triggers and coping strategies) to better contextualize their experiences and create actional plans for mitigating crises [54,56,61]. We recommend that additional linkages between the SPCs be included to further personalize users' experiences.

Despite the integration of SPI components within mHealth suicide safety planning app designs, we also identified important gaps in the literature that warrant the attention of app designers, researchers, and mental health professionals who may use this type of technology within their clinical practice. For instance, researchers have consistently emphasized the importance of completing the initial safety plan alongside a knowledgeable clinician [42,54,58,61] to ensure that at-risk users and loved ones understand the components and purpose of a safety plan. However, many of the analyzed apps allowed users to complete the safety plan without the recommended clinical support, and in some cases, they lacked disclaimers. Therefore, additional guidance from a professional when using mHealth suicide safety planning apps would further serve to assist users and ensure that the safety planning process is carried out as intended.

This review also found that most of the apps did not go beyond the traditional SPCs of paper-based protocols to integrate more interactive features that could potentially improve adherence or engagement. For instance, daily or weekly check-ins have been shown to improve adherence in other mHealth contexts, such as for smoking cessation [69] and the management of schizophrenia [70]. Visualization graphs of patterns or trends in suicide warning signs, triggers, and coping behaviors logged over time may serve to increase engagement and improve outcomes, as visualizing behavior change over time has been recommended in other mHealth contexts [71], such as alcohol reduction [72]. Furthermore, other meaningful ways to actively and continuously engage one's support contacts (eg, clinicians, parents, and family members) and to reinforce the use of healthful coping strategies would be an advantageous direction for future exploration in mHealth app design. Beyond general support contacts, prior research has found that parental support is a significant protective factor against youth suicide [73,74]. For youth, in particular, it may be advantageous to include parents, family members, or other trusted adults in the mHealth suicide safety planning process to increase uptake, enhance help-seeking and coping behaviors, and reinforce ways to keep one's environment safe. However, future research would need to carefully design and evaluate such interventions to ensure they are effective before making these interventions widely available through the dissemination of mHealth apps for suicide safety planning.

Another variation across the apps was that some apps provided default values for suicide SPCs (eg, suggested coping strategies), whereas others did not. Therefore, an area of future research could be to study whether providing default values is beneficial or detrimental to the safety planning process. Finally, rather than training focused on the technical aspects of using the mHealth app, there is a need to include psychoeducation for suicide safety planning [75], especially related to coping strategies and lethal means restriction, which should be modeled as a collaborative process between at-risk users and their support systems [76].

Usability and Design Considerations for mHealth Suicide Safety Planning Apps

Overall, our review highlights three important recommendations to consider when designing safety planning mHealth apps (RQ2): the need to (1) encourage end user collaboration in the design and implementation of the intervention, (2) incorporate personalization or customization capabilities, and (3) develop appropriate privacy safeguards to prevent liability and address other safety concerns that may arise when integrating mental health care and technology. A key strength of most studies in our review was the interdisciplinary collaboration between app developers, computer scientists, and clinical researchers that facilitated the design, development, and evaluation of the various mHealth suicide safety planning apps. In addition, multiple stakeholders were included in the design process, including individuals at risk of suicide, clinicians, usability experts, parents, and extended family members. Only in 1 instance, end users engaged who were not considered part of the target population of at-risk users (eg, students). We strongly recommend that future research continue to include researchers from across multiple disciplines (eg, psychology, public health, social work, medicine, computer science, and human-computer interaction), intended end users, and mental health professionals across each stage of the research process. For instance, researchers from different disciplines may be able to raise important threats to validity during the research design process that could lead to more robust study designs.

A key weakness highlighted within several studies was limited uptake or sustained use of the mHealth suicide safety planning apps over time. Such findings shed suspicion on the feasibility of this type of intervention being effective outside of research, regardless of the high usability and acceptability ratings. Some studies attributed lack of use to the reduction of suicidal behaviors over time, but others suggested that the suicide safety planning process, as designed to be carried out within the apps, was only suited for in-crisis situations and not appropriate for sustained use over time. Although this may be the case, it is also possible that the lack of interactive or engaging features within the apps made them less appealing to users. Being able to customize and personalize app features may help to enhance the user's experience and increase app engagement. Many of the apps included social distractions (ie, music and pictures) or other features, such as diary cards, which might help increase overall app engagement during noncrisis periods. However, as suicidality is episodic, future research should be conducted to understand how different modalities or features (eg, mood tracking, journaling, mindfulness, and art) could be combined with suicide safety planning in a complementary way for long-term use and engagement. Future work should also consider leveraging advanced technologies and assessments, such as artificial intelligence and ecological momentary assessments [77,78], that could be used to anticipate heightened suicide risk and prompt users to engage in the mHealth app suicide safety planning process when they need it most.

Threats to Validity and Inconclusive Clinical Outcomes Associated With the Use of mHealth Suicide Safety Planning Apps

This review provides some preliminary evidence suggesting that suicide safety planning via mHealth apps could be an easy-to-use mechanism to provide individualized care to those who may otherwise go unserved due to common treatment barriers (RQ3), such as poor accessibility to service providers, lack of knowledge about suicide, and stigmatizing beliefs about help seeking [20-24]. At the same time, several threats to validity were uncovered by our assessment of risk bias, which can inform directions for future research. First, the robustness of the qualitative studies could be improved by stating the positionality of the researchers as well as a clear justification for the design of the mHealth apps. In some cases, articles were published by interdisciplinary teams, whereas in other cases, authors appeared to be from a single discipline (eg, computer science). Details about the composition and expertise of the research team are important, as well-implemented mHealth apps require interdisciplinary skill sets that span clinical, design-based, and technical expertise. Furthermore, the quantitative studies analyzed in our review were constrained by small sample sizes and no published RCTs. Among the pre-posttest studies conducted thus far, the clinical outcomes were inconclusive.

As such, RCTs with control groups, random assignment, and repeated measure outcomes assessed over time are needed in the future to evaluate the efficacy of using suicide safety planning mHealth apps compared with traditional paper-based safety plans [54,57], specifically related to reducing suicidal urges and behaviors and increasing use of coping strategies, as well as increased engagement in crisis and mental health services after the crisis. When doing so, researchers should recruit larger samples to ensure that the results are conclusive and can be generalized to the populations of interest. Furthermore, additional use metrics collected by the apps to track behavioral data associated with using different app features, such as user engagement with the 6 components of the SPI developed by Stanley and Brown [31], should be considered to better understand the potential mediating factors and behaviors that may influence clinical outcomes. Although the usability of the apps would be an important consideration to control for in future studies, it is necessary to move beyond such measures to determine the efficacy of mHealth apps in reducing suicide-related outcomes. In summary, the inclusion of more advanced study design methodologies and recommendations from lessons learned in future mHealth apps could serve to mitigate suicide risk and promote overall safety.

Limitations and Future Research

This systematic review included 14 peer-reviewed articles that designed, developed, and evaluated mHealth apps for suicide safety planning. There are several limitations of this study that should be addressed in future research. First, although our search process was comprehensive, it is possible that our keywords missed relevant articles and mHealth apps that should have been included in the review. Second, as many of the apps described in the articles were not publicly available for download, we requested access from the corresponding authors to conduct our review. In 2 cases, we were unable to gain access to the apps; therefore, our analysis was based on the description of those apps based on the published paper. As such, it may be possible that some features were not described in the original papers; thus, they were not included in our review. Future research should also consider conducting a systematic feature analysis of mHealth suicide safety planning apps that are publicly available for download but not studied within the peer-reviewed literature. Finally, a limited number of published RCTs at the time of the review restricted our ability to report on app use and suicide-related outcomes. As such, the main call-to-action from this review is the need to move beyond usability studies of newly developed mHealth suicide safety planning apps to robust clinical research designs to examine their efficacy in reducing suicidality among at-risk user populations.

Conclusions

Overall, most articles included in this review did little to evaluate the efficacy of mHealth suicide safety planning apps beyond usability assessments, signaling that these apps and corresponding research are still in their infancy in terms of validating clinical outcomes. Although most of the mHealth safety planning apps included in our review are not yet downloadable and broadly available for public use, the prevalence and popularity of mHealth suicide prevention and mental health support apps on the open market that have been deployed without rigorous peer-reviewed research is a concern. As such, there is a critical need for future research to ensure that mHealth apps for suicide safety planning integrate the lessons learned from empirical user-based and clinical research, are upheld to high ethical mental health care standards, and show clinical efficacy for reducing suicidality before the apps are released to end users. This is especially true given the delicate and important goal of preventing suicide among at-risk populations. It is promising to see that future randomized clinical trials have been registered to build upon this important preliminary work on mHealth suicide safety planning apps.

Acknowledgments

The authors would like to thank the following authors who provided access to their apps to conduct this feature analysis: Niels Buus and JL Stovgaard Larsen (MYPLAN); Glenn Melvin (BeyondNow); James Duggan (SafePlan); and Lea Meier, Caroline Gurtner, and François von Kaenel (SERO).

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 182 KB - mental_v11i1e52763_app1.pdf]

Multimedia Appendix 2
Example search strategy.
[DOCX File , 18 KB - mental_v11i1e52763_app2.docx]

Multimedia Appendix 3
Detailed summary of the selected articles and key findings (N=14).
[DOCX File , 40 KB - mental_v11i1e52763_app3.docx]

Multimedia Appendix 4
Critical appraisal results.
[DOCX File , 20 KB - mental_v11i1e52763_app4.docx]

References

1. WISQARS™ — web-based injury statistics query and reporting system. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/injury/wisqars/index.html [accessed 2023-01-11]
2. Curtin SC, Heron M. Death rates due to suicide and homicide among persons aged 10-24: United States, 2000-2017. NCHS Data Brief 2019 Oct(352):1-8. [Medline: 31751202]
3. Congressional Black Caucus. Ring the alarm: the crisis of black youth suicide in America. The National Action Alliance for Suicide Prevention. 2019 Dec. URL: https://theactionalliance.org/resource/ring-alarm-crisis-black-youth-suicide-america [accessed 2023-06-05]
4. Ramchand R, Gordon JA, Pearson JL. Trends in suicide rates by race and ethnicity in the United States. JAMA Netw Open 2021 May 03;4(5):e2111563 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.11563] [Medline: 34037735]
5. Gaylor EM, Krause KH, Welder LE, Cooper AC, Ashley C, Mack KA, et al. Suicidal thoughts and behaviors among high school students - youth risk behavior survey, United States, 2021. MMWR Suppl 2023 Apr 28;72(1):45-54 [FREE Full text] [doi: 10.15585/mmwr.su7201a6] [Medline: 37104546]
6. Plemmons G, Hall M, Doupnik S, Gay J, Brown C, Browning W, et al. Hospitalization for suicide ideation or attempt: 2008-2015. Pediatrics 2018 Jun;141(6):e20172426. [doi: 10.1542/peds.2017-2426] [Medline: 29769243]

7. Johns MM, Lowry R, Haderxhanaj LT, Rasberry CN, Robin L, Scales L, et al. Trends in violence victimization and suicide risk by sexual identity among high school students - youth risk behavior survey, United States, 2015-2019. MMWR Suppl 2020 Aug 21;69(1):19-27 [FREE Full text] [doi: 10.15585/mmwr.su6901a3] [Medline: 32817596]
8. Peña JB, Wyman PA, Brown CH, Matthieu MM, Olivares TE, Hartel D, et al. Immigration generation status and its association with suicide attempts, substance use, and depressive symptoms among latino adolescents in the USA. Prev Sci 2008 Dec;9(4):299-310 [FREE Full text] [doi: 10.1007/s11121-008-0105-x] [Medline: 18855139]
9. Johns MM, Lowry R, Andrzejewski J, Barrios LC, Demissie Z, McManus T, et al. Transgender identity and experiences of violence victimization, substance use, suicide risk, and sexual risk behaviors among high school students - 19 states and large urban school districts, 2017. MMWR Morb Mortal Wkly Rep 2019 Jan 25;68(3):67-71 [FREE Full text] [doi: 10.15585/mmwr.mm6803a3] [Medline: 30677012]
10. Stone DM, Jones CM, Mack KA. Changes in suicide rates - United States, 2018-2019. MMWR Morb Mortal Wkly Rep 2021 Feb 26;70(8):261-268 [FREE Full text] [doi: 10.15585/mmwr.mm7008a1] [Medline: 33630824]
11. Jin SS, Dolan TM, Cloutier AA, Bojdani E, DeLisi L. Systematic review of depression and suicidality in child and adolescent (CAP) refugees. Psychiatry Res 2021 Aug;302:114025. [doi: 10.1016/j.psychres.2021.114025] [Medline: 34090083]
12. Fortune S, Stewart A, Yadav V, Hawton K. Suicide in adolescents: using life charts to understand the suicidal process. J Affect Disord 2007 Jun;100(1-3):199-210. [doi: 10.1016/j.jad.2006.10.022] [Medline: 17182107]
13. Moskos M, Olson L, Halbern S, Keller T, Gray D. Utah youth suicide study: psychological autopsy. Suicide Life Threat Behav 2005 Oct;35(5):536-546. [doi: 10.1521/suli.2005.35.5.536] [Medline: 16268770]
14. Pagura J, Fotti S, Katz LY, Sareen J, Swampy Cree Suicide Prevention Team. Help seeking and perceived need for mental health care among individuals in Canada with suicidal behaviors. Psychiatr Serv 2009 Jul;60(7):943-949. [doi: 10.1176/ps.2009.60.7.943] [Medline: 19564225]
15. Sankey M, Lawrence R. Brief report: classification of adolescent suicide and risk-taking deaths. J Adolesc 2005 Dec;28(6):781-785. [doi: 10.1016/j.adolescence.2005.03.005] [Medline: 16291510]
16. Pisani AR, Wyman PA, Petrova M, Schmeelk-Cone K, Goldston DB, Xia Y, et al. Emotion regulation difficulties, youth-adult relationships, and suicide attempts among high school students in underserved communities. J Youth Adolesc 2013 Jun;42(6):807-820 [FREE Full text] [doi: 10.1007/s10964-012-9884-2] [Medline: 23666604]
17. Rajappa K, Gallagher M, Miranda R. Emotion dysregulation and vulnerability to suicidal ideation and attempts. Cogn Ther Res 2011 Nov 9;36:833-839. [doi: 10.1007/s10608-011-9419-2]
18. Hung P, Busch SH, Shih YW, McGregor AJ, Wang S. Changes in community mental health services availability and suicide mortality in the US: a retrospective study. BMC Psychiatry 2020 Apr 25;20(1):188 [FREE Full text] [doi: 10.1186/s12888-020-02607-y] [Medline: 32334552]
19. Hester RD. Lack of access to mental health services contributing to the high suicide rates among veterans. Int J Ment Health Syst 2017 Aug 18;11:47 [FREE Full text] [doi: 10.1186/s13033-017-0154-2] [Medline: 28828036]
20. Curtis C. Youth perceptions of suicide and help-seeking: ‘they'd think I was weak or “mental”’. J Youth Stud 2010 Jul 28;13(6):699-715. [doi: 10.1080/13676261003801747]
21. Aguirre Velasco A, Cruz IS, Billings J, Jimenez M, Rowe S. What are the barriers, facilitators and interventions targeting help-seeking behaviours for common mental health problems in adolescents? A systematic review. BMC Psychiatry 2020 Jun 11;20(1):293 [FREE Full text] [doi: 10.1186/s12888-020-02659-0] [Medline: 32527236]
22. Czyz EK, Horwitz AG, Eisenberg D, Kramer A, King CA. Self-reported barriers to professional help seeking among college students at elevated risk for suicide. J Am Coll Health 2013;61(7):398-406 [FREE Full text] [doi: 10.1080/07448481.2013.820731] [Medline: 24010494]
23. Hom MA, Stanley IH, Joiner TE. Evaluating factors and interventions that influence help-seeking and mental health service utilization among suicidal individuals: a review of the literature. Clin Psychol Rev 2015 Aug;40:28-39. [doi: 10.1016/j.cpr.2015.05.006] [Medline: 26048165]
24. Wilson CJ, Deane FP. Adolescent opinions about reducing help-seeking barriers and increasing appropriate help engagement. J Educ Psychol Consult 2001;12(4):345-364. [doi: 10.1207/s1532768xjepc1204_03]
25. Ougrin D, Tranah T, Stahl D, Moran P, Asarnow JR. Therapeutic interventions for suicide attempts and self-harm in adolescents: systematic review and meta-analysis. J Am Acad Child Adolesc Psychiatry 2015 Feb;54(2):97-107.e2. [doi: 10.1016/j.jaac.2014.10.009] [Medline: 25617250]
26. Pfeiffer PN, King C, Ilgen M, Ganoczy D, Clive R, Garlick J, et al. Development and pilot study of a suicide prevention intervention delivered by peer support specialists. Psychol Serv 2019 Aug;16(3):360-371 [FREE Full text] [doi: 10.1037/ser0000257] [Medline: 30382743]
27. Evidence. Zero Suicide. URL: https://zerosuicide.edc.org/evidence [accessed 2023-10-25]
28. Ferguson M, Rhodes K, Loughhead M, McIntyre H, Procter N. The effectiveness of the safety planning intervention for adults experiencing suicide-related distress: a systematic review. Arch Suicide Res 2022;26(3):1022-1045. [doi: 10.1080/13811118.2021.1915217] [Medline: 33913799]
29. Nuij C, van Ballegooijen W, de Beurs D, Juniar D, Erlangsen A, Portzky G, et al. Safety planning-type interventions for suicide prevention: meta-analysis. Br J Psychiatry 2021 Aug;219(2):419-426. [doi: 10.1192/bjp.2021.50] [Medline: 35048835]

30. Conti EC, Jahn DR, Simons KV, Edinboro LP, Jacobs ML, Vinson L, et al. Safety planning to manage suicide risk with older adults: case examples and recommendations. Clin Gerontol 2020;43(1):104-109 [FREE Full text] [doi: 10.1080/07317115.2019.1611685] [Medline: 31096885]
31. Stanley B, Brown GK. Safety planning intervention: a brief intervention to mitigate suicide risk. Cogn Behav Pract 2012 May;19(2):256-264. [doi: 10.1016/j.cbpra.2011.01.001]
32. Chesin MS, Stanley B, Haigh EA, Chaudhury SR, Pontoski K, Knox KL, et al. Staff views of an emergency department intervention using safety planning and structured follow-up with suicidal veterans. Arch Suicide Res 2017 Jan 02;21(1):127-137. [doi: 10.1080/13811118.2016.1164642] [Medline: 27096810]
33. Stanley IH, Hom MA, Joiner TE. A systematic review of suicidal thoughts and behaviors among police officers, firefighters, EMTs, and paramedics. Clin Psychol Rev 2016 Mar;44:25-44. [doi: 10.1016/j.cpr.2015.12.002] [Medline: 26719976]
34. Stanley B, Brown GK, Brenner LA, Galfalvy HC, Currier GW, Knox KL, et al. Comparison of the safety planning intervention with follow-up vs usual care of suicidal patients treated in the emergency department. JAMA Psychiatry 2018 Sep 01;75(9):894-900 [FREE Full text] [doi: 10.1001/jamapsychiatry.2018.1776] [Medline: 29998307]
35. Labouliere CD, Stanley B, Lake AM, Gould MS. Safety planning on crisis lines: feasibility, acceptability, and perceived helpfulness of a brief intervention to mitigate future suicide risk. Suicide Life Threat Behav 2020 Feb;50(1):29-41 [FREE Full text] [doi: 10.1111/sltb.12554] [Medline: 31112330]
36. Czyz EK, King CA, Biermann BJ. Motivational interviewing-enhanced safety planning for adolescents at high suicide risk: a pilot randomized controlled trial. J Clin Child Adolesc Psychol 2019;48(2):250-262. [doi: 10.1080/15374416.2018.1496442] [Medline: 30142300]
37. Stewart KL, Darling EV, Yen S, Stanley B, Brown GK, Weinstock LM. Dissemination of the safety planning intervention (SPI) to university counseling center clinicians to reduce suicide risk among college students. Arch Suicide Res 2020;24(sup1):75-85. [doi: 10.1080/13811118.2018.1531797] [Medline: 30300102]
38. Rudd MD, Joiner TE, Rajab MH. Treating Suicidal Behavior: An Effective, Time-Limited Approach. New York City, NY: Guilford Publications; 2001.
39. Asarnow JR, Berk M, Hughes JL, Anderson NL. The SAFETY program: a treatment-development trial of a cognitive-behavioral family treatment for adolescent suicide attempters. J Clin Child Adolesc Psychol 2015;44(1):194-203 [FREE Full text] [doi: 10.1080/15374416.2014.940624] [Medline: 25255931]
40. Bryan CJ, Mintz J, Clemans TA, Leeson B, Burch TS, Williams SR, et al. Effect of crisis response planning vs. contracts for safety on suicide risk in U.S. Army Soldiers: a randomized clinical trial. J Affect Disord 2017 Apr 01;212:64-72. [doi: 10.1016/j.jad.2017.01.028] [Medline: 28142085]
41. Little V, Neufeld J, Cole AR. Integrating safety plans for suicidal patients into patient portals: challenges and opportunities. Psychiatr Serv 2018 Jun 01;69(6):618-619. [doi: 10.1176/appi.ps.201700458] [Medline: 29493413]
42. Nuij C, van Ballegooijen W, Ruwaard J, de Beurs D, Mokkenstorm J, van Duijn E, et al. Smartphone-based safety planning and self-monitoring for suicidal patients: rationale and study protocol of the CASPAR (Continuous Assessment for Suicide Prevention And Research) study. Internet Interv 2018 May 05;13:16-23 [FREE Full text] [doi: 10.1016/j.invent.2018.04.005] [Medline: 30206514]
43. Kennard BD, Biernesser C, Wolfe KL, Foxwell AA, Craddock Lee SJ, Rial KV, et al. Developing a brief suicide prevention intervention and mobile phone application: a qualitative report. J Technol Hum Serv 2015 Oct 01;33(4):345-357 [FREE Full text] [doi: 10.1080/15228835.2015.1106384] [Medline: 26977137]
44. Moscardini EH, Hill RM, Dodd CG, Do C, Kaplow JB, Tucker RP. Suicide safety planning: clinician training, comfort, and safety plan utilization. Int J Environ Res Public Health 2020 Sep 04;17(18):6444 [FREE Full text] [doi: 10.3390/ijerph17186444] [Medline: 32899637]
45. Czyz EK, Arango A, Healy N, King CA, Walton M. Augmenting safety planning with text messaging support for adolescents at elevated suicide risk: development and acceptability study. JMIR Ment Health 2020 May 25;7(5):e17345 [FREE Full text] [doi: 10.2196/17345] [Medline: 32160150]
46. Anderson M, Jiang J. Teens, social media and technology 2018. Pew Research Center. 2018 May 31. URL: https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/ [accessed 2024-02-28]
47. Comer JS. Introduction to the special series: applying new technologies to extend the scope and accessibility of mental health care. Cogn Behav Pract 2015 Aug;22(3):253-257. [doi: 10.1016/j.cbpra.2015.04.002]
48. Lal S. E-mental health: promising advancements in policy, research, and practice. Healthc Manage Forum 2019 Mar;32(2):56-62 [FREE Full text] [doi: 10.1177/0840470418818583] [Medline: 30739487]
49. Turvey CL, Roberts LJ. Recent developments in the use of online resources and mobile technologies to support mental health care. Int Rev Psychiatry 2015;27(6):547-557. [doi: 10.3109/09540261.2015.1087975] [Medline: 26523397]
50. Shingleton RM, Richards LK, Thompson-Brenner H. Using technology within the treatment of eating disorders: a clinical practice review. Psychotherapy (Chic) 2013 Dec;50(4):576-582 [FREE Full text] [doi: 10.1037/a0031815] [Medline: 23527906]
51. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]

52. Ryan R, Horey D, Oliver S, McKenzie J, Prictor M, Santesso N, et al. Cochrane Consumers and Communication Group Standard protocol text and additional guidance for review authors. La Trobe University. 2019 Mar. URL: https://opal.latrobe.edu.au/articles/journal_contribution/Standard_Protocol_Text_and_Additional_Guidance_for_Review_Authors/6692969 [accessed 2023-05-19]
53. Andreasson K, Krogh J, Bech P, Frandsen H, Buus N, Stanley B, et al. MYPLAN -mobile phone application to manage crisis of persons at risk of suicide: study protocol for a randomized controlled trial. Trials 2017 Apr 11;18(1):171 [FREE Full text] [doi: 10.1186/s13063-017-1876-9] [Medline: 28399909]
54. Buus N, Erlangsen A, River J, Andreasson K, Frandsen H, Larsen JL, et al. Stakeholder perspectives on using and developing the MYPLAN suicide prevention mobile phone application: a focus group study. Arch Suicide Res 2020;24(1):48-63. [doi: 10.1080/13811118.2018.1489319] [Medline: 29985770]
55. Jeong YW, Chang HJ, Kim JA. Development and feasibility of a safety plan mobile application for adolescent suicide attempt survivors. Comput Inform Nurs 2020 Aug;38(8):382-392. [doi: 10.1097/CIN.0000000000000592] [Medline: 32079816]
56. Meier L, Gurtner C, Nuessli S, Miletic M, Bürkle T, Durrer M. SERO - a new mobile app for suicide prevention. Stud Health Technol Inform 2022 May 16;292:3-8. [doi: 10.3233/SHTI220310] [Medline: 35575841]
57. Melia R, Francis K, Duggan J, Bogue J, O'Sullivan M, Young K, et al. Using a safety planning mobile app to address suicidality in young people attending community mental health services in Ireland: protocol for a pilot randomized controlled trial. JMIR Res Protoc 2023 Feb 21;12:e44205 [FREE Full text] [doi: 10.2196/44205] [Medline: 36809171]
58. Melvin GA, Gresham D, Beaton S, Coles J, Tonge BJ, Gordon MS, et al. Evaluating the feasibility and effectiveness of an Australian safety planning smartphone application: a pilot study within a tertiary mental health service. Suicide Life Threat Behav 2019 Jun;49(3):846-858 [FREE Full text] [doi: 10.1111/sltb.12490] [Medline: 29999193]
59. Muscara F, Ng O, Crossley L, Lu S, Kalisch L, Melvin G, et al. The feasibility of using smartphone apps to manage self-harm and suicidal acts in adolescents admitted to an inpatient mental health ward. Digit Health 2020 Nov 26;6:2055207620975315 [FREE Full text] [doi: 10.1177/2055207620975315] [Medline: 33294207]
60. Nuij C, van Ballegooijen W, de Beurs D, de Winter RF, Gilissen R, O'Connor RC, et al. The feasibility of using smartphone apps as treatment components for depressed suicidal outpatients. Front Psychiatry 2022 Sep 27;13:971046 [FREE Full text] [doi: 10.3389/fpsyt.2022.971046] [Medline: 36238944]
61. O'Grady C, Melia R, Bogue J, O'Sullivan M, Young K, Duggan J. A mobile health approach for improving outcomes in suicide prevention (SafePlan). J Med Internet Res 2020 Jul 30;22(7):e17481 [FREE Full text] [doi: 10.2196/17481] [Medline: 32729845]
62. Skovgaard Larsen JL, Frandsen H, Erlangsen A. MYPLAN - a mobile phone application for supporting people at risk of suicide. Crisis 2016 May;37(3):236-240. [doi: 10.1027/0227-5910/a000371] [Medline: 26831213]
63. Pauwels K, Aerts S, Muijzers E, De Jaegere E, van Heeringen K, Portzky G. BackUp: development and evaluation of a smart-phone application for coping with suicidal crises. PLoS One 2017 Jun 21;12(6):e0178144 [FREE Full text] [doi: 10.1371/journal.pone.0178144] [Medline: 28636617]
64. Larkin C, Djamasbi S, Boudreaux ED, Varzgani F, Garner R, Siddique M, et al. ReachCare mobile apps for patients experiencing suicidality in the emergency department: development and usability testing using mixed methods. JMIR Form Res 2023 Jan 27;7:e41422 [FREE Full text] [doi: 10.2196/41422] [Medline: 36705961]
65. Tufanaru C, Munn Z, Aromataris E, Campbell J, Hopp L. Chapter 3: systematic reviews of effectiveness. In: Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis. Adelaide, South Australia: Joanna Briggs Institute; 2020.
66. Lockwood C, Munn Z, Porritt K. Qualitative research synthesis: methodological guidance for systematic reviewers utilizing meta-aggregation. Int J Evid Based Healthc 2015 Sep;13(3):179-187. [doi: 10.1097/XEB.0000000000000062] [Medline: 26262565]
67. De Jaegere E, Portzky G, van den Berg M. Ethical guidelines for technology-based suicide prevention programmes. European Regions Enforcing Actions against Suicide. 2014. URL: https://www.researchgate.net/publication/260297369_Ethical_Guidelines_for_Technology-Based_Suicide_Prevention_Programmes [accessed 2024-02-28]
68. Brooke J. SUS: a 'quick and dirty' usability scale. In: Usability Evaluation In Industry. Boca Raton, FL: CRC Press; 1996.
69. McClure JB, Anderson ML, Bradley K, An LC, Catz SL. Evaluating an adaptive and interactive mHealth smoking cessation and medication adherence program: a randomized pilot feasibility study. JMIR Mhealth Uhealth 2016 Aug 03;4(3):e94 [FREE Full text] [doi: 10.2196/mhealth.6002] [Medline: 27489247]
70. Stubbe DE. Enhancing adherence: using mobile health technology to improve self-management for individuals with schizophrenia. Focus (Am Psychiatr Publ) 2020 Oct;18(4):424-427 [FREE Full text] [doi: 10.1176/appi.focus.20200028] [Medline: 33343253]
71. Short CE, DeSmet A, Woods C, Williams SL, Maher C, Middelweerd A, et al. Measuring engagement in eHealth and mHealth behavior change interventions: viewpoint of methodologies. J Med Internet Res 2018 Nov 16;20(11):e292 [FREE Full text] [doi: 10.2196/jmir.9397] [Medline: 30446482]
72. Bell L, Garnett C, Qian T, Perski O, Williamson E, Potts HW. Engagement with a behavior change app for alcohol reduction: data visualization for longitudinal observational study. J Med Internet Res 2020 Dec 11;22(12):e23369 [FREE Full text] [doi: 10.2196/23369] [Medline: 33306026]

73. Hatchel T, Merrin GJ, Espelage AD. Peer victimization and suicidality among LGBTQ youth: the roles of school belonging, self-compassion, and parental support. J LGBT Youth 2019;16(2):134-156. [doi: 10.1080/19361653.2018.1543036]
74. Flouri E, Buchanan A. The protective role of parental involvement in adolescent suicide. Crisis 2002;23(1):17-22. [doi: 10.1027//0227-5910.23.1.17] [Medline: 12650218]
75. Abbott-Smith S, Ring N, Dougall N, Davey J. Suicide prevention: what does the evidence show for the effectiveness of safety planning for children and young people? - a systematic scoping review. J Psychiatr Ment Health Nurs 2023 Oct;30(5):899-910. [doi: 10.1111/jpm.12928] [Medline: 37052321]
76. Asarnow JR, Zullo L, Ernestus SM, Venables CW, Goldston DB, Tunno AM, et al. "Lock and protect": development of a digital decision aid to support lethal means counseling in parents of suicidal youth. Front Psychiatry 2021 Oct 6;12:736236 [FREE Full text] [doi: 10.3389/fpsyt.2021.736236] [Medline: 34690841]
77. Auerbach RP, Lan R, Galfalvy H, Alqueza KL, Cohn JF, Crowley RN, et al. Intensive longitudinal assessment of adolescents to predict suicidal thoughts and behaviors. J Am Acad Child Adolesc Psychiatry 2023 Sep;62(9):1010-1020. [doi: 10.1016/j.jaac.2023.03.018] [Medline: 37182586]
78. Bhatt P, Liu J, Gong Y, Wang J, Guo Y. Emerging artificial intelligence-empowered mHealth: scoping review. JMIR Mhealth Uhealth 2022 Jun 09;10(6):e35053 [FREE Full text] [doi: 10.2196/35053] [Medline: 35679107]

Abbreviations

JBI: Joanna Briggs Institute
mHealth: mobile health
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
RQ: research question
SPC: safety planning component
SPI: safety planning intervention

Edited by J Torous; submitted 14.09.23; peer-reviewed by HL Tam, B Leckning; comments to author 06.10.23; revised version received 19.12.23; accepted 31.12.23; published 28.03.24.

Please cite as:
Gryglewicz K, Orr VL, McNeil MJ, Taliaferro LA, Hines S, Duffy TL, Wisniewski PJ
Translating Suicide Safety Planning Components Into the Design of mHealth App Features: Systematic Review
JMIR Ment Health 2024;11:e52763
URL: https://mental.jmir.org/2024/1/e52763
doi:10.2196/52763
PMID:38546711

Original Paper

Feasibility, Acceptability, and Preliminary Efficacy of a Smartphone App–Led Cognitive Behavioral Therapy for Depression Under Therapist Supervision: Open Trial

Sabine Wilhelm[1], PhD; Emily E Bernstein[1], PhD; Kate H Bentley[1], PhD; Ivar Snorrason[1], PhD; Susanne S Hoeppner[1], PhD; Dalton Klare[1], MA; Jennifer L Greenberg[1], PsyD; Hilary Weingarden[1], PhD; Thomas H McCoy[1], MD; Oliver Harrison[2], MD

[1]Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States

[2]Koa Health, London, United Kingdom

Corresponding Author:
Sabine Wilhelm, PhD
Massachusetts General Hospital, Harvard Medical School
185 Cambridge Street
Suite 2000
Boston, MA, 02114
United States
Phone: 1 617 724 6146
Email: swilhelm@mgh.harvard.edu

Abstract

Background: Major depressive disorder affects approximately 1 in 5 adults during their lifetime and is the leading cause of disability worldwide. Yet, a minority receive adequate treatment due to person-level (eg, geographical distance to providers) and systems-level (eg, shortage of trained providers) barriers. Digital tools could improve this treatment gap by reducing the time and frequency of therapy sessions needed for effective treatment through the provision of flexible, automated support.

Objective: This study aimed to examine the feasibility, acceptability, and preliminary clinical effect of Mindset for Depression, a deployment-ready 8-week smartphone-based cognitive behavioral therapy (CBT) supported by brief teletherapy appointments with a therapist.

Methods: This 8-week, single-arm open trial tested the Mindset for Depression app when combined with 8 brief (16-25 minutes) video conferencing visits with a licensed doctoral-level CBT therapist (n=28 participants). The app offers flexible, accessible psychoeducation, CBT skills practice, and support to patients as well as clinician guidance to promote sustained engagement, monitor safety, and tailor treatment to individual patient needs. To increase accessibility and thus generalizability, all study procedures were conducted remotely. Feasibility and acceptability were assessed via attrition, patient expectations and feedback, and treatment utilization. The primary clinical outcome measure was the clinician-rated Hamilton Depression Rating Scale, administered at pretreatment, midpoint, and posttreatment. Secondary measures of functional impairment and quality of life as well as maintenance of gains (3-month follow-up) were also collected.

Results: Treatment credibility (week 4), expectancy (week 4), and satisfaction (week 8) were moderate to high, and attrition was low (n=2, 7%). Participants self-reported using the app or practicing (either on or off the app) the CBT skills taught in the app for a median of 50 (IQR 30-60; week 4) or 60 (IQR 30-90; week 8) minutes per week; participants accessed the app on an average 36.8 (SD 10.0) days and completed a median of 7 of 8 (IQR 6-8) steps by the week 8 assessment. The app was rated positively across domains of engagement, functionality, aesthetics, and information. Participants' depression severity scores decreased from an average Hamilton Depression Rating Scale score indicating moderate depression (mean 19.1, SD 5.0) at baseline to a week 8 mean score indicating mild depression (mean 10.8, SD 6.1; d=1.47; P<.001). Improvement was also observed for functional impairment and quality of life. Gains were maintained at 3-month follow-up.

Conclusions: The results show that Mindset for Depression is a feasible and acceptable treatment option for individuals with major depressive disorder. This smartphone-led treatment holds promise to be an efficacious, scalable, and cost-effective treatment option. The next steps include testing Mindset for Depression in a fully powered randomized controlled trial and real-world clinical settings.

Trial Registration: ClinicalTrials.gov NCT05386329; https://clinicaltrials.gov/study/NCT05386329?term=NCT05386329

(JMIR Ment Health 2024;11:e53998) doi:10.2196/53998

KEYWORDS

depressive disorder; depressive; depression; open trial; open trials; single arm; smartphone; cognitive behavioral therapy; cognitive behavioural therapy; CBT; psychotherapy; psychoeducation; digital health; mobile applications; mHealth; mobile health; app; apps; application; applications; psychiatry; psychiatric; feasibility; acceptability; usability; satisfaction; user experience; mental

Introduction

Major depressive disorder (MDD), characterized by hallmark symptoms of persistent depressed mood and loss of interest in activities [1], is highly prevalent. In 2020, an estimated 21 million adults in the United States were impacted (8.4% population prevalence) [2]. The rates of elevated depressive symptoms have continued to rise since the COVID-19 pandemic, now affecting nearly 1 in 3 adults [3]. Depression is the leading cause of disability worldwide [4] and is associated with economic costs exceeding US $326.2 billion in the United States alone [5]. Despite the substantial personal and societal impact of MDD, a minority of individuals meeting diagnostic criteria—let alone those at risk or with subthreshold symptoms—receive care; even fewer receive minimally adequate treatment (estimates range from 3% in low- to middle-income countries and 23% in high-income countries [6]), such as cognitive behavioral therapy (CBT), the most widely studied and recommended psychotherapy [7]. At a systems level, the limited availability of trained clinicians is a substantial contributor to low treatment utilization [8]. There simply are not and will not be enough clinicians to meet current demands for mental health care. Additionally, many people do not seek treatment for depression due to obstacles such as geographic distance from care providers, high costs, and stigma [8,9]. Given the prevalence of MDD and the substantial treatment gaps that exist, there is a clear need for low barrier, more widely accessible, effective treatments for MDD.

The technology could bridge such gaps. Smartphone apps or other digital tools could reduce the time and frequency of sessions by supplementing clinician effort with automated, validated support that can be used flexibly between sessions [10,11]. However, standalone apps are not sufficient or appealing for many patients [12,13]. The majority of apps, even those that are grounded in empirically supported treatments, have high dropout rates, which limits their effectiveness [14-16]. The absence of concurrent human support is often cited as the major reason for nonadherence or nonengagement [13,17-19]. Some engagement is likely a minimum requirement for an app-based therapy to be effective; guidance from a trained provider should further mitigate issues of comprehension, personalization, problem-solving, and interference from comorbid or life concerns [20]. Equivalent effects of face-to-face CBT and internet-delivered CBT for depression have been found for treatments that are *therapist guided*, meaning patients are in contact with a therapist throughout treatment (eg, weekly sessions, check-in phone calls, asynchronous, messaging, and weekly feedback emails) [21-23]. Moreover, many users simply want access to a therapist and are less willing to engage in self-directed digital treatments [24,25]. Thus, a digital service that combines mobile-based CBT with brief remote individual sessions with a clinician (ie, teletherapy monitoring) has the potential to greatly enhance the scalability of high-quality app-based treatment, particularly for moderately and severely ill patients while reducing clinician burden and cost [26-28].

The purpose of this study was to conduct an open trial to test the feasibility, acceptability, and efficacy of the Mindset for Depression app (a novel, smartphone-based CBT program) with brief video-conferencing appointments with a therapist. We hypothesized that the treatment would be feasible and acceptable. We also hypothesized that treatment would yield statistically significant reductions in depression symptom severity (primary clinical outcome) as well as improvements in functioning and quality of life (secondary clinical outcomes) from baseline to posttreatment (week 8). The treatment was tested for patients with moderate to severe depression: those who would typically be referred for one-on-one outpatient therapy [29].

Methods

Study Design

This open trial tested the Mindset for Depression app when combined with brief (16-25 minutes) video-conferencing visits with a CBT therapist over 8 weeks. The primary outcomes were feasibility, acceptability, and preliminary efficacy, as measured by change in depression symptom severity. To increase accessibility and thus generalizability, all study procedures were conducted remotely.

Ethical Considerations

The study was approved by the institutional review board of Massachusetts General Hospital (2020P001958). All participants provided informed consent prior to the initiation of study procedures and were given the ability to opt out at any point. Data were deidentified to protect participants' privacy. Participants were compensated US $25 at mid-treatment, end of treatment, and 3-month follow-up assessments.

Participants

Eligible participants, recruited between May 2022 and February 2023, were at least 18 years old, living in Massachusetts, presenting with a current primary *Diagnostic and Statistical Manual of Mental Disorders: 5th Edition* (*DSM-5*) diagnosis of MDD, and experiencing at least moderately severe symptoms (Patient Health Questionnaire-9 [PHQ-9] score ≥ 10). Participants taking psychotropic medication were on a stable dose for at least 2 months prior to enrollment and were asked to remain on the same stable dose throughout the study period.

Exclusion criteria included 4 or more prior sessions of CBT for depression (assessed via self-report and interview with an independent evaluator), current severe substance use disorder, lifetime bipolar disorder or psychosis, acute and active suicidal ideation as indicated by clinical judgment, a score ≥ 2 on the past month suicidal ideation subscale of the Columbia-Suicide Severity Rating Scale [30], concurrent psychological treatment, and inability to engage with treatment (eg, did not own a supported smartphone).

Procedure

Treatment

The Mindset for Depression app provides key CBT-derived content for adults with MDD and was designed to be used in conjunction with a therapist over 8 weeks. The duration of CBT trials typically ranges from 6 to 20 sessions [31]. Mindset for Depression was built in collaboration between researchers at the Massachusetts General Hospital and Koa Health. The app-based format allows participants to review CBT content and accompanying skills practice exercises at their convenience and own pace and with support from their therapist.

The app and clinician dashboard were developed through collaborative, user-centered design, integrating perspectives from clinicians (MDs and psychologists with expertise in MDD and CBT), digital health researchers, patients with MDD and experience with CBT and other therapies, engineers, and designers. Through this approach, the product being tested was deployment ready (eg, built on a commercial platform, able to be quickly scaled and professionally maintained to minimize technical difficulties, and ensure compliance with up-to-date privacy and security standards) and therefore well positioned to succeed outside of research studies [32].

CBT Modules

The app delivers content in 8 steps, corresponding to the 8 weeks of treatment. A summary of these steps is visualized in Table 1. Participants were also allowed access to the app during the 3-month follow-up period. Core CBT skills included across treatment include psychoeducation, cognitive restructuring and core beliefs, behavioral activation, mindfulness, and relapse prevention [33,34]. Step 1 comprises psychoeducation about MDD and the CBT model and background and skills practice for identifying and restructuring "thinking traps," or maladaptive automatic thoughts [35,36]. Step 2 focuses on the short- and long-term impact of withdrawal and avoidance on mood and provides a structure for recording daily activities and monitoring associated moods. Step 3 introduces behavioral activation and scheduling and provides guidance for identifying valued or new activities and setting specific, measurable, achievable, relevant, and time-bound (SMART) goals [37,38]. Activity scheduling and monitoring (ie, recording completed activities and associated mood ratings), with an emphasis on personal values and meaning, continue for the remainder of treatment. Step 4 introduces mindfulness (present-focused and nonjudgmental awareness) and offers a guided mindful breathing audio exercise [39]. Steps 5 and 6 provide users with additional mindfulness approaches, including grounding and letting go of unhelpful thoughts. Step 7 builds on prior cognitive skills and delves into the definition of core beliefs, their relationship to automatic thoughts and feelings, and strategies to identify and challenge them (eg, downward arrow technique and building self-esteem). Step 8 concludes with relapse prevention by helping users consolidate treatment skills, anticipate future challenges, and plan for continued practice and flexible use of skills. Example screenshots from the smartphone app are included in Figure 1.

Figure 1. Screenshots from the Mindset for Depression smartphone app. CBT: cognitive behavioral therapy; SMART: specific, measurable, achievable, relevant, and time-bound.

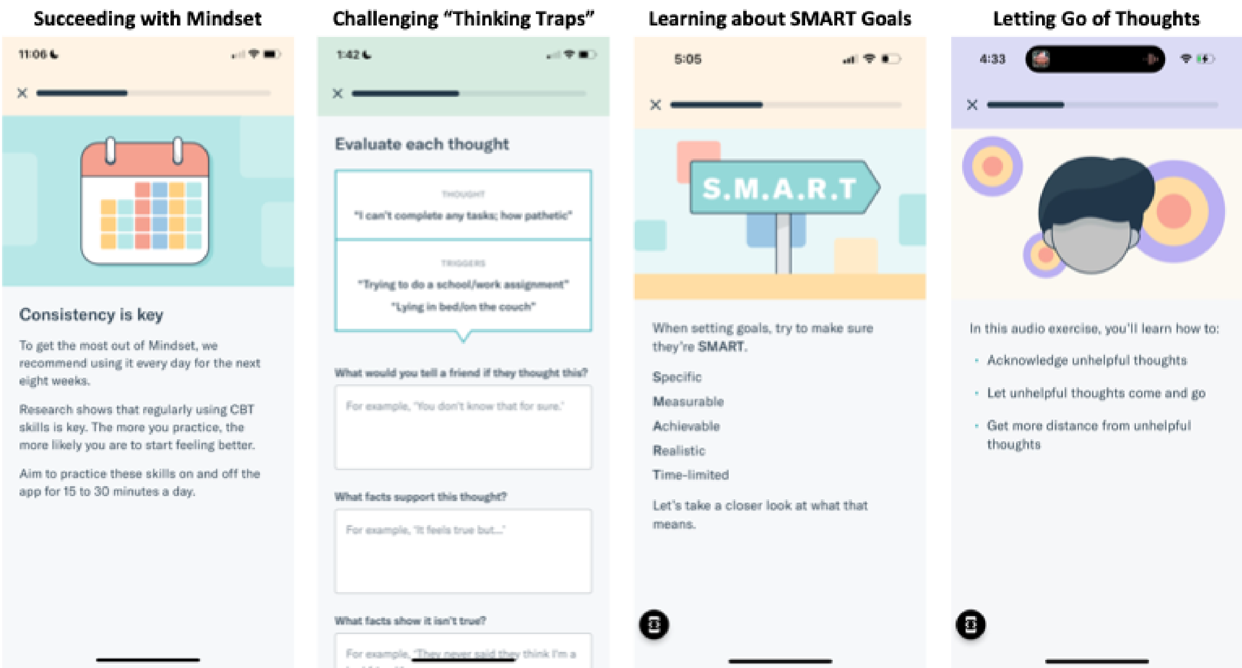


Table 1. Summary of steps in the Mindset for Depression program.

	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7	Step 8
Psychoeducation	✓							
Cognitive restructuring	✓							
Relationship between behavior and mood		✓						
Behavioral activation			✓	✓	✓	✓	✓	✓
Mindfulness				✓	✓	✓		
Modifying core beliefs and building self-esteem							✓	
Relapse prevention								✓

Therapists

Each participant was matched with a licensed doctoral-level therapist. Therapists were trained in and actively practicing CBT for MDD and provided with study-specific training in using the Mindset app and therapist dashboard prior to beginning the trial. To ensure proficiency, therapists were required to complete the Massachusetts General Hospital Psychiatry Academy CBT training course and pass (>90% correct) both the corresponding CBT knowledge test and an MDD knowledge test. Weekly supervision from the principal investigator (expert in CBT) was also provided. To ensure ongoing high-quality treatment, including that implementation fidelity targets were met and non-CBT techniques were absent, sessions were audio recorded, and 40 of the 224 planned sessions (18.9% of the 212 sessions ultimately conducted) were randomly selected and rated for competency and treatment adherence by an independent rater. Adherence raters were experienced in CBT for MDD and further trained and supervised. Core elements of each treatment session (5-6 items) were rated for adherence on a 7-point scale (1=not at all to 7=completely adherent) and then a global rating of adherence was assigned. The full adherence scale is included in Multimedia Appendix 1. Competence was rated on 12 aspects (32 items) of CBT for MDD (eg, positive outlook, knowledge, clear communication, empathy, flexibility, and empowering the patient). Each item was scored on a 5-point scale (1=not at all to 5=completely competent) and then a global rating of competence was assigned. Overall, adherence and competence were high, with 100% of all rated sessions evaluated as “completely” adherent and 100% of all rated sessions evaluated as “mostly” or “completely” competent.

Therapists offered each patient 8 video-conferencing appointments (16-25 minutes; via HIPAA [Health Insurance Portability and Accountability Act]–compliant video conference) to be conducted weekly. This duration of appointment corresponds to a clinician billing code (CPT-90832), helping to ensure that the reimbursement of clinician time would not become a barrier to scale-up following the research. As needed, because of a therapist’s or patient’s schedule, up to 2 sessions were able to be scheduled per week. Throughout the treatment, participants were able to communicate with their therapists between sessions through asynchronous in-app secure messaging. Sessions were meant to support a patient’s progress through the app-led treatment. In this way, the model mimicked the “flipped classroom,” a new pedagogical approach shown to improve student learning [40]. In a flipped classroom, students watch or read lectures and complete initial practice problems asynchronously, reserving valuable classroom time for active problem-solving with an instructor. As such, sessions were intended to monitor risk as needed, help participants set goals, enhance motivation, clarify and practice the skills learned via the Mindset app to best meet the patient’s needs, brainstorm ideas for homework, and problem-solve treatment barriers that arose. Therapists were instructed to work within a CBT framework and not to introduce other treatment modalities. Such fidelity was monitored in weekly supervision, via therapist self-checks included within session records (“Did you use any of the following non-CBT techniques? [check *all* that apply]”) and via adherence ratings (ie, the degree to which forbidden content was introduced). The therapist dashboard was a separate web-based portal wherein therapists could receive and respond to messages and track participant progress in the app.

Assessments

Assessments were conducted by master’s or doctoral-level independent evaluators who were not involved in treatment, were complemented by participant self-report, and occurred at baseline, mid-treatment (week 4), end of treatment (week 8), and follow-up (3 months posttreatment). Evaluators completed training on all clinician-administered measures and were required to maintain high reliability (>0.75 intraclass correlation coefficient), with a gold standard expert rater; 18.9% (20/106) of randomly selected assessments were rated to prevent rater drift. Evaluators were not privy to participants’ progress in treatment (eg, app content reviewed and session notes). Adverse events, life events, and changes in medication or outside treatment were surveyed at each assessment or when a patient reported to study staff.

Measures Descriptions

Baseline Diagnostic Assessment

The Mini International Neuropsychiatric Interview was used to establish eligibility and characterize the sample. It is a reliable, validated semistructured diagnostic assessment of *DSM-5* psychiatric disorders [41].

Feasibility and Acceptability

Participants completed the self-reported measures as follows. The Credibility/Expectancy Questionnaire (CEQ) [42], completed at baseline and week 4, is a 6-item, self-reported

Likert-type questionnaire that assesses patients' judgments about the credibility of the treatment rationale and treatment expectancy. Items on both subscales are summed together for total outcome scores that can range from 3 to 27, where higher scores mean higher treatment credibility and higher outcome expectancy. We assessed the internal consistency of scales with coefficient omega (McDonald ω), given the heterogeneity of variances across scale items; coefficient ω can be interpreted in the same way as Cronbach α. The internal consistency of the credibility items in this sample ranged from ω=0.69 at baseline to ω=0.77 at week 4; for the expectancy items, internal consistency ranged from ω=0.82 at baseline to ω=0.96 at week 4. The Mobile Application Rating Scale User Version (uMARS) [43], administered at week 8, collects evaluations of mobile health apps. The 26 items assess participants' evaluations of engagement, functionality, aesthetics, information quality, app subjectivity quality, and perceived impact. Items are rated on differently worded 5-point Likert scales ranging from 1 (inadequate) to 5 (excellent). An overall app rating score can be calculated as the mean score of the first 4 subscales (engagement, functionality, aesthetics, and information quality; a range of 1-5), where higher scores indicate higher overall perceived app quality. In this sample, the internal consistency of the 4 subscales used in the overall mean scores were ω=0.76 for engagement, ω=0.83 for functionality, ω=0.72 for aesthetics, and ω=0.62 for information quality, with an overall item consistency of ω=0.83. The Client Satisfaction Questionnaire (CSQ) [44], completed at weeks 4 and 8, is an 8-item questionnaire assessing satisfaction with clinical services received. Each item uses a 4-point Likert scale. Items are summed for a total score ranging from 8 to 32, with higher scores indicating greater satisfaction. The internal consistency of the CSQ was ω=0.93 at week 4 and ω=0.89 at week 8. Treatment use was assessed with a single question: "On average, how much time (in minutes) do you spend using the app or practicing skills from the app in total, per week?" Answers were collected as the number of minutes in integer format, where more time spent on and off the app was interpreted as greater treatment use. In addition, app use data were collected automatically based on the actions participants completed in the app. Due to technical issues, 2 participants' app use data were inadvertently not recorded. The internal consistency values for the CEQ credibility subscale at baseline and the uMARS information quality subscale fell below 0.7 and are a noted limitation.

Clinician-Administered Measures

The primary measure of MDD symptom severity was the (clinician-rated) Hamilton Depression Rating Scale (HAM-D) [45]. Considered a gold standard means of assessing symptom severity in patients who are depressed, it contains 21 items that are rated on a mixture of 3- and 5-point Likert scales. The first 17 items are summed for the total score, which can range from 0 to 52. Higher scores indicate greater depression severity. The internal consistency of the HAM-D in this sample ranged from ω=0.79 at baseline to ω=0.93 at the 3-month follow-up (question 17 was necessarily omitted from internal consistency calculations due to the absence of variability in responses; all participants received a score of 0 for this "insight" item ["Acknowledges being depressed and ill"] at all assessment points with the exception of 1 participant at the 3-month follow-up). To evaluate treatment response and remission, we used criteria of HAM-D score reductions of ≥50% for treatment response, HAM-D score reductions of ≥25% but < 50% for partial response, and HAM-D scores ≤7 to indicate remission [46-48]. An expert rater reviewed 18.9% (20 of the 106 assessments that were ultimately completed) of HAM-D assessments. Inter-rater reliability was excellent (HAM-D: intraclass correlation coefficient (1,1)=0.91).

Self-Reported Measures

Participants completed the following secondary measures of symptoms and functioning at each assessment: (1) The Work and Social Adjustment Scale (WSAS) [49] is a 5-item, self-reported measure of impairment in occupational, social, and family domains. Items are measured on 9-point Likert scales ranging from 0 (no impairment at all) to 8 (very severe impairment). The items are summed for a total score ranging from 0 to 40, where higher scores mean higher functional impairment. The internal consistency of the WSAS ranged from ω=0.84 at baseline to ω=0.93 at the 3-month follow-up. (2) The Quality of Life Enjoyment and Satisfaction Questionnaire-Short Form (Q-LES-Q-SF) [50] is a 16-item self-reported measure of subjective quality of life. Each question is rated on a 5-point Likert scale ranging from 1 (very poor) to 5 (very good). Questions 1-14 are then summed to a total score, and the total score is reported as a percentage maximum possible, such that the final percent score range is 0% to 100%; higher scores correspond to greater ratings of quality of life. The internal consistency of the Q-LES-Q-SF ranged from ω=0.84 at baseline to ω=0.90 at the 3-month follow-up. (3) The PHQ-9 [51] is a self-reported measure of the past week's depression severity. It includes 9 Likert scale items mapping onto *DSM-5* symptom criteria and ranging from 0 (not at all) to 3 (every day). The internal consistency of the PHQ-9 ranged from ω=0.70 at baseline to ω=0.89 at week 8.

Data Analysis

Power Analysis

With 28 participants enrolled, we had > 80% power to detect pre- to posttreatment effect sizes of d ≥1.37 (very large effect sizes), assuming 30% dropout, a pre- to posttreatment correlation of 0.18, and a doubling of the SD from pre- to posttreatment. The pre- to posttreatment correlation estimate was based on the mean pooled correlation between pretest and posttest HAM-D scores in 14 CBT trials for adult depression [52], and the estimate of the detec effect size was based on a single degree of freedom contrast in a paired means test implemented in SAS for Windows (version 9.4; SAS Institute).

Feasibility and Acceptability

We examined feasibility and acceptability by reporting (1) dropout rates and reasons (defined as participants not completing an end point HAM-D), (2) patient satisfaction (CSQ), (3) patient feedback (uMARS), (4) patient credibility and expectancy ratings (CEQ), and (5) treatment use. We computed means and SDs for the number of app steps completed, the number of days on which participants completed any actions in the app, the number of messages participants sent their therapist in the app, the number of sessions participants completed with their therapists, and session time spent per patient per week by therapists. For measures collected at least twice (ie, CSQ, CEQ credibility and expectancy, and treatment use), we used generalized linear mixed models with repeated measures to examine if these self-reported ratings changed over time.

Preliminary Efficacy and Secondary Outcomes

Analyses were first completed using our intent-to-treat sample (participants who completed a baseline assessment) and then repeated with our "per-protocol" sample (participants who completed posttreatment assessments and did not change psychiatric medications or begin psychotherapy during the study; n=24, 86%). We examined the preliminary efficacy of Mindset for Depression plus brief video-conferencing appointments with a therapist on symptoms and well-being outcomes using mixed model analyses with repeated measures (baseline, mid-treatment, posttreatment, and 3-month follow-up) modeled using an unstructured covariance matrix. We then compared pre- to posttreatment differences using a 2-tailed α of .05 to evaluate preliminary efficacy. We similarly compared pretreatment to end of follow-up estimates to estimate whether changes remained significant by the end of the follow-up. Means are presented as raw means with SDs, while differences between assessments are presented as model-estimated means with CIs (LSM differences [95% CI]) unless otherwise specified. Effect sizes were calculated as Hedges g_{ave}, which takes the correlation of within-participant scores into account [53]. Analyses were conducted for changes in depression symptoms (HAM-D and PHQ-9), functional impairment (WSAS), and quality of life (Q-LES-Q-SF). All analyses were completed using the SAS software (version 9.4, SAS Institute Inc).

Results

Overview

Study participants (N=28) were predominantly female (n=21, 75%), White (n=20, 71%), and single (n=19, 70%), with a mean age of 33.5 (SD 10.9) years. The majority of participants had college or advanced degrees, were employed full time, and came from urban or suburban locations (Tables 2 and 3). Nearly half of the participants (n=13, 45%) had one or more comorbid psychiatric diagnoses, and the average duration of MDD was 15.5 (SD 12.7) years.

Table 2. Baseline demographics of participants enrolled in the Mindset open trial.

Demographics	Values
Age (years), mean (SD)	33.5 (10.9)
Sex at birth, n (%)	
Female	75 (21)
Male	25 (7)
Gender identity, n (%)	
Women	75 (21)
Men	25 (7)
Sexual orientation, n (%)	
Straight or heterosexual	71 (20)
Bisexual	18 (5)
Lesbian, gay, or homosexual	4 (1)
Other	7 (2)
Hispanic ethnicity, n (%)	18 (5)
Race, n (%)	
Asian or Pacific Islander	11 (3)
Black	4 (1)
White	71 (20)
Other	14 (4)
Education, n (%)	
Less than or equal to a high school graduate	14 (4)
Technical school or some college	18 (5)
College graduate	39 (11)
Graduate or professional school	29 (8)
Marital status, n (%)	
Single, never married	68 (19)
Married	18 (5)
Partnered	7 (2)
Separated or widowed	7 (2)
Employment, n (%)	
Full time (≥35 hours per week)	86 (24)
Student	7 (2)
Unemployed	4 (1)
Retired	4 (1)
Household income (US $), n (%)	
$34,999 or less	11 (3)
$35,000-74,999	25 (7)
$75,000-149,999	50 (14)
$150,000 or more	14 (4)
Geographic location, n (%)	
Urban	43 (12)
Suburban	46 (13)
Rural	11 (3)

Table 3. Baseline clinical characteristics of participants enrolled in the Mindset open trial.

Characteristics	Values
Duration of MDD[a] (years), mean (SD)	15.5 (12.7)
Current psychiatric comorbidities (*DSM-5*[b] diagnoses)[c], n (%)	
Agoraphobia	7 (2)
Alcohol use disorder	11 (3)
Generalized anxiety disorder	21 (6)
Social anxiety disorder	14 (4)
Other	18 (5)
Number of psychiatric comorbidities, n (%)	
None	54 (15)
1	36 (10)
2	0 (0)
3 or more	11 (3)
Current psychotropic medication[c], n (%)	
None	57 (16)
SRI[d]	21 (6)
Non-SRI antidepressant	18 (5)
Other psychotropic medication[e]	14 (4)

[a]MDD: major depressive disorder.
bDSM-5: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition.
[c]Percentage sums may exceed 100% because participants could report more than one diagnosis or be on more than 1 s psychotropic medication.
[d]SRI: serotonin reuptake inhibitor.
[e]This included anticonvulsants; no participant reported taking antipsychotics.

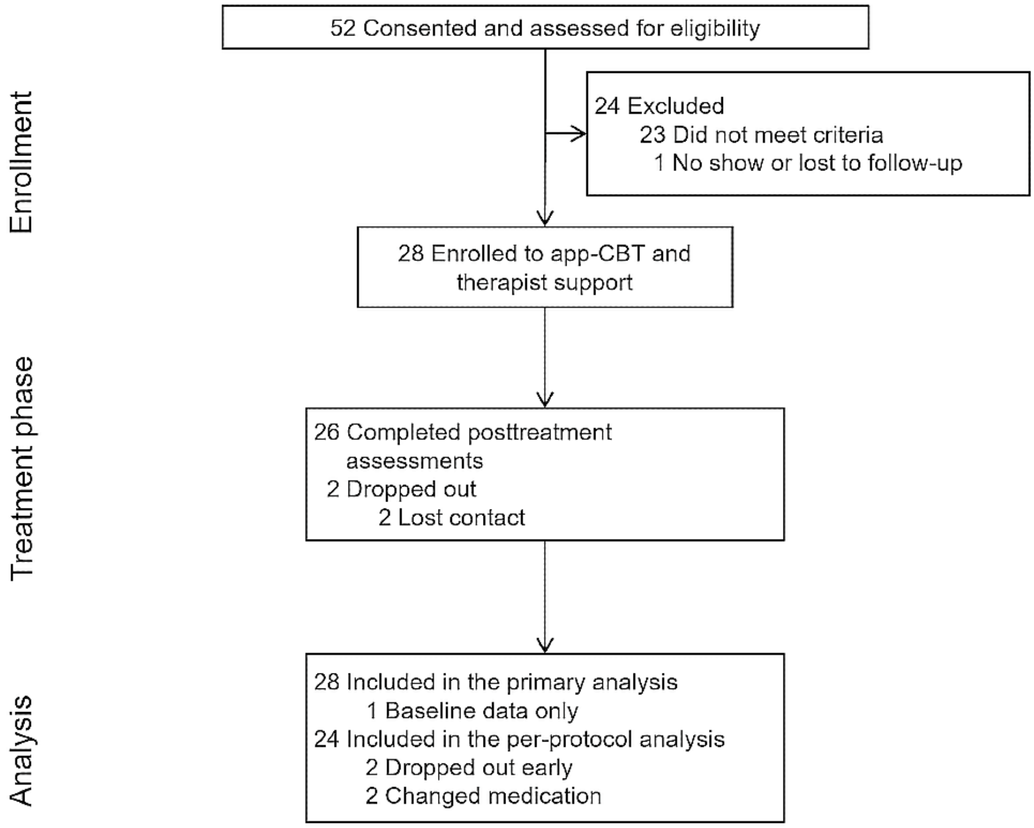
Feasibility and Acceptability

Credibility and expectancy scores were moderate to high at pre- and mid-treatment. The ratings did not differ significantly between timepoints. Mean credibility ratings were 18.9 (SD 3.1) at pretreatment and 19.3 (SD 3.8) at mid-treatment (LSM difference 0.4, 95% CI –1.4 to 2.2; *P*=.63; g_{ave}=0.13). Mean expectancy ratings were 13.8 (SD 3.3) at pretreatment and 15.0 (SD 5.4) at mid-treatment (LSM difference 1.2, 95% CI –0.7-3.0; *P*=.22; g_{ave}=0.28). Of the 28 participants, 2 (7%) dropped out of the study prior to their posttreatment assessment at week 8 (Figure 2): 1 during the first week of treatment and 1 after the midpoint assessment; both participants were lost to follow-up despite repeated contact attempts and neither provided a reason for drop out. One more participant was lost to follow-up after the posttreatment assessment. Among the 26 participants who completed the posttreatment assessment, patient satisfaction was high and did not change significantly from mid-treatment (CSQ total score mean 26.3, SD 4.0) to posttreatment (mean 27.2, SD 3.3; LSM difference 1.0, 95% CI –0.1 to 2.2; *P*=.07; g_{ave}=0.24). Conservatively counting the 2 dropouts as not satisfied, 89% (16/28) were very or mostly (9/28) satisfied and 93% (26/28) would recommend the Mindset for Depression program.

With respect to app use and satisfaction, participants reported practicing skills from the app on their smartphone and offline for a median of 50 (IQR 30-60) minutes per week up to mid-treatment and 60 (IQR 30-90) minutes per week between mid- and posttreatment. Based on passively collected app use data (n=26), participants accessed the app on 36.8 (SD 10.0) days, completed a median of 7 (IQR 6-8) steps out of 8 steps by the week 8 assessment, and sent a median of 0 (IQR 0-4) between session messages to their therapist through the app. Five participants completed the last assigned step after the week 8 assessment, bringing step completion to a median of 8 (IQR 6-8) by the end of follow-up (3 months), with 58% (15/26) participants completing the final step by then. Participants' overall ratings of the app quality, rated on the 1 (inadequate) to 5 (excellent) scale of the uMARS, was high (mean 4.3, SD 0.4); ratings of the app's functionality (mean 4.5, SD 0.6), aesthetics (mean 4.6, SD 0.4), and information (mean 4.6, SD 0.4) were higher than those of the engagement subscale (mean 3.6, SD 0.6). Participants reported a mean overall star rating of 4.0 (SD 0.5) but were less inclined to endorse that they would be willing to pay for the app (mean 2.5, SD 1.2).

With respect to therapist support, participants attended an average of 7.6 (SD 1.5) of the possible 8 brief sessions, each of which lasted approximately 24.5 (SD 1.1) minutes. In the sessions, therapists mainly covered behavioral strategies (mean 10.9, SD 3.4 minutes), cognitive strategies (mean 6.9, SD 2.7 minutes), psychoeducation (mean 2.5, SD 2.9 minutes), and mindfulness strategies (mean 2.4, SD 1.2 minutes), with only a little time (<2 minutes on average) spent on explicit motivational strategies, risk management, and technical issues. Participants had a mean homework completion rate (per therapist report) of 82.7% (SD 13.8%).

Figure 2. Flow of participants through the 8-week open trial of the Mindset Depression app with brief therapist visits on the web for people with a primary diagnosis of major depressive disorder. Reasons for ineligibility include diagnosis of bipolar disorder or severe substance use disorder, PHQ-9 score < 10, past CBT for MDD, acute, active suicidal ideation, and MDD not being the primary diagnosis. CBT: cognitive behavioral therapy.



Preliminary Efficacy and Secondary Outcomes

Over the course of the 8-week treatment, participants' depression severity decreased significantly on both the clinician-rated (HAM-D: P<.001; g_{ave}=1.47) and self-reported measures (PHQ-9: P<.001; g_{ave}=1.89; 4). Concurrently, participants' self-rated functional impairment decreased (WSAS: P<.001; g_{ave}=1.29), and their self-rated quality of life increased (Q-LES-Q-SF: P<.001; g_{ave}=1.74). These changes persisted through the 3-month follow-up, with effect sizes remaining largely the same (Table 4). The results did not differ meaningfully in the per-protocol analyses; HAM-D, PHQ-9, WSAS, and Q-LES-Q scores all improved with statistically significant and large effect sizes (Table 5).

Conservatively counting dropouts as having not responded to treatment, 46% (13/28) participants responded and another 7% (2/28) partially responded at posttreatment. Regarding remission, and again counting dropouts also as not remitting, 36% (10/28) achieved remission at posttreatment. Based on the per-protocol sample, 54% (13/24) participants fully responded to treatment and 8% (2/24) partially responded to treatment; 42% (10/24) achieved remission by the posttreatment assessment. By the end of follow-up and counting dropouts as not responding and not remitting, 50% (14/28) participants had responded to treatment, and an additional 21% (6/28) of participants had partially responded to treatment; 36% (10/28) participants were in remission. In the per-protocol sample at the end of follow-up and excluding the additional participant lost to follow-up, 61% (14/23) of participants had responded to treatment, and 26% (6/23) of participants had partially responded to treatment; 43% (10/23) participants were in remission at the 3-month follow-up assessment.

Table 4. Baseline, mid-treatment (week 4), end of treatment (week 8), and follow-up (week 20) estimated mean scores on key clinical outcome measures.

Outcome measure	Baseline, LSM[a] (SE)	Week 8, LSM (SE)	Week 20, LSM (SE)	Estimated difference (week 0-8), LSM (95% CI)	P value	Effect size (week 0-8)[b], Hedges g_{av}	Estimated difference (week 0-20), LSM (95%CI)	P value	Effect size (week 0-20)[a], Hedges g_{av}
HAM-D[c] total scores	19.1 (0.9)	11.3 (1.2)	10.4 (1.6)	–7.8 (–10.5 to –5.2)	<.001	1.47	–8.7 (–12.0 to –5.4)	<.001	1.44
PHQ-9-[d] total scores	15.1 (0.7)	7.1 (1.0)	6.9 (0.9)	–8.0 (–9.9 to –6.1)	<.001	1.89	–8.2 (–10.2 to –6.3)	<.001	1.97
WSAS[e] total scores	23.1 (1.5)	13.2 (1.5)	12.5 (1.8)	–9.9 (–13.5 to –6.4)	<.001	1.29	–10.6 (–14.5 to –6.8)	<.001	1.26
Q-LES-Q-SF[f] % scores	40.9 (2.2)	62.5 (2.6)	62.8 (2.8)	21.5 (14.5 to 28.6)	<.001	1.74	21.9 (15.0 to 28.8)	<.001	1.73

[a]LSM: least squares mean.

[b]Within-group effect sizes were calculated as Hedges g_{ave} for differences from baseline to week 8 or 20, respectively, using raw means data.

[c]HAM-D: Hamilton Depression Rating Scale (score range 0 to 52, where higher scores indicate greater depression severity).

[d]PHQ-9: Patient Health Questionnaire-9 item (score range 0 to 27, where higher scores indicate greater depression severity).

[e]WSAS: Work and Social Adjustment Scale (score range 0 to 40, where higher scores mean higher functional impairment).

[f]Q-LES-Q-SF: Quality of Life Enjoyment and Satisfaction Questionnaire- Short Form (percent score range 0% to 100%, where higher scores correspond to greater ratings of quality of life).

Table 5. Baseline, mid-treatment (week 4), end-of-treatment (week 8), and follow-up (week 20) estimated mean scores on key clinical outcome measures in the per-protocol sample (n=24).

Outcome measure	Baseline, LSM[a] (SE)	Week 8, LSM (SE)	Week 20, LSM (SE)	Estimated difference (week 0-8), LSM (95% CI)	P value	Effect size (week 0-8)[b], Hedges g_{av}	Estimated difference (week 0-20), LSM (95% CI)	P value	Effect size (week 0-20)[a], Hedges g_{av}
HAM-D[c] total scores	18.7 (1.0)	10.3 (1.2)	8.6 (1.3)	–8.4 (–11.1 to –5.6)	<.001	1.52	–10.1 (–13.1 to –7.1)	<.001	1.79
PHQ-9[d] total scores	14.7 (0.8)	6.4 (1.0)	6.1 (0.9)	–8.3 (–10.3 to –6.2)	<.001	1.89	–8.6 (–10.6 to –6.6)	<.001	2.03
WSAS[e] total scores	22.0 (1.6)	12.1 (1.4)	10.7 (1.6)	–9.8 (–13.7 to –6.0)	<.001	1.32	–11.2 (–15.2 to –7.3)	<.001	1.42
Q-LES-Q-SF[f] % scores	41.7 (2.3)	63.7 (2.8)	65.3 (2.7)	22.0 (14.5 to 29.6)	<.001	1.73	23.6 (16.7 to 30.5)	<.001	1.92

[a]LSM = least squares mean.

[b]Within-group effect sizes were calculated as Hedges g_{ave} for differences from baseline to week 8 or 20, respectively, using raw means for completers only.

[c]HAM-D: Hamilton Depression Rating Scale (score range 0 to 52, where higher scores indicate greater depression severity).

[d]PHQ-9: Patient Health Questionnaire-9 item (score range 0 to 27, where higher scores indicate greater depression severity).

[e]WSAS: Work and Social Adjustment Scale (score range 0 to 40, where higher scores mean higher functional impairment).

[f]Q-LES-Q-SF: Quality of Life Enjoyment and Satisfaction Questionnaire-Short Form (percent score range 0% to 100%, where higher scores correspond to greater ratings of quality of life).

Adverse Events and Medication Changes

Overall, 17 out of 28 participants reported a total of 33 adverse events during the 8-week treatment phase of the trial in categories such as psychiatric symptoms (n=21, 64%; eg, increased suicidal ideation, depression, or anxiety, sleep difficulties, and emotional distress), infections (n=6, 18%; eg, COVID-19, shingles, and illness), physical injuries (n=3, 9%), and other (n=3, 9%). All adverse events were identified as either mild (new event that did not interfere with activities of daily living; 25/33, 75.8%) or moderate (new event that posed some interference or required intervention to prevent interference; 8/33, 24.2%). No serious adverse events occurred in this trial. By assessing how likely it was that reported adverse events were related to treatment, most events were found to be definitely unrelated (19/33, 57.6%), followed by unlikely to be related (5/33, 15.2%) or possibly related (9/33, 27.3%). A waxing and waning course of MDD symptoms and suicidal ideation is common in MDD. However, there were no adverse events indicating significant clinical deterioration in the trial and no principal investigator–initiated withdrawals. Two participants changed psychotropic medications during the treatment phase: 1 participant increased the dosage of their medication for anxiety and 1 participant discontinued their medication for depression. There were no reported therapy changes during the treatment phase. In the 3-month follow-up phase of the trial, 7 participants reported an additional 10 adverse events in the categories of psychiatric symptoms (4/10, 40%), general disorders (2/10, 20%), and other (4/10, 40%), which were found to be definitely unrelated (5/10, 50%), unlikely to be related (1/10, 10%), and possibly related (4/10, 40%) to treatment. Also during the follow-up period, 5 participants changed psychotropic medications and 3 participants started individual therapy or counseling (non-CBT; for mood or anxiety, and traumatic event or PTSD).

Discussion

Principal Findings

In this study, we examined the feasibility, acceptability, and preliminary clinical impact of Mindset for Depression, an 8-week app-based CBT with therapist support. The results support Mindset for Depression as a viable treatment option for individuals with moderate to severe MDD. Treatment was feasible to deliver in a setting and acceptable to patients who varied widely in age, severity of the symptoms, and other clinical and demographic dimensions, as indicated by high retention rates (27/29, 93%), favorable satisfaction ratings (CSQ), and positive user feedback (uMARS). The results also showed that the treatment was efficacious. There was a significant reduction in clinician-rated (HAM-D) depression severity with a large effect size as well as significant improvement in functioning and quality of life. After just 8 weeks, about half of the participants were rated as treatment responders and a third were in remission, and these changes were maintained throughout the 3-month follow-up. These results are similar to face-to-face psychotherapy [54] and comparable to guided internet-delivered CBTs, notable given the short treatment duration, younger age, and relatively higher severity of the sample, all of which are associated with lower odds of response and remission in digital treatment [55].

Encouragingly, app ratings were above average for mental health apps [56]. Compared to other mHealth for depression or anxiety [57-60] and treatment-as-usual for depression [61], overall treatment satisfaction scores (CSQ) were also excellent. This was achieved despite clinician time that was below typical courses of CBT for depression (8 sessions averaging 24 minutes vs upwards of 20 sessions lasting on average 45-50 minutes) [62]. Indeed, much of the time-consuming didactic content (eg, psychoeducation about the CBT model) was administered by the app through readings, videos, and practice questions, conserving clinician time for more personalized skills' review, practice, and tailoring and addressing risk issues. App usage data were excellent, with most participants reaching the final step by the end of 8 weeks as intended and reporting regularly practicing skills on or off the app each week. Moreover, participants rarely used the messaging function between sessions (eg, seeking additional clarification or encouragement), which would be unbillable clinician time. In this way, Mindset for Depression has the potential to improve the reach of CBT therapists and hopefully reduce treatment gaps, particularly for underserved communities [63,64].

These results are important because the high cost and limited availability of trained clinicians are major barriers to the dissemination of traditional psychotherapy [65]. Our findings add to an emerging literature demonstrating the potential of guided smartphone-based CBT to mitigate these challenges. However, although numerous seemingly efficacious therapist-supported digital treatments have been created for depression [10], few are available outside of research settings or integrated into a health care system [25]. Created in collaboration with an industry partner to accelerate the dissemination pipeline and allow for ongoing technical maintenance and improvements, Mindset for Depression is commercially available and poised to be truly scalable and successful in real clinical settings. Setting it apart, Mindset for Depression was collaboratively developed with a design team, clinicians, and people with lived experience as well as rigorously applied user interface and experience best practices for mobile platforms. Critically, "users" in the user-centered design process included both patients and clinicians. This approach aligns clinical and engagement incentives so that one is not delivered at the expense of the other and yields an easy to use, streamlined, and effective treatment program. Concretely, this translated to pacing content (delivering or unlocking intervention components in a stepwise manner to encourage practice and mastery of concepts and skills that build on one another), shorter activity lengths, creating a professional and approachable tone, and inclusion of feedback loops. As standardized content is delivered via the app in each step, supporting clinicians are able to prioritize personalizing treatment and use their specialized skillsets, such as addressing unique barriers to motivation, engagement, or response. Critical next steps would be to directly evaluate the program and its readiness to scale in larger scale effectiveness trials and real-world settings.

The results also provide important guidance for improving the program. First, participant feedback (eg, uMARS engagement subscale) indicates that increased customization and interactivity could improve the app's appeal. This is consistent with the larger literature showing user preferences for apps with such features and negative reactions to apps whose content is repetitive and not personally relevant [66]. Although most participants shared positive views of the Mindset app, including indicating that they would recommend the app to friends, there was a mixed response regarding their willingness to pay for the Mindset app. It is unclear to what extent this reflects (1) that this question was asked after treatment and thus patients no longer felt the need to use the app; patients were meant to complete all therapeutic content within the 8-week treatment period; (2) a gap between what patients find beneficial and what they are willing to pay for; other studies have similarly found a reluctance to pay for mental health apps [66,67]; or (3) whether the app and concurrent therapist support were experienced as critically linked, and thus the app alone was not as valued. Indeed, half of participants indicated weekly brief sessions were the exact right amount of therapist contact and only 2 would have preferred less contact. Resolving this question will be necessary for developing a commercially sustainable implementation plan. Moreover, future iterations would benefit from broadening outcomes of interest. For example, beyond reducing symptoms of depression or other mental health concerns, an optimal intervention would also foster positive emotions and thriving and perhaps target common comorbidities, such as sleep difficulties or substance use. These outcomes should be captured in future studies and additional treatment components integrated as appropriate.

Limitations

The study has limitations that should be considered. First, the study has the inherent limitation of an open trial. Without a control group, we cannot conclusively determine that the treatment causes improvements in symptoms. Future controlled large-scale trials are needed. Second, the patient sample was self-selected, recruitment platforms were diverse, and our study therapists were trained in the use of digital therapeutics. Thus, patient and clinician stakeholders might have been biased toward individuals who are motivated by app-based therapy. Future research in real-world clinical settings is warranted. Third, although the large proportion of White women in the sample is consistent with past work and higher MDD prevalence and rate of treatment seeking in women [10], greater representation of patients with other racial and gender identities would strengthen our conclusions and ongoing treatment improvements. Fourth, we had adequate power to detect moderate to large treatment effects; a larger replication is needed to explore moderators and mediators, which are important for tiered care models. Finally, a longer follow-up period and health economics metrics would be required to see the full time and cost-savings potential of Mindset for Depression.

Conclusions

Mindset for Depression offers flexible app-led psychoeducation, skills practice, and support to patients with complementary clinician guidance to promote sustained engagement, monitor safety, and tailor treatment further to individual patient needs. The findings show that Mindset for Depression is a feasible, acceptable, and efficacious tool for adults with MDD. The hope is that such a program could be one cost-effective solution to barriers to psychotherapy dissemination and significantly increase access to evidence-based care. Although these initial results are very promising, more work remains to personalize the amount of therapist support and dose of treatment individuals receive to optimize treatment and increase rates of response and remission. The next steps include testing Mindset for Depression in a fully powered randomized controlled trial as well as the real-world clinical settings in which it is deployed.

Acknowledgments

The research was supported by Koa Health. Investigators from Massachusetts General Hospital (MGH) (SW, EEB, KHB, JLG, and HW) developed the Mindset app in collaboration with technologists and designers from Koa Health via a user-centered, iterative design approach. The investigators from MGH were responsible for the study design and the execution of the study. Koa Health supported the technical activities related to the development and deployment of the app. Koa Health had no role in the recruitment of participants and had no access to data during the trial. The sponsor was able to make minor edits to the draft of the paper. HW was supported in part by the National Institute of Mental Health of the National Institutes of Health under award number K23MH119372. KHB was supported in part by the National Institute of Mental Health of the National Institutes of Health under award number K23MH120436. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We would like to thank the following individuals for their invaluable work on the project: Lauren Fisher, Margaret Hall, Brynn Huguenel, Adam Jaroszewski, Geneva Jonathan, Jennifer Ragan, and Jasmine Williams.

Conflicts of Interest

SW has received royalties from Elsevier Publications, Guilford Publications, New Harbinger Publications, Springer, and Oxford University Press. SW has also received speaking honoraria from various academic institutions and foundations, including a research award from the National Alliance on Mental Illness. In addition, she received honoraria for her role on the Scientific Advisory Board for One-Mind (PsyberGuide), Koa Health Inc, Noom Inc, and Jimini Health. SW has received research support from Koa Health. EEB has received research support from Koa Health. EEB has a consulting agreement with Otsuka Pharmaceutical Development & Commercialization Inc and is on the Scientific Advisory Board for AugMend Health Inc. KHB has received research support from Koa Health. IS has received research support from Koa Health. SSH has received research support from Koa Health. DK has received research support from Koa Health. JLG has received research support from Koa Health. She has received speaking honoraria from L'Oreal (for a presentation at a SkinCeuticals cosmetic surgery and dermatology conference) and RBC Consultants for the CeraVe Psychodermatology Advisory Board. HW has received research support from Koa Health and the National Institute of Mental Health (NIMH; K23MH119372). HW has a consulting agreement with Hello Therapeutics, Inc. THM has received research support from Koa Health. He has received unrelated editorial honoraria from Springer Nature as well as grants to his institution from InterSystems, National Institute of Mental Health, National Institute of Nursing Research, National Institute on Aging, and National Human Genome Research Institute. OH is the founder/chief executive officer of Koa Health, a digital mental health company that collaborated with SW and her team at MGH to build Mindset. OH also serves on the World Health Organization Roster of Experts for Digital Health, sits on the Board of EMPOWER (a nonprofit organization promoting the training of community health workers to provide mental health care), and is a member of the expert panel for implementing the Wellcome Trust's mental health strategy. Dr Harrison is a Royal Society Entrepreneur in Residence in healthcare Artificial Intelligence (AI) at Oxford University.

Multimedia Appendix 1
Mindset Therapist Adherence Scale.
[DOCX File , 20 KB - mental_v11i1e53998_app1.docx]

References

1. Diagnostic and Statistical Manual of Mental Disorders, 5th Edition. Arlington, VA: American Psychiatric Publishing; 2013.
2. 2020 National Survey on Drug Use and Health (NSDUH) releases. Substance Abuse and Mental Health Services Administration. 2020. URL: https://www.samhsa.gov/data/release/2020-national-survey-drug-use-and-health-nsduh-releases [accessed 2024-03-12]
3. Ettman CK, Cohen GH, Abdalla SM, Sampson L, Trinquart L, Castrucci BC, et al. Persistent depressive symptoms during COVID-19: a national, population-representative, longitudinal study of U.S. adults. Lancet Reg Health Am 2022;5:100091 [FREE Full text] [doi: 10.1016/j.lana.2021.100091] [Medline: 34635882]
4. Friedrich MJ. Depression is the leading cause of disability around the world. JAMA 2017;317(15):1517. [doi: 10.1001/jama.2017.3826] [Medline: 28418490]
5. Greenberg PE, Fournier AA, Sisitsky T, Simes M, Berman R, Koenigsberg SH, et al. The economic burden of adults with major depressive disorder in the United States (2010 and 2018). Pharmacoeconomics 2021;39(6):653-665 [FREE Full text] [doi: 10.1007/s40273-021-01019-4] [Medline: 33950419]
6. Moitra M, Santomauro D, Collins PY, Vos T, Whiteford H, Saxena S, et al. The global gap in treatment coverage for major depressive disorder in 84 countries from 2000-2019: a systematic review and Bayesian meta-regression analysis. PLoS Med 2022;19(2):e1003901 [FREE Full text] [doi: 10.1371/journal.pmed.1003901] [Medline: 35167593]
7. López-López JA, Davies SR, Caldwell DM, Churchill R, Peters TJ, Tallon D, et al. The process and delivery of CBT for depression in adults: a systematic review and network meta-analysis. Psychol Med 2019;49(12):1937-1947 [FREE Full text] [doi: 10.1017/S003329171900120X] [Medline: 31179960]
8. Andrade LH, Alonso J, Mneimneh Z, Wells JE, Al-Hamzawi A, Borges G, et al. Barriers to mental health treatment: results from the WHO World Mental Health surveys. Psychol Med 2014;44(6):1303-1317 [FREE Full text] [doi: 10.1017/S0033291713001943] [Medline: 23931656]
9. Webb CA, Rosso IM, Rauch SL. Internet-based cognitive-behavioral therapy for depression: current progress and future directions. Harv Rev Psychiatry 2017;25(3):114-122 [FREE Full text] [doi: 10.1097/HRP.0000000000000139] [Medline: 28475503]
10. Børtveit L, Dechsling A, Sütterlin S, Nordgreen T, Nordahl-Hansen A. Guided internet-delivered treatment for depression: scoping review. JMIR Ment Health 2022;9(10):e37342 [FREE Full text] [doi: 10.2196/37342] [Medline: 36194467]
11. Torous J, Levin ME, Ahern DK, Oser ML. Cognitive behavioral mobile applications: clinical studies, marketplace overview, and research agenda. Cogn Behav Pract 2017;24(2):215-225. [doi: 10.1016/j.cbpra.2016.05.007]
12. Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry 2019;18(3):325-336 [FREE Full text] [doi: 10.1002/wps.20673] [Medline: 31496095]
13. Mohr DC, Azocar F, Bertagnolli A, Choudhury T, Chrisp P, Frank R, et al. Banbury forum consensus statement on the path forward for digital mental health treatment. Psychiatr Serv 2021;72(6):677-683 [FREE Full text] [doi: 10.1176/appi.ps.202000561] [Medline: 33467872]
14. Wasil AR, Gillespie S, Shingleton R, Wilks CR, Weisz JR. Examining the reach of smartphone apps for depression and anxiety. Am J Psychiatry 2020;177(5):464-465 [FREE Full text] [doi: 10.1176/appi.ajp.2019.19090905] [Medline: 32354266]
15. Lattie EG, Schueller SM, Sargent E, Stiles-Shields C, Tomasino KN, Corden ME, et al. Uptake and usage of Intellicare: a publicly available suite of mental health and well-being apps. Internet Interv 2016;4(2):152-158 [FREE Full text] [doi: 10.1016/j.invent.2016.06.003] [Medline: 27398319]
16. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
17. Mohr DC, Cuijpers P, Lehman K. Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions. J Med Internet Res 2011;13(1):e30 [FREE Full text] [doi: 10.2196/jmir.1602] [Medline: 21393123]
18. Mohr DC, Schueller SM, Tomasino KN, Kaiser SM, Alam N, Karr C, et al. Comparison of the effects of coaching and receipt of app recommendations on depression, anxiety, and engagement in the IntelliCare platform: factorial randomized controlled trial. J Med Internet Res 2019;21(8):e13609 [FREE Full text] [doi: 10.2196/13609] [Medline: 31464192]
19. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018;21(3):116-119 [FREE Full text] [doi: 10.1136/eb-2018-102891] [Medline: 29871870]

20. Muñoz RF. The efficiency model of support and the creation of digital apothecaries. Clin Psychol (New York) 2017;24(1):46-49. [doi: 10.1111/cpsp.12174]
21. Andersson G, Topooco N, Havik O, Nordgreen T. Internet-supported versus face-to-face cognitive behavior therapy for depression. Expert Rev Neurother 2016;16(1):55-60. [doi: 10.1586/14737175.2015.1125783] [Medline: 26610160]
22. Andersson G, Cuijpers P, Carlbring P, Riper H, Hedman E. Guided internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: a systematic review and meta-analysis. World Psychiatry 2014;13(3):288-295 [FREE Full text] [doi: 10.1002/wps.20151] [Medline: 25273302]
23. Carlbring P, Andersson G, Cuijpers P, Riper H, Hedman-Lagerlöf E. Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. Cogn Behav Ther 2018;47(1):1-18 [FREE Full text] [doi: 10.1080/16506073.2017.1401115] [Medline: 29215315]
24. Shechter A, Diaz F, Moise N, Anstey DE, Ye S, Agarwal S, et al. Psychological distress, coping behaviors, and preferences for support among New York healthcare workers during the COVID-19 pandemic. Gen Hosp Psychiatry 2020;66:1-8 [FREE Full text] [doi: 10.1016/j.genhosppsych.2020.06.007] [Medline: 32590254]
25. Bernstein EE, Weingarden H, Wolfe EC, Hall MD, Snorrason I, Wilhelm S. Human support in app-based cognitive behavioral therapies for emotional disorders: scoping review. J Med Internet Res 2022;24(4):e33307 [FREE Full text] [doi: 10.2196/33307] [Medline: 35394434]
26. Chang S, Gray L, Torous J. Smartphone app engagement and clinical outcomes in a hybrid clinic. Psychiatry Res 2023;319:115015. [doi: 10.1016/j.psychres.2022.115015] [Medline: 36549096]
27. Lakhtakia T, Torous J. Current directions in digital interventions for mood and anxiety disorders. Curr Opin Psychiatry 2022;35(2):130-135. [doi: 10.1097/YCO.0000000000000772] [Medline: 34966117]
28. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FMG, Furukawa TA, Cuijpers P, Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021;78(4):361-371 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
29. Depression in adults: treatment and management [NICE guideline NG222]. National Institute for Health and Care Excellence. URL: https://www.nice.org.uk/guidance/ng222/ [accessed 2024-03-12]
30. Posner K, Brent D, Lucas C, Gould M, Stanley B, Brown G, et al. Columbia-Suicide Severity Rating Scale (C-SSRS). The Columbia Lighthouse Project. 2008. URL: https://cssrs.columbia.edu/wp-content/uploads/C-SSRS_Pediatric-SLC_11.14.16.pdf [accessed 2024-03-12]
31. Forde F, Frame M, Hanlon P, MacLean G, Nolan D, Shajahan P, et al. Optimum number of sessions for depression and anxiety. Nurs Times 2005;101(43):36-40. [Medline: 16276843]
32. Mohr DC, Weingardt KR, Reddy M, Schueller SM. Three problems with current digital mental health research . . . and three things we can do about them. Psychiatr Serv 2017;68(5):427-429 [FREE Full text] [doi: 10.1176/appi.ps.201600541] [Medline: 28412890]
33. Ciharova M, Furukawa TA, Efthimiou O, Karyotaki E, Miguel C, Noma H, et al. Cognitive restructuring, behavioral activation and cognitive-behavioral therapy in the treatment of adult depression: a network meta-analysis. J Consult Clin Psychol 2021;89(6):563-574. [doi: 10.1037/ccp0000654] [Medline: 34264703]
34. Sauer-Zavala S, Bentley KH, Steele SJ, Tirpak JW, Ametaj AA, Nauphal M, et al. Treating depressive disorders with the Unified Protocol: a preliminary randomized evaluation. J Affect Disord 2020;264:438-445 [FREE Full text] [doi: 10.1016/j.jad.2019.11.072] [Medline: 31759672]
35. Beck AT. Cognitive Therapy of Depression. New York: Guilford Publications; 1979.
36. Bentley KH, Bernstein EE, Wallace B, Mischoulon D. Treatment for anxiety and comorbid depressive disorders: transdiagnostic cognitive-behavioral strategies. Psychiatr Ann 2021;51(5):226-230 [FREE Full text] [doi: 10.3928/00485713-20210414-01] [Medline: 34433988]
37. Dimidjian S, Martell CR, Herman-Dunn R, Hubley S. Behavioral activation for depression. In: Barlow DH, editor. Clinical Handbook of Psychological Disorders: A Step-by-Step Treatment Manual, 5th Edition. New York: Guilford Publications; 2014:353-393.
38. Sturmey P. Behavioral activation is an evidence-based treatment for depression. Behav Modif 2009;33(6):818-829. [doi: 10.1177/0145445509350094] [Medline: 19933444]
39. MacKenzie MB, Kocovski NL. Mindfulness-based cognitive therapy for depression: trends and developments. Psychol Res Behav Manag 2016;9:125-132 [FREE Full text] [doi: 10.2147/PRBM.S63949] [Medline: 27274325]
40. Akçayır G, Akçayır M. The flipped classroom: a review of its advantages and challenges. Comput Educ 2018;126:334-345. [doi: 10.1016/j.compedu.2018.07.021]
41. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 1998;59(Suppl 20):22-33;quiz 34. [Medline: 9881538]
42. Devilly GJ, Borkovec TD. Psychometric properties of the credibility/expectancy questionnaire. J Behav Ther Exp Psychiatry 2000;31(2):73-86. [doi: 10.1016/s0005-7916(00)00012-4] [Medline: 11132119]

43. Stoyanov SR, Hides L, Kavanagh DJ, Wilson H. Development and validation of the user Version of the Mobile Application Rating Scale (uMARS). JMIR Mhealth Uhealth 2016;4(2):e72 [FREE Full text] [doi: 10.2196/mhealth.5849] [Medline: 27287964]
44. Attkisson CC, Zwick R. The client satisfaction questionnaire. Psychometric properties and correlations with service utilization and psychotherapy outcome. Eval Program Plann 1982;5(3):233-237. [doi: 10.1016/0149-7189(82)90074-x] [Medline: 10259963]
45. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry 1960;23(1):56-62 [FREE Full text] [doi: 10.1136/jnnp.23.1.56] [Medline: 14399272]
46. Rosenblat JD, Lee Y, McIntyre RS. The effect of pharmacogenomic testing on response and remission rates in the acute treatment of major depressive disorder: a meta-analysis. J Affect Disord 2018;241:484-491. [doi: 10.1016/j.jad.2018.08.056] [Medline: 30149336]
47. Nierenberg AA, DeCecco LM. Definitions of antidepressant treatment response, remission, nonresponse, partial response, and other relevant outcomes: a focus on treatment-resistant depression. J Clin Psychiatry 2001;62(Suppl 16):5-9. [Medline: 11480882]
48. Frank E, Prien RF, Jarrett RB, Keller MB, Kupfer DJ, Lavori PW, et al. Conceptualization and rationale for consensus definitions of terms in major depressive disorder. Remission, recovery, relapse, and recurrence. Arch Gen Psychiatry 1991;48(9):851-855. [doi: 10.1001/archpsyc.1991.01810330075011] [Medline: 1929776]
49. Mundt JC, Marks IM, Shear MK, Greist JH. The work and social adjustment scale: a simple measure of impairment in functioning. Br J Psychiatry 2002;180:461-464 [FREE Full text] [doi: 10.1192/bjp.180.5.461] [Medline: 11983645]
50. Endicott J, Nee J, Harrison W, Blumenthal R. Quality of life enjoyment and satisfaction questionnaire: a new measure. Psychopharmacol Bull 1993;29(2):321-326. [Medline: 8290681]
51. Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary care evaluation of mental disorders. Patient health questionnaire. JAMA 1999;282(18):1737-1744 [FREE Full text] [doi: 10.1001/jama.282.18.1737] [Medline: 10568646]
52. Cuijpers P, Cristea IA, Karyotaki E, Reijnders M, Huibers MJH. How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence. World Psychiatry 2016;15(3):245-258 [FREE Full text] [doi: 10.1002/wps.20346] [Medline: 27717254]
53. Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 2013;4:863 [FREE Full text] [doi: 10.3389/fpsyg.2013.00863] [Medline: 24324449]
54. Cuijpers P, Karyotaki E, Weitz E, Andersson G, Hollon SD, van Straten A. The effects of psychotherapies for major depression in adults on remission, recovery and improvement: a meta-analysis. J Affect Disord 2014;159:118-126. [doi: 10.1016/j.jad.2014.02.026] [Medline: 24679399]
55. Karyotaki E, Ebert DD, Donkin L, Riper H, Twisk J, Burger S, et al. Do guided internet-based interventions result in clinically relevant changes for patients with depression? An individual participant data meta-analysis. Clin Psychol Rev 2018;63:80-92. [doi: 10.1016/j.cpr.2018.06.007] [Medline: 29940401]
56. Hudson G, Negbenose E, Neary M, Jansli SM, Schueller SM, Wykes T, et al. Comparing professional and consumer ratings of mental health apps: mixed methods study. JMIR Form Res 2022;6(9):e39813 [FREE Full text] [doi: 10.2196/39813] [Medline: 36149733]
57. Schlicker S, Baumeister H, Buntrock C, Sander L, Paganini S, Lin J, et al. A web- and mobile-based intervention for comorbid, recurrent depression in patients with chronic back pain on sick leave (Get.Back): pilot randomized controlled trial on feasibility, user satisfaction, and effectiveness. JMIR Ment Health 2020;7(4):e16398 [FREE Full text] [doi: 10.2196/16398] [Medline: 32293577]
58. Netter AL, Beintner I, Brakemeier EL. Adding an app-based intervention to the cognitive behavioral analysis system of psychotherapy in routine outpatient psychotherapy treatment: proof-of-concept study. JMIR Form Res 2022;6(8):e35482 [FREE Full text] [doi: 10.2196/35482] [Medline: 35943764]
59. Sharma G, Schlosser L, Jones BDM, Blumberger DM, Gratzer D, Husain MO, et al. Brief app-based cognitive behavioral therapy for anxiety symptoms in psychiatric inpatients: feasibility randomized controlled trial. JMIR Form Res 2022;6(11):e38460 [FREE Full text] [doi: 10.2196/38460] [Medline: 36322113]
60. Gotthardt M, Striegl J, Loitsch C, Weber G. Voice assistant-based CBT for depression in students: effects of empathy-driven dialog management. In: Miesenberger K, Kouroupetroglou G, Mavrou K, Manduchi R, Rodriguez MC, Penáz P, editors. Computers Helping People with Special Needs. ICCHP-AAATE 2022. Lecture Notes in Computer Science, Vol 13341. Cham: Springer; 2022:451-461.
61. Richards DA, Hill JJ, Gask L, Lovell K, Chew-Graham C, Bower P, et al. Clinical effectiveness of collaborative care for depression in UK primary care (CADET): cluster randomised controlled trial. BMJ 2013;347:f4913 [FREE Full text] [doi: 10.1136/bmj.f4913] [Medline: 23959152]
62. Cuijpers P, Berking M, Andersson G, Quigley L, Kleiboer A, Dobson KS. A meta-analysis of cognitive-behavioural therapy for adult depression, alone and in comparison with other treatments. Can J Psychiatry 2013;58(7):376-385 [FREE Full text] [doi: 10.1177/070674371305800702] [Medline: 23870719]

63. Ramos G, Chavira DA. Use of technology to provide mental health care for racial and ethnic minorities: evidence, promise, and challenges. Cogn Behav Pract 2022;29(1):15-40. [doi: 10.1016/j.cbpra.2019.10.004]
64. Bae H, Shin H, Ji HG, Kwon JS, Kim H, Hur JW. App-based interventions for moderate to severe depression: a systematic review and meta-analysis. JAMA Netw Open 2023;6(11):e2344120 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.44120] [Medline: 37983028]
65. Cavanagh K. Geographic inequity in the availability of cognitive behavioural therapy in England and Wales: a 10-year update. Behav Cogn Psychother 2014;42(4):497-501. [doi: 10.1017/S1352465813000568] [Medline: 23829940]
66. Balaskas A, Schueller SM, Cox AL, Rashleigh C, Doherty G. Examining young adults daily perspectives on usage of anxiety apps: a user study. PLOS Digit Health 2023;2(1):e0000185 [FREE Full text] [doi: 10.1371/journal.pdig.0000185] [Medline: 36812622]
67. Bakker D, Kazantzis N, Rickwood D, Rickard N. Development and pilot evaluation of smartphone-delivered cognitive behavior therapy strategies for mood- and anxiety-related problems: MoodMission. Cogn Behav Pract 2018;25(4):496-514. [doi: 10.1016/j.cbpra.2018.07.002]

Abbreviations

CBT: cognitive behavioral therapy
CEQ: Credibility/Expectancy Questionnaire
CSQ: Client Satisfaction Questionnaire
DSM-5: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
HAM-D: Hamilton Depression Rating Scale
HIPAA: Health Insurance Portability and Accountability Act
MDD: major depressive disorder
PHQ-9: Patient Health Questionnaire-9
Q-LES-Q-SF: The Quality of Life Enjoyment and Satisfaction Questionnaire-Short Form
SMART: specific, measurable, achievable, relevant, and time-bound
uMARS: Mobile Application Rating Scale User Version
WSAS: Work and Social Adjustment Scale

Edited by J Torous; submitted 26.10.23; peer-reviewed by A Brahmer, NH Zainal; comments to author 25.11.23; revised version received 19.12.23; accepted 03.01.24; published 09.04.24.

Please cite as:
Wilhelm S, Bernstein EE, Bentley KH, Snorrason I, Hoeppner SS, Klare D, Greenberg JL, Weingarden H, McCoy TH, Harrison O
Feasibility, Acceptability, and Preliminary Efficacy of a Smartphone App–Led Cognitive Behavioral Therapy for Depression Under Therapist Supervision: Open Trial
JMIR Ment Health 2024;11:e53998
URL: https://mental.jmir.org/2024/1/e53998
doi:10.2196/53998
PMID:38592771

Original Paper

Examining the Efficacy of Extended Reality–Enhanced Behavioral Activation for Adults With Major Depressive Disorder: Randomized Controlled Trial

Margot Paul[1], PsyD; Kim Bullock[1], MD; Jeremy Bailenson[2], PhD; David Burns[1], MD

[1]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA, United States

[2]Department of Communication, Stanford University, Stanford, CA, United States

Corresponding Author:
Margot Paul, PsyD
Department of Psychiatry and Behavioral Sciences
Stanford University School of Medicine
401 Quarry Rd
Stanford, CA, 94305
United States
Phone: 1 781 572 4136
Email: mdpaul@stanford.edu

Abstract

Background: Major depressive disorder (MDD) is a global concern with increasing prevalence. While many evidence-based psychotherapies (EBPs) have been identified to treat MDD, there are numerous barriers to patients accessing them. Virtual reality (VR) has been used as a treatment enhancement for a variety of mental health disorders, but few studies have examined its clinical use in treating MDD. Behavioral activation (BA) is a simple yet effective and established first-line EBP for MDD that has the potential to be easily enhanced and adapted with VR technology. A previous report by our group explored the feasibility and acceptability of VR-enhanced BA in a small clinical proof-of-concept pilot. This study examines the clinical efficacy of a more immersive extended reality (XR)–enhanced BA (XR-BA) prototype. This is the first clinical efficacy test of an XR-BA protocol.

Objective: This study examined whether XR-BA was feasible and efficacious in treating MDD in an ambulatory telemedicine clinic.

Methods: A nonblinded between-subject randomized controlled trial compared XR-BA to traditional BA delivered via telehealth. The study used a previously established, brief 3-week, 4-session BA EBP intervention. The experimental XR-BA participants were directed to use a Meta Quest 2 (Reality Labs) VR headset to engage in simulated pleasant or mastery activities and were compared to a control arm, which used only real-life mastery or pleasant activities as between-session homework. The Patient Health Questionnaire (PHQ)–9 was the primary outcome measure. Independent-sample and paired-sample *t* tests (2-tailed) were used to determine statistical significance and confirmed using structural equation modeling.

Results: Overall, 26 participants with MDD were randomized to receive either XR-BA (n=13, 50%) or traditional BA (n=13, 50%). The mean age of the 26 participants (n=6, 23% male; n=19, 73% female; n=1, 4% nonbinary or third gender) was 50.3 (SD 17.3) years. No adverse events were reported in either group, and no substantial differences in dropout rates or homework completion were observed. XR-BA was found to be statistically noninferior to traditional BA ($t_{18.6}$=−0.28; *P*=.78). Both the XR-BA (t_9=2.5; *P*=.04) and traditional BA (t_{10}=2.3; *P*=.04) arms showed a statistically significant decrease in PHQ-9 and clinical severity from the beginning of session 1 to the beginning of session 4. There was a significant decrease in PHQ-8 to PHQ-9 scores between the phone intake and the beginning of session 1 for the XR-BA group (t_{11}=2.6; *P*=.03) but not the traditional BA group (t_{11}=1.4; *P*=.20).

Conclusions: This study confirmed previous findings that XR-BA may be a feasible, non-inferior, and acceptable enhancement to traditional BA. Additionally, there was evidence that supports the potential of XR to enhance expectation or placebo effects. Further research is needed to examine the potential of XR to improve access, outcomes, and barriers to MDD care.

Trial Registration: ClinicalTrials.gov NCT05525390; https://clinicaltrials.gov/study/NCT05525390

(JMIR Ment Health 2024;11:e52326) doi:10.2196/52326

KEYWORDS

virtual reality; extended reality; major depressive disorder; behavioral activation; depression; Meta Quest 2

Introduction

Background

Major depressive disorder (MDD) is a global concern with increasing cases worldwide [1]. Depressive disorders are the most significant contributors to nonfatal health loss worldwide, with a 37.9% increase in their economic burden from 2010 to 2020 [1,2]. MDD is associated with suicide, which is one of the leading causes of death in young adults. Although many evidence-based psychotherapies (EBPs) for MDD exist, less than 1 in 4 people in low- to middle-income countries receive these treatments [1]. Thus, creating solutions for access to care is a priority in the treatment of MDD.

Due to its simplicity and efficacy, behavioral activation (BA) is one of the most widely used first-line EBP for MDD [3]. The behavioral theory underpinning this intervention postulates that depression is due to an avoidance pattern of increasingly less frequent engagement in pleasurable or mastery activities [3,4]. BA provides the tools for reversing this pattern through intentional scheduling of and engaging in positive and mastery activities. Despite BA's usefulness and effectiveness, few patients with MDD ever obtain access to BA due to barriers and obstacles such as physical limitations, financial constraints to accessing activities, social isolation, mental health stigma, or a lack of trained providers [5,6].

Extended reality (XR) is a term that is used to describe all current and future immersive technologies, including virtual reality (VR) and augmented reality. XR is becoming increasingly popular, with approximately 1 in 5 consumers in the United States using it in 2020 and an estimated 70.8 million people in the United States using it at least once per month in 2023 [7]. Immersive technologies such as VR are being used to solve multiple barriers to mental health care, such as improving access to content not readily accessible in real life (IRL). The use of VR to enhance the treatment of anxiety and trauma disorders has been reported on for the past 3 decades owing to its ability to easily and reliably provide controlled cue desensitization [8].

While there is a preponderance of support illustrating XR's efficacy in enhancing EBP and reducing barriers to care, surprisingly few clinical studies have examined its use directly in populations with MDD [8,9]. While several studies have examined the use of VR to treat mood disorders, to our knowledge, only 1 clinical trial focusing on using VR-enhanced BA (VR-BA) in a population with MDD has been completed to date [9,10]. This previous feasibility study, completed by our group, observed evidence of improvement in MDD outcomes using VR to simulate pleasant activities during a brief BA intervention [10].

This study is an extension of our previous VR-BA feasibility study and remains in line with the international working group's methodological framework for developing the design, implementation, analysis, interpretation, and communication of trials of novel VR behavioral health treatments [11]. We iterated our previous VR-BA prototype based on pilot-testing and feedback to a more immersive, embodied, and autonomous XR prototype (XR-enhanced BA [XR-BA]) using a commercially available Meta Quest 2 (Reality Labs) VR headset. On the basis of user feedback, we made the XR-BA protocol identical to traditional in vivo or IRL BA by allowing patients to freely decide which XR pleasant activities to engage in rather than being restricted to a list of curated pleasant experiences. This allowed for a more personalized customization and a higher sense of autonomy by users and mirrors the elements associated with traditional BA. It also more closely mimics traditional BA by refraining from confining participants to preselected VR choices. While the devices were not Health Insurance Portability and Accountability Act–compliant at this time, participants were informed of the potential privacy risks associated with Meta potentially tracking their use. However, the authors and the institutional review board considered this risk akin to that of traditional BA, where activities can be observed by others or tracked through purchases. Furthermore, no mental health–specific data were collected on the device.

This study went beyond examining simple feasibility and tested the clinical efficacy of XR-BA compared to traditional BA. We specifically examined whether XR-BA is efficacious in reducing clinical depressive symptoms and shows noninferiority to traditional BA delivered without VR in an ambulatory MDD sample.

Objectives

The first aim of this study was to test the safety and feasibility of our newest prototype (XR-BA) using an embodied and interactive VR headset to engage with maximum free choice during a brief BA protocol guided by a telehealth clinician. It was hypothesized that XR-BA would be safe, feasible, and acceptable for outpatients with MDD receiving remote care.

The second aim of this study was to examine the efficacy of XR-BA compared to that of a traditional brief BA protocol for MDD. We predicted from prior work that XR-BA would not be inferior to traditional brief BA in reducing symptoms of MDD as measured using the Patient Health Questionnaire–9 (PHQ-9). The decision to compare XR-BA to a traditional BA protocol rather than a sham VR control was because engagement with VR itself can be considered a pleasant activity within BA, and thus could be a confounding variable within a sham control. Thus, with traditional BA being the gold EBP standard with known efficacy, it seemed a more meaningful comparison. If XR-BA can be as efficacious as traditional BA and more accessible to those with barriers to IRL pleasant activities, then it may prove to be an impactful enhancement to treatment.

Methods

Recruitment

Recruitment took place remotely via Zoom (Zoom Video Communications)–delivered telehealth sessions between

December 19, 2022, and July 24, 2023. The trial was registered on ClinicalTrials.gov (ID NCT05525390).

Participants were recruited locally via study flyers posted in the Stanford School of Medicine Department of Psychiatry and Behavioral Sciences located in Palo Alto, California, United States. The description of the study was also electronically listed on Stanford University's website for currently recruiting studies, ClinicalTrials.gov, and Craigslist. Without solicitation, a private web-based company called *Power* included our study on its website and connected participants with this study without any formal agreement, consent, or payment from our research group.

The inclusion criteria were as follows: age of ≥18 years; ability to speak English; and meeting of the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*, criteria for MDD. The exclusion criteria were as follows: substance use disorder in the previous year, diagnosis of any psychotic or bipolar I disorder, seizure in the previous 6 months or untreated epilepsy, current suicidal urges or intent, current nonsuicidal self-injury or parasuicidal behavior, changing psychotherapy treatment within the last 4 months before study entry, or changing psychotropic medication within 2 months of study entry. This study offered no compensation for participation.

The initial screening procedure consisted of 2 steps: an initial phone screening and a face-to-face Zoom intake session. During the initial phone screening, callers were assessed for preliminary eligibility using the Patient Health Questionnaire–8 (PHQ-8) and a brief screening questionnaire and were given the opportunity to ask questions about the study (Multimedia Appendix 1). If the initial eligibility criteria were met, as determined by the answers to the questionnaire and a PHQ-8 score of ≥10, potential participants were securely emailed a consent form to read, review, and sign at their leisure [12] (Multimedia Appendix 2). Potential participants were informed that they could reach out to the clinician with any questions before signing the consent form. After potential participants securely returned their signed consent forms, a Zoom intake session was held to determine complete study eligibility and obtain demographic information (Multimedia Appendix 3). Complete study eligibility was determined using the clinician-administered Mini-International Neuropsychiatric Interview [13]. The previously published case report and feasibility study provide further details [10,14].

Enrollment and Randomization

When a participant met the full study eligibility criteria and was enrolled to take part in the study, they were randomly assigned to 1 of the 2 study arms in a single-blind fashion using permuted blocks of 4 in sealed envelopes. Participants were notified of their randomization outcome via secure email before session 1.

Procedure

A clinical psychologist met with each participant for 30 to 50 minutes once per week for 4 sessions over Zoom to administer a brief BA therapy protocol. At the beginning of each session, all participants were verbally administered the PHQ-9. If item 9 was endorsed, a risk assessment was conducted in real time, and safety measures were taken in accordance with the risk. Both arms followed the protocol for brief BA based on the guidance of the published literature [15,16]. No participants were provided with a stipend for activities. All sessions followed the previously established protocol detailed in the case report and feasibility study [10,14].

Experimental Arm (XR-BA)

The XR-BA participants were shipped a VR Meta Quest 2 headset before the first session with a prepaid return label. This headset has a resolution of 1832 × 1920 pixels; support for a 60-, 72-, and 90-Hz refresh rate; and room scale [17].

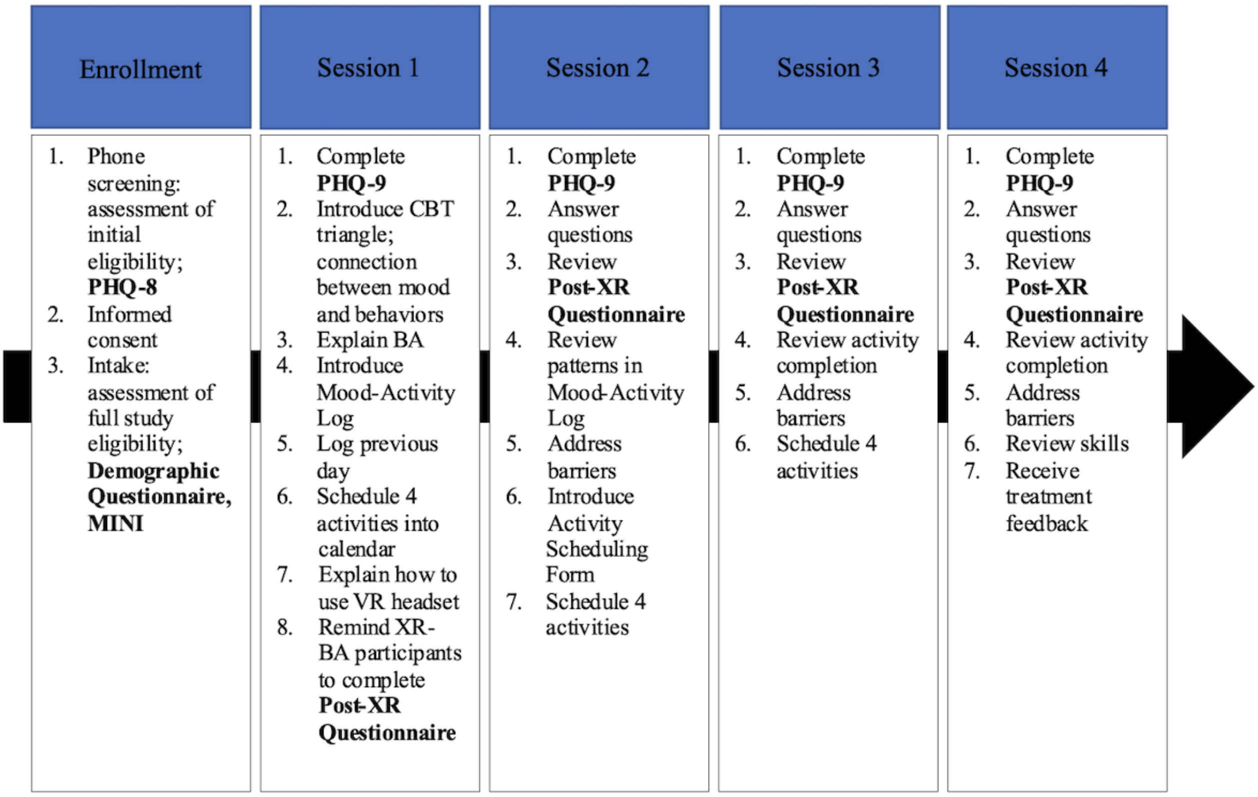
The Meta Quest 2 headsets did not have any software preloaded or prechosen. XR-BA participants were provided with an XR activity list similar to the Pleasant Events Schedule from traditional BA. The items included in the list were determined by asking subject matter experts to provide quality activity choices based on the categories provided within the Meta Quest 2 headset. While the list included different category options and ideas from those within the Meta Quest 2 headset, the clinician clarified that participants could choose any activity offered within the headset even if it was not included in the list (Multimedia Appendix 4) [18]. In between each session, participants were asked to complete ≥4 XR activities per week and 1 post-XR questionnaire pertaining to all completed XR activities from the week to assess spatial presence, simulator sickness (tolerability), and technology acceptability (Multimedia Appendix 5). This questionnaire was sent out to the participant and returned to the clinician via a secure email.

Control Arm (Traditional BA)

The participants in the control group followed the same protocol as those in the XR-BA arm except that they were not provided with a VR headset, were emailed the Pleasant Events Schedule, were asked to choose and complete ≥4 activities IRL, and were not administered the post-XR questionnaire.

The previously published case report and feasibility study provide more details [10,14]. The study timeline is shown in Figure 1.

Figure 1. Study timeline. BA: behavioral activation; CBT: cognitive behavioral therapy; MINI: Mini-International Neuropsychiatric Interview; PHQ-8: Patient Health Questionnaire–8; PHQ-9: Patient Health Questionnaire–9; VR: virtual reality; XR: extended reality; XR-BA: extended reality–enhanced behavioral activation.



Measures

The feasibility, or the degree to which XR could successfully be integrated into the brief BA protocol, was measured by commenting on qualitative barriers to use observed. Barriers were assessed by rates of dropout, adverse events, number of times the headset was used, and level of presence experienced in the headset [11]. The level of presence was obtained via participant reports using a Likert scale of 0 (*not at all*) to 4 (*very strongly*) for each question; with 3 questions, there was a possibility of yielding a score between 0 and 12.

The acceptability of the XR-BA treatment was measured via participant reports using the Technology Acceptance Model, with the agreement choices on a Likert scale ranging from 0 (*strongly disagree*) to 4 (*strongly agree*) for each question; with 3 or 4 questions per category, there was a possibility of yielding a score between 0 and 12 or 0 and 16, respectively.

The tolerability of the XR-BA treatment was measured via participant reports using the Simulator Sickness Questionnaire (SSQ), with the agreement choices on a Likert scale ranging from 0 (*no more than usual*) to 3 (*severely more than usual*); with 16 items, there was a possibility of yielding a score between 0 and 48.

The efficacy of the XR-BA treatment was measured via participant reports using the PHQ-8 and PHQ-9, with the agreement choices on a Likert scale ranging from 0 (*not at all*) to 3 (*nearly every day*); with 8 or 9 questions, there was a possibility of yielding a score between 0 and 24 or 0 and 27, respectively.

The previously published case report [14] provides a more in-depth description and background of the following measures: demographic questionnaire, Mini-International Neuropsychiatric Interview, PHQ-8, PHQ-9, presence scale, Technology Acceptance Model, and SSQ.

Of note, agitation (ie, the brief agitation measure) was not used as a measure of tolerability in this study. In addition, unlike the previous study, the number of times the headset was used was not determined from the device itself; rather, it was obtained via participant self-report.

Statistical Analyses

Overview

This study was a 2-arm nonblinded between-participant randomized controlled trial (RCT) testing the feasibility and efficacy of using XR-simulated activities compared to IRL pleasurable or mastery activities during a brief BA intervention for MDD. The Holter critical number in structural equation modeling (SEM) with the analysis of moment structures (AMOS; version 28.0; IBM Corp) [19] was used to determine whether a sample size of 26 would be needed to disprove the model if it were incorrect.

Feasibility

The average total presence for intention-to-treat (ITT) participants and protocol completers was calculated. The average presence experienced was also calculated as a percentage by dividing the average score by 12 (the maximum score). The number of questions in each category determined the outcome range (either 0-12 for 3 questions or 0-16 for 4 questions). The average percentage of acceptance was also calculated by dividing the average score by the maximum score within the outcome range. To determine the degree of acceptance, as labeled on the scale, the average score was then scaled back depending on the number of questions. For example, the "Perceived Usefulness" category included 3 questions, yielding a potential range of 0 to 12, so an average score of 10 would be divided by 3 to assess the degree of acceptance (in this case, it would yield a score of 3.33, which would correlate to "agree" on the Likert scale). Physical tolerability of the VR headset was assessed via participant reports using the SSQ, which was broken down by symptom and used a Likert scale ranging from 0 (*no more than usual*) to 3 (*severely more than usual*) for each item. The percentage of physical intolerability was calculated by dividing the average scores by the highest potential score (48).

Efficacy

To assess the clinical efficacy of the XR-BA treatment compared with the traditional BA treatment group, the participants' depression scores were measured using the PHQ-8 from the initial phone screening and the PHQ-9 from the 4 session time points. Independent-sample *t* tests (2-tailed) were used to compare the means between the 2 groups, and paired-sample *t* tests (2-tailed) were used to compare the means within each group. In addition, SEM AMOS was used to confirm the results because of its ability to compare competing models using nested tests, compare parameter estimates across groups, and estimate missing data models using full-information maximum likelihood [19-21]. SEM is widely used in the social sciences and was chosen for this study given its ability to adeptly manage missing data and exhibit greater statistical power compared to conventional multiple regression analyses [22], which was important given this study's relatively low sample size. The chi-square statistic was used to evaluate model fit [23]. Changes in chi-square values relative to changes in *df* (chi-square difference tests) were used to compare nested models. These results were also confirmed using traditional linear growth models [24].

Ethical Considerations

Ethics approval was obtained from the Stanford University institutional review board (protocol 66488) and participants provided informed consent before beginning the study. Participant data were deidentified and participants were informed of the potential privacy risks associated with Meta potentially tracking their use. No mental health–specific data were collected on the device. Participants were not compensated for participation.

Results

Participant Demographics

The sample consisted of 26 adults (mean age 50.3, SD 17.3 y; n=6, 23% male; n=19, 73% female; and n=1, 4% nonbinary or third gender), with 21 (81%; mean age 47.9, SD 17.7 y; n=5, 24% male; n=15, 71% female; and n=1, 5% nonbinary or third gender) completing the full protocol. The Holter critical number in SEM AMOS was used to determine that a sample size of 26 would be needed to disprove the model if it were incorrect. There was no significant difference in age ($t_{23.6}$=1.34; *P*=.19) or sex ($t_{23.4}$=0.71; *P*=.49) between the groups.

Figure 2 shows the CONSORT (Consolidated Standards of Reporting Trials) diagram, and Table 1 provides additional participant demographic information.

Figure 2. CONSORT (Consolidated Standards of Reporting Trials) diagram. BA: behavioral activation; MINI: Mini-International Neuropsychiatric Interview; XR-BA: extended reality–enhanced behavioral activation.

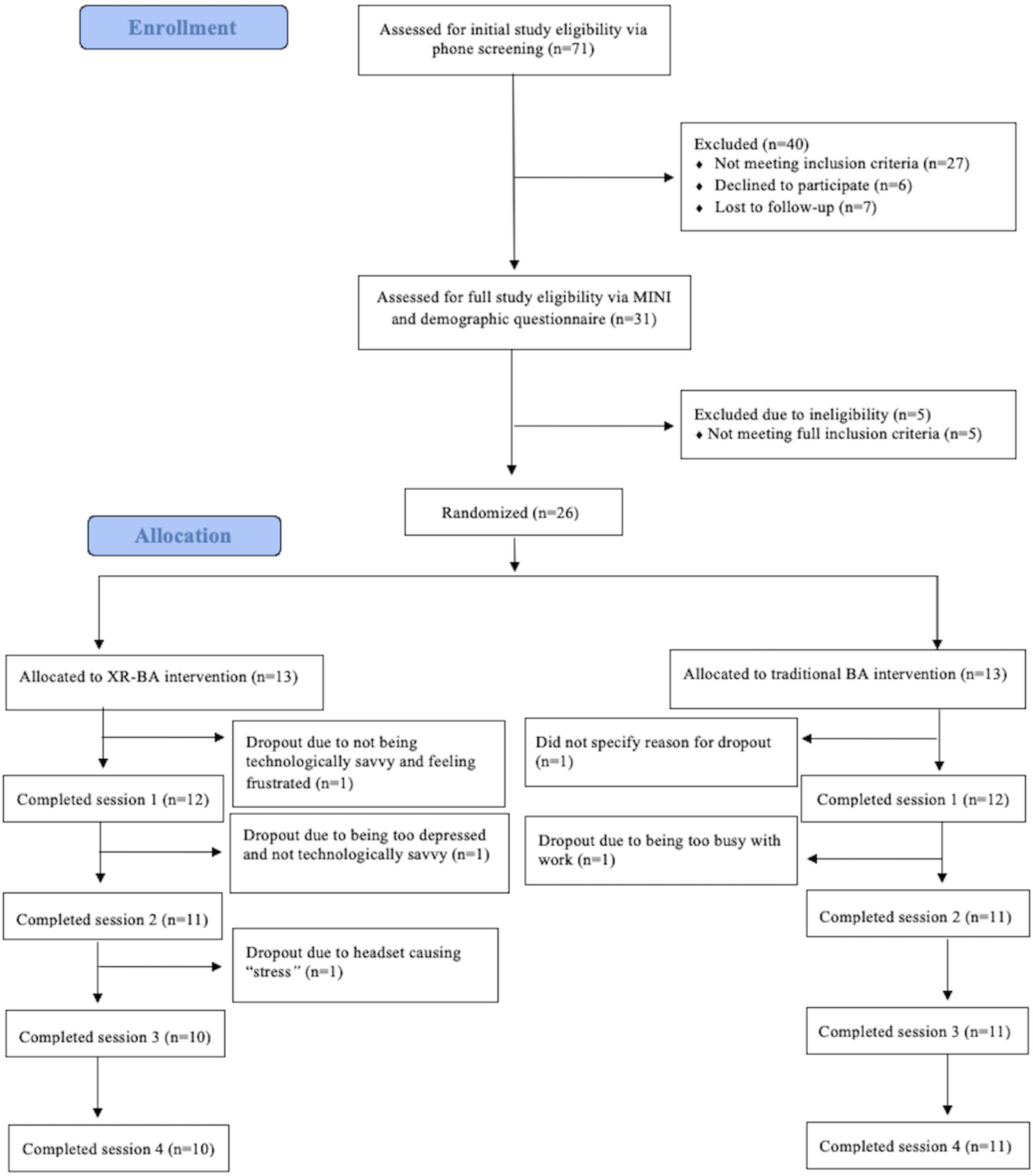


Table 1. Participant demographics (N=26).

Characteristic	XR-BA[a] (n=13), n (%)	Traditional BA[b] (n=13), n (%)	Total, n (%)
Gender			
Male	1 (8)	5 (38)	6 (23)
Female	11 (85)	8 (62)	19 (73)
Nonbinary or third gender	1 (8)	0 (0)	1 (4)
Age group (y)			
20 to 29	3 (23)	2 (15)	5 (19)
30 to 39	3 (23)	0 (0)	3 (12)
40 to 49	1 (8)	2 (15)	3 (12)
50 to 59	3 (23)	3 (23)	6 (23)
60 to 69	1 (8)	5 (38)	6 (23)
70 to 79	2 (15)	1 (8)	3 (12)
Race or ethnicity			
Asian	1 (8)	1 (8)	2 (8)
Black	0 (0)	1 (8)	1 (4)
Hispanic or Latino	0 (0)	1 (8)	1 (4)
Indian	1 (8)	2 (15)	3 (12)
Mexican	1 (8)	0 (0)	1 (4)
Non-Hispanic White	10 (77)	8 (62)	18 (69)
Previous mental health treatment			
Yes	12 (92)	12 (92)	24 (92)
No	1 (8)	1 (8)	2 (8)
Current mental health treatment			
Yes	11 (85)	6 (46)	17 (65)
Psychotherapy only	1 (9)	1 (17)	2 (12)
Psychotropic medications only	3 (27)	3 (50)	6 (35)
Psychotherapy and medications	7 (64)	2 (33)	9 (53)
No	2 (15)	7 (54)	9 (35)
Previous experience using VR[c]			
0 times	9 (69)	9 (69)	18 (69)
1 to 4 times	3 (23)	3 (23)	6 (23)
5 to 9 times	1 (8)	1 (8)	2 (8)
≥10 times	0 (0)	0 (0)	0 (0)
Purpose of previous VR use			
Gaming	3 (75)[d]	2 (50)[d]	5 (62)[e]
Treatment	0 (0)[d]	0 (0)[d]	0 (0)[e]
Research	1 (25)[d]	2 (50)[d]	3 (38)[e]

[a]XR-BA: extended reality–enhanced behavioral activation.

[b]BA: behavioral activation.

[c]VR: virtual reality.

[d]n=4.

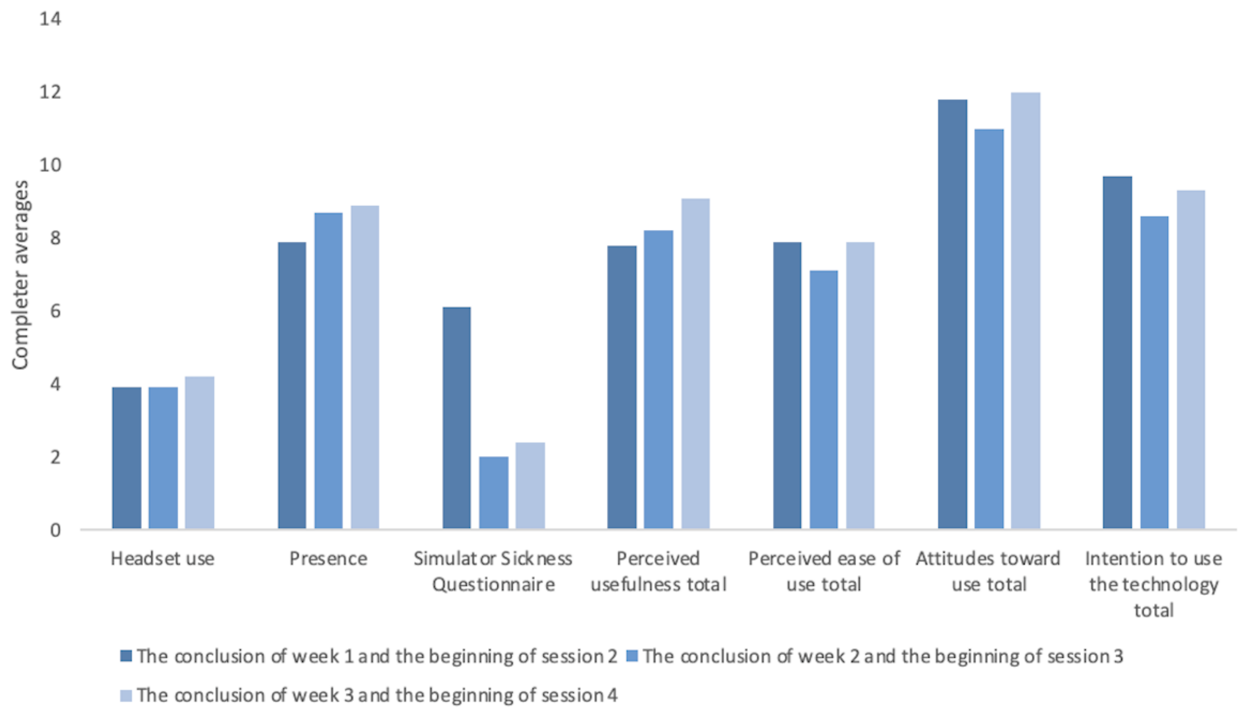
[e]n=8.

XR-BA Prototype Feasibility

The completion rates were 77% (10/13) in the XR-BA arm and 85% (11/13) in the traditional BA arm. No participants reported any serious adverse events. The participants in the XR-BA arm used the headset, on average, slightly less than suggested (encouraged a minimum of 12 times), with the completer average being 12 (SD 2.67) and the ITT participant average being 11.18 (SD 3.71). Only 8% (1/13) of the participants did not submit a post-XR questionnaire during 1 week of treatment. This participant reported that she did not use the headset during that week due to being busier than usual with work deadlines and feeling physically ill.

The average total presence rating of the ITT XR-BA participants was 68% (8.1/12; SD 2.5%), whereas the average rating of all the XR-BA completers was 71% (8.5/12; SD 2.2%). The participant who gave the lowest presence rating (3.7/12, 31%, SD 1.15%) shared that "tactile" sensations, such as feeling the sun on her skin, were important to her, and consequently, the VR did not feel immersive. Participants who completed the protocol on average indicated progressively higher levels of presence each subsequent week (Figure 3), although statistical significance was not analyzed.

Figure 3. Headset use and post–extended reality questionnaire results by week among protocol completers.



XR-BA Acceptability

Overall, the participants who completed the protocol were "neutral" on or "agreed" with the VR treatment being acceptable, with an average rating of 2.8 (SD 0.21; where 2=neutral and 3=agree) on the Likert scale and 71% (37/52) acceptability (Table 2). The participant who gave the lowest acceptability rating (26.7/52, 51%) described that the learning curve of the headset and the discomfort from the weight of the headset made VR less enjoyable. Participants who completed the protocol indicated a higher level of perceived usefulness of VR on average after each subsequent week of use (Figure 3). Between the conclusion of week 1 and the beginning of session 2 and the conclusion of week 3 and the beginning of session 4, participants who completed the protocol reported a lower level of desire to continue using the headset after treatment on average (Figure 3). While participants were provided with the XR activity list (Multimedia Appendix 4), there were certain activities from the list that participants specifically mentioned enjoying, such as YouTube 360° videos (8/13, 62%), *Tripp* (5/13, 38%), *Liminal* (5/13, 38%), *Beat Saber* (2/13, 15%), and *Painting* (2/13, 15%).

Table 2. Extended reality–enhanced behavioral activation acceptability.

	Perceived usefulness[a] (0-12; 3 items), mean (SD)	Perceived ease of use[a] (0-12; 3 items), mean (SD)	Attitudes toward use[b] (0-16; 4 items), mean (SD)	Intention to use the technology[a] (0-12; 3 items), mean (SD)
Completer average	8.4 (2.2)	7.7 (2.6)	11.7 (2.6)	9.2 (1.9)
ITT[c] average	8.1 (2.3)	7.5 (2.5)	11.1 (3.0)	8.4 (3.3)

[a]Domains comprising the Technology Acceptance Model (higher numbers indicate greater acceptability). Perceived usefulness, perceived ease of use, and intention to use the technology comprised 3 items with a range of 0 (*strongly disagree*) to 4 (*strongly agree*) for each item.

[b]Attitudes toward use comprised 4 items with a range of 0 (*strongly disagree*) to 4 (*strongly agree*) for each item.

[c]ITT: intention to treat.

XR-BA Tolerability

Physical tolerability was determined using the SSQ. Possible responses for the 16 items ranged from 0 (*no more than usual*) to 3 (*severely more than usual*). Lower numbers indicate greater tolerability. The average overall physical tolerability of those who completed the protocol and the ITT participants was 92% (44/48) and 92% (44.4/48), respectively. *Eyestrain* was the most common symptom of physical intolerability. *Burping* and *increased salivation* were the least common symptoms of physical intolerability, with 8% (1/13) of the participants endorsing burping after week 1 of headset use and 8% (1/13) of the participants endorsing increased salivation after week 1 of headset use. The participant who endorsed relatively higher overall average simulator sickness symptoms (rating of 17/48) compared to other participants experienced most of these symptoms after the first week (rating of 30/48) of headset use. This participant shared that the headset felt uncomfortable and heavy on her head and she experienced symptoms of nausea when she was immersed in any activity that had a quick-moving image. Upon trying other activities within the headset, such as the slower-moving *Liminal* and YouTube 360° videos, this participant's symptoms reduced to a rating of 4 out of 48. Overall, participants who completed the protocol experienced a decrease in simulator sickness symptoms between the conclusion of weeks 1 and 3 on average (Figure 3).

Clinical Efficacy

Participants in both study arms showed a 4-point decrease in PHQ-9 scores between sessions 1 and 4, with participants in the XR-BA arm experiencing a 4.4-point decrease and participants in the traditional BA arm experiencing a 3.7-point decrease. There was no significant difference in improvement in PHQ-9 scores between the study arms ($t_{18.6}$=−0.28; *P*=.78; Table 3).

Protocol completers in the XR-BA arm went from an average of moderately severe (15.8, SD 2.86; phone intake) to moderate (session 1: 12.8, SD 3.46; session 2: 10.6, SD 4.14; session 3: 10.5, SD 4.81) to mild (8.4, SD 3.72; session 4) symptoms of depression. The average decrease of 7.4 points on the PHQ-9 between the initial phone screening and session 4 was statistically significant (t_9=4.2; *P*=.002) and represented a clinically significant change in severity level from *moderately severe* to *mild* (>5) [25] (Table 4).

Participants in the traditional BA arm remained at an average of moderately severe (16.0, SD 3.38) between the phone intake and the beginning of session 1 (14.5, SD 3.50), and their symptoms of depression decreased to moderate (10.7, SD 4.63) by session 4. This average decrease of 5.3 points on the PHQ-8 and PHQ-9 between the initial phone screening and session 4 was also statistically significant (t_{10}=2.88; *P*=.02) and represented a change in clinical severity from *moderately severe* to *moderate* (Table 5).

There was a significant decrease in PHQ-9 scores between the phone intake and the beginning of session 1 in the XR-BA group (t_{11}=2.6; *P*=.03) but not in the traditional BA group (t_{11}=1.4; *P*=.20). These results indicate that participants in the XR-BA arm showed a significant decrease in PHQ-9 scores even before the treatment began.

To determine whether the participants showed a further statistically significant decrease in PHQ-9 scores between the beginning of sessions 1 and 4, paired-sample *t* tests were run. These results illustrated that participants in both the XR-BA (t_9=2.5; *P*=.04) and the traditional BA (t_{10}=2.3; *P*=.04) arms experienced a significant decrease in PHQ-9 scores between the start and end of the study (Tables 4 and 5).

Table 3. Independent-sample *t* test—extended reality–enhanced behavioral activation (XR-BA) versus traditional behavioral activation (BA) and test of significance of the difference between the 2 groups.

	XR-BA (n=10), mean (SD)	Traditional BA (n=11), mean (SD)	*t* test (*df*)	2-sided *P* value
Difference between PHQ-8[a] at intake and PHQ-9[b] at session 4	7.40 (5.54)	5.27 (6.07)	−0.84 (19)	.41
Difference between PHQ-9 at session 1 and PHQ-9 at session 4	4.40 (5.66)	3.73 (5.37)	−0.28 (19)	.78

[a]PHQ-8: Patient Health Questionnaire–8.

[b]PHQ-9: Patient Health Questionnaire–9.

Table 4. Paired-sample *t* test (2-tailed)—testing the significance of extended reality–enhanced behavioral activation participants' average Patient Health Questionnaire scores between various time points (n=13).

	Pretest assessment		Posttest assessment		*t* test (*df*)	2-sided *P* value
	Participants, n (%)	Values, mean (SD)	Participants, n (%)	Values, mean (SD)		
PHQ-8[a] at phone intake and PHQ-9[b] at session 4	10 (77)	15.80 (2.86)	10 (77)	8.40 (3.72)	4.22 (9)	.002
PHQ-8 at phone intake and PHQ-9 at session 1	12 (92)	15.67 (2.90)	12 (92)	13.25 (3.82)	2.59 (11)	.03
PHQ-9 at session 1 and PHQ-9 at session 4	10 (77)	12.80 (3.46)	10 (77)	8.40 (3.72)	2.46 (9)	.04
PHQ-9 at session 1 and PHQ-9 at session 2	11 (85)	13.45 (3.93)	11 (85)	11.55 (5.03)	1.70 (10)	.12
PHQ-9 at session 2 and PHQ-9 at session 3	10 (77)	10.60 (4.14)	10 (77)	10.50 (4.81)	0.07 (9)	.95
PHQ-9 at session 3 and PHQ-9 at session 4	10 (77)	10.50 (4.81)	10 (77)	8.40 (3.72)	1.41 (9)	.19

[a]PHQ-8: Patient Health Questionnaire–8.

[b]PHQ-9: Patient Health Questionnaire–9.

Table 5. Paired-sample *t* test (2-tailed)—testing the significance of traditional behavioral activation participants' average Patient Health Questionnaire scores between various time points (n=13).

	Pretest assessment		Posttest assessment		*t* test (*df*)	2-sided *P* value
	Participants, n (%)	Values, mean (SD)	Participants, n (%)	Values, mean (SD)		
PHQ-8[a] at phone intake and PHQ-9[b] at session 4	11 (85)	16.00 (3.38)	11 (85)	10.73 (4.63)	2.88 (10)	.02
PHQ-8 at phone intake and PHQ-9 at session 1	12 (92)	16.00 (3.22)	12 (92)	14.75 (3.49)	1.37 (11)	.20
PHQ-9 at session 1 and PHQ-9 at session 4	11 (85)	14.45 (3.50)	11 (85)	10.73 (4.63)	2.30 (10)	.04
PHQ-9 at session 1 and PHQ-9 at session 2	11 (85)	14.45 (3.50)	11 (85)	12.09 (3.59)	2.95 (10)	.01
PHQ-9 at session 2 and PHQ-9 at session 3	11 (85)	12.09 (3.59)	11 (85)	11.82 (2.64)	0.25 (10)	.81
PHQ-9 at session 3 and PHQ-9 at session 4	11 (85)	11.82 (2.64)	11 (85)	10.73 (4.63)	1.17 (10)	.27

[a]PHQ-8: Patient Health Questionnaire–8.

[b]PHQ-9: Patient Health Questionnaire–9.

Discussion

Feasibility

The results of this study provide evidence that XR-BA in this MDD telehealth treatment setting was a safe, feasible, tolerable, and acceptable modification to a brief BA protocol. The attrition rate of 23% (3/13) of the participants in the XR-BA arm of the study is comparable with that of other VR studies [26,27], lower than that of many RCTs of internet-based interventions for depression [28], and lower than that of a small-sample pilot RCT exploring exercise as a treatment for depression [29]. Importantly, no participant in the XR-BA treatment arm dropped out of the study because of serious adverse events, and no serious adverse events were reported throughout the study.

While participants in the previous VR-BA study completed on average more VR activities than recommended, the participants in this study did not meet their total recommended headset use of ≥4 activities each week. This was due to many participants reporting that the headset was difficult to use and it feeling like an overwhelming task to learn. Participants remarked that they would have used the headset more often if they had increased familiarity. In this vein, participants reported that the headset became more enjoyable and useful over time, which aligns with research that states that the easier to use the device, the more acceptable it is to users [30]. When working with people unfamiliar with VR, future prototypes of VR-BA may want to opt for designs that allow for simplicity, preloaded experiences, decreased choices, and rapid onboarding skill acquisition.

Participants noted several barriers and XR challenges that may have impacted their attempts to use this modality for BA and mood improvement. The learning curve for using the headset device was surprisingly burdensome. Our previous VR-BA prototype chose a simpler, less immersive headset preloaded with activity choices, but this study chose to use a more immersive and interactive headset with higher quality and range of choices of pleasant and mastery activities. This increase in variety and autonomy to simulate traditional BA came with an increasing cost to the user, with each novel experience entailing unique and new technical XR challenges and requiring new skills. This observation aligns with research indicating that it is important to learn *how* to use VR before learning *in* VR [31]. Spending time teaching participants how to use the XR headset was contraindicated in a research study due to the creation of a confound when compared to traditional BA. Yet, in practice and outside of clinical trials, it may be necessary to do so at this time when technical onboarding to commercial headsets is still complex and challenging for the average person. However, the challenges of onboarding may just as likely provide an opportunity to engage in a mastery or pleasant activity when struggling and finally gaining access to the XR headset and may actually attenuate BA with the focus on this acquisition of onboarding skills.

One participant noted that the ability to choose any activity on the headset led to "decision paralysis," an interesting juxtaposition to the previous study, which had a limited selection of 37 preselected videos and where feedback stated a desire to have more activity options. While activity ideas were provided, when using XR for activity engagement, it may be helpful to provide an even more detailed database of activity options similar to the list of adult pleasant activities [32].

Considering participant feedback from the previous study that noted that the requirement to complete a post-VR questionnaire after each use was a hindrance and burden, this study only asked participants to complete 1 post-XR questionnaire a week. While only 20% (1/5) of the participants in the previous study completed a post-VR questionnaire for each VR activity, all participants in this study (13/13, 100%) completed a post-XR questionnaire during the weeks in which they used the device. Participants in this study subsequently did not comment on the administrative burden of completing the post-XR questionnaire; however, they did acknowledge that having all the tracking and scheduling accessible via the web or through an app would make it more convenient for them to remember and complete all the required tasks.

Participants in this study rated their presence as higher on average than participants in the previous study, a finding that is consistent with research suggesting that achieving a strong sense of presence is more influenced by interactivity than by realism [33]. Participants noted feeling so present while using the headset that they made comments such as the following: "[it was] good to be able to go elsewhere [in VR] since I don't have a car," "it is nice to be able to take a break from my kids and be present at home, but not be," and "I was so immersed in the VR that I lost track of time." In addition, presence ratings increased week to week on average, consistent with participant reports that the more familiar they became with the device, the more immersed they felt.

The acceptability ratings in this study were comparatively lower than those recorded for the device used in the previous study. Nevertheless, a noteworthy observation from this study is that acceptance levels in the domains of *Perceived usefulness* and *Attitudes toward use* exhibited an average increase between the conclusion of week 1 and the beginning of session 2 and the conclusion of week 3 and the beginning of session 4. It would be intriguing to extend the study timeline and ascertain whether this trend of escalating acceptance continues, potentially surpassing the ratings for the simpler headset. It would be equally fascinating to explore whether the gradual rise in acceptance over time corresponds to more substantial improvements in mood over the same period. This is particularly relevant considering that some participants mentioned that they would have used the device more frequently if they had not perceived the learning curve as a hindrance. Furthermore, participants qualitatively indicated that the *Intention to use the technology* rating was lower given the cost and lack of affordability of the Meta Quest 2 headset.

The participants rated the protocol as largely physically tolerable, and no participants dropped out because of adverse effects. While the ratings of physical tolerability were the same (92%-93%) between the 2 studies, the participants in this study qualitatively reported more simulator sickness. Participants particularly noted that they found the Meta Quest 2 headset itself to be "heavy" and "uncomfortable" on their faces. In addition, consistent with the research on simulator sickness, participants noted that they experienced more symptoms of simulator sickness while partaking in activities with a faster-moving image compared to those with a slower-moving image [34,35]. However, also aligned with previous research, participants quantitatively and qualitatively reported a habituation effect where their simulator sickness symptoms largely decreased over time [35]. All participants reported that their symptoms were quickly resolved upon removal of the headset and did not persist.

While this study expanded upon the previous study by increasing the sample size and using a more immersive, interactive headset that offered a wider range of activity options, it would be interesting to conduct a similar study that also uses a mobile app to decrease the administrative burden for providers and patients and streamline the homework process. It is postulated that the focus on homework in BA is essential for successful treatment outcomes. Research has demonstrated that homework completion is significantly related to a decrease in symptoms [36]. Specifically, the behavioral task of completing pleasant activities contributed most strongly to decreasing symptoms of depression [36]. Hence, addressing barriers to completing homework tasks is pivotal for optimizing treatment results.

Given that participants in the previous study noted decreased headset use due to administrative constraints and that participants in this study independently expressed the value of a tracking and reminder app for homework compliance, the next crucial phase involves evaluating whether implementing a mobile app that consolidates scheduling and activity tracking can enhance homework completion rates. This, in turn, could potentially lead to more accurate and consistent homework adherence, thereby further reducing depressive symptoms and enhancing mood, ultimately maximizing the effectiveness of treatment outcomes.

Finally, this study solely used XR as a method of engaging in BA. In subsequent studies, it would be interesting to conduct the therapy in VR for both arms rather than over Zoom. Given that telehealth has become increasingly popular, it would be fascinating to note the feasibility, acceptability, and tolerability of conducting the entire session in VR. It would also be interesting to measure how this may affect the effectiveness of the intervention.

Efficacy

The XR-BA protocol was found to be noninferior to a brief BA protocol for MDD in this small study. Participants in both the traditional and XR-BA arms experienced a statistically significant reduction in depression symptoms between the initial phone screening and session 4 and between sessions 1 and 4, as well as significant reductions in clinical severity between the initial phone screening and session 4.

Only the XR-BA arm showed a statistically significant decrease in symptoms between the phone screening and session 1. Given the unblinded nature of this study, this is not surprising. These results may indicate that participants in the XR-BA arm had an enhanced expectancy or placebo effect due to the novelty or implicit beliefs surrounding technology and mental health treatment. The novelty of the treatment and anticipation that it would be helpful may have led to increased levels of hope and a decrease in depressive symptoms [37,38]. It was observed that the participants who learned that they were randomized into the XR-BA arm expressed more excitement than those who were randomized into the traditional BA arm.

While our previous study suggested the possibility of a greater reduction in symptoms of depression among participants in the VR-BA arm compared to the traditional BA arm, this study did not demonstrate any such superiority as symptom reduction was not statistically or clinically different between the groups. The noninferiority of XR-BA may be attributed to both the positives and negatives of using VR, as noted by participants. Similar to the previous study, participants shared that they found VR to be “novel,” using VR showed them that they could enjoy activities again, and VR inspired them to engage in real-life activities. The latter fact was true among both participants who found VR to be a positive experience (ie, watching a YouTube 360° video of a beach inspired them to visit the beach in person) and a participant who did not enjoy VR because of preferences for tactile experiences and consequently made an increased effort to go outside to feel the sun on their skin. Many participants also noted that the XR-BA helped improve their mood by taking them to a new place in an immersive way, thereby increasing their attention and decreasing distraction, allowing for a fully mindful experience in the present moment.

The finding that XR-BA was as efficacious in reducing symptoms of MDD as a brief traditional BA protocol is critical. Patients can use VR to improve their mood if they encounter barriers to engaging in activities IRL. Participants commented that “VR is easier and more convenient than having to go places,” “I have been able to visit a few places I have always wanted to travel, so I noticed being so absorbed [by the places],” “VR has a larger realm of possibilities. In the real world I need to check hours [that events are occurring/open] and the weather,” and “I would recommend [using VR] to a friend if they didn’t want to do therapy,” which qualitatively supports the notion that VR can help decrease barriers to in-person activity engagement. These statements further corroborate the previous study’s suggestion that clinicians may be justified in using VR as a first step in BA for patients who may not have access, motivation, or desire to engage in activities IRL.

Limitations

This study aimed to recruit and enroll 40 participants, and recruitment took place remotely via Zoom-delivered telehealth sessions between December 19, 2022, and July 24, 2023. The study ended recruitment in July 2023 given that the necessary number of participants to yield a powered result had been enrolled. Although many of the enumerated findings are promising, this study has several limitations. First, the quantitative and qualitative measures were subjective and completed by the participants. Participants in both the XR-BA and traditional BA arms self-reported their completed activity and mood scores, which may have introduced inaccurate reporting. Specifically, there were no objective measures used to evaluate XR experiences, which resulted in participant self-report of the activities chosen and length of time in XR, which many participants did not document in the moment, leading to potentially inaccurate reporting. In addition, although the PHQ-9 is a standard self-report measure, the questions were read aloud for participants to answer rather than being delivered in a standard written format. This method may have resulted in less accurate reporting if the participants felt inclined to respond in a certain way.

Furthermore, as there were no official follow-ups, it is unknown whether the mood gains that the participants reported were long-lasting. This study had a relatively short duration. As mentioned previously, participants remarked on the learning curve of the headset, and both qualitative and quantitative data illustrated that the headset became more acceptable and tolerable each week. Thus, in a longer trial, participants may experience greater mood gains as they become more familiar with the headset. In addition, a participant in the XR-BA arm of the study was unable to use the headset between sessions 2 and 3 owing to both a heavy work week and being physically ill. This participant expressed sadness about this outcome and a desire to expand the study timeline to have more time with the headset. Furthermore, the study’s short duration may have led to mood changes due to factors external to the study, such as a relatively heavy or light work week or an illness. Finally, many participants expressed that there were few free trials or options within the XR headset. Some participants reported that they would be more willing to purchase activities if they were able to keep the headset or if the study were longer so that they had more time with their purchase. Overall, participants in both study arms expressed a desire to lengthen the study timeline and noted that the 3-week, 4-session protocol felt too short.

Another limitation was the nonblinded nature of the study. Participants randomized into the XR-BA arm expressed greater excitement than those randomized into the traditional BA arm, which may have led to an initially greater decrease in depressive symptoms on the former. Further studies should invest in a system of double blinding to confirm these results.

Finally, as in our previous study, recruitment was a large obstacle. Although the goal was to randomize 40 participants with MDD into either of the study arms, only 26 participants were randomized because other potential participants were excluded based on ineligibility, declining to participate, or being lost to follow-up. It is notable that other VR and depression studies have had similar or smaller sample sizes [39,40]. It is important to recognize that this could underscore an inherent challenge in depression studies, where health state and conditional altruism are large contributing factors to participation interest [41]. In addition, the small sample size hindered our ability to address other interesting research questions, such as whether different subtypes or severity of MDD would yield different effects from the treatment. For example, would individuals with more severe MDD respond better or worse to XR-BA than those with a milder case compared to traditional BA? These are questions that would need to be answered in a study with a larger sample size. Moreover, given the diverse nature of the disorder, the findings might not universally apply to all those dealing with symptoms of depression.

Conclusions

The findings of this study support our previous report that using XR as a substitute for IRL pleasant and mastery activities in a brief BA protocol for individuals diagnosed with MDD is feasible, acceptable, and tolerable. This remained true even when using a more difficult and interactive headset that posed technical and physical challenges.

This study also expanded on our feasibility trial to perform the first known efficacy trial of XR-BA. This study demonstrated that XR-BA may not be inferior to traditional BA as it was equally and statistically efficacious in improving symptoms of depression in an MDD sample as measured using the PHQ-9. It also suggested that XR-BA may have enhanced the placebo or expectation effects of BA treatment.

The results of this study demonstrate that it may not be unreasonable for clinicians to suggest the use of VR-simulated pleasant activities to patients when delivering BA as VR-simulated pleasant activities may offer solutions to some of the common problems and barriers encountered when using BA. When deciding on a clinical approach, professionals may need to weigh the advantages and disadvantages of using simpler versus complex headsets. Given that this study and the previous VR-BA study both illustrate clinical effectiveness and that the feedback on the previous study's preloaded headset was more favorable compared with this study's software-agnostic Meta Quest 2 headset, it could be concluded that a simpler device would be preferred by patients at this time. Furthermore, despite this study providing the opportunity for participants to choose among a multitude of VR options, the most commonly chosen VR activity was YouTube 360° videos largely given its simplicity and lack of cost. Regardless of the hardware or software specifications, this study supports the notion that the use of VR may enhance mood in those living with MDD when used in conjunction with individual therapy delivering BA principles and protocols. More research on the implementation of such an approach is needed to understand how to most effectively leverage this technology in depressive disorders.

Future and more extensive controlled studies may want to explore further whether XR can increase expectation or placebo effects during MDD treatments or have other enhancing qualities to the delivery of BA for MDD.

Acknowledgments

This study was partially funded by the 2022 Wu Tsai Neurosciences Institute and Stanford Byers Center for Biodesign Neuroscience:Translate award.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Telephone screening questions.
[DOCX File , 13 KB - mental_v11i1e52326_app1.docx]

Multimedia Appendix 2
Consent form.
[DOCX File , 60 KB - mental_v11i1e52326_app2.docx]

Multimedia Appendix 3
Demographic questionnaire.
[DOCX File , 17 KB - mental_v11i1e52326_app3.docx]

Multimedia Appendix 4
Extended reality activity list.

[DOCX File , 19 KB - mental_v11i1e52326_app4.docx]

Multimedia Appendix 5
Post–extended reality questionnaire.
[DOCX File , 20 KB - mental_v11i1e52326_app5.docx]

Multimedia Appendix 6
CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) V 1.6.1 checklist.
[PDF File (Adobe PDF File), 3474 KB - mental_v11i1e52326_app6.pdf]

References

1. Depression and other common mental disorders: global health estimates. World Health Organization. 2017. URL: https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf [accessed 2023-08-12]
2. Greenberg PE, Fournier AA, Sisitsky T, Simes M, Berman R, Koenigsberg SH, et al. The economic burden of adults with major depressive disorder in the United States (2010 and 2018). Pharmacoeconomics 2021 Jun;39(6):653-665 [FREE Full text] [doi: 10.1007/s40273-021-01019-4] [Medline: 33950419]
3. Clinical practice review for major depressive disorder. Anxiety & Depression Association of America. URL: https://adaa.org/resources-professionals/practice-guidelines-mdd [accessed 2023-08-12]
4. Dimidjian S, Barrera MJ, Martell C, Muñoz RF, Lewinsohn PM. The origins and current status of behavioral activation treatments for depression. Annu Rev Clin Psychol 2011;7:1-38 [FREE Full text] [doi: 10.1146/annurev-clinpsy-032210-104535] [Medline: 21275642]
5. Collins KA, Westra HA, Dozois DJ, Burns DD. Gaps in accessing treatment for anxiety and depression: challenges for the delivery of care. Clin Psychol Rev 2004 Sep;24(5):583-616 [FREE Full text] [doi: 10.1016/j.cpr.2004.06.001] [Medline: 15325746]
6. Wang PS, Lane M, Olfson M, Pincus HA, Wells KB, Kessler RC. Twelve-month use of mental health services in the United States: results from the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005 Jun;62(6):629-640 [FREE Full text] [doi: 10.1001/archpsyc.62.6.629] [Medline: 15939840]
7. Lin Y. 10 virtual reality statistics every marketer should know in 2023 [infographic]. Oberlo. 2023 Mar 17. URL: https://www.oberlo.com/blog/virtual-reality-statistics [accessed 2023-08-12]
8. Park M, Kim DJ, Lee U, Na EJ, Jeon HJ. A literature overview of virtual reality (VR) in treatment of psychiatric disorders: recent advances and limitations. Front Psychiatry 2019 Jul 19;10:505 [FREE Full text] [doi: 10.3389/fpsyt.2019.00505] [Medline: 31379623]
9. Baghaei N, Chitale V, Hlasnik A, Stemmet L, Liang HN, Porter R. Virtual reality for supporting the treatment of depression and anxiety: scoping review. JMIR Ment Health 2021 Sep 23;8(9):e29681 [FREE Full text] [doi: 10.2196/29681] [Medline: 34554097]
10. Paul M, Bullock K, Bailenson J. Virtual reality behavioral activation for adults with major depressive disorder: feasibility randomized controlled trial. JMIR Ment Health 2022 May 06;9(5):e35526 [FREE Full text] [doi: 10.2196/35526] [Medline: 35404830]
11. Birckhead B, Khalil C, Liu X, Conovitz S, Rizzo A, Danovitch I, et al. Recommendations for methodology of virtual reality clinical trials in health care by an international working group: iterative study. JMIR Ment Health 2019 Jan 31;6(1):e11973 [FREE Full text] [doi: 10.2196/11973] [Medline: 30702436]
12. Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr;114(1-3):163-173 [FREE Full text] [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
13. Sheehan DV, Lecrubier Y, Harnett Sheehan K, Janavs J, Weiller E, Keskiner A, et al. The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability. Eur Psychiatry 1997;12(5):232-241 [FREE Full text] [doi: 10.1016/s0924-9338(97)83297-x]
14. Paul M, Bullock K, Bailenson J. Virtual reality behavioral activation as an intervention for major depressive disorder: case report. JMIR Ment Health 2020 Nov 03;7(11):e24331 [FREE Full text] [doi: 10.2196/24331] [Medline: 33031046]
15. Jacobson NS, Martell CR, Dimidjian S. Behavioral activation treatment for depression: returning to contextual roots. Clin Psychol Sci Pract 2001;8(3):255-270 [FREE Full text] [doi: 10.1093/clipsy.8.3.255]
16. Funderburk JS, Pigeon WR, Shepardson RL, Maisto SA. Brief behavioral activation intervention for depressive symptoms: patient satisfaction, acceptability, engagement, and treatment response. Psychol Serv 2020 Nov;17(4):443-451 [FREE Full text] [doi: 10.1037/ser0000328] [Medline: 30714752]
17. Meta quest 2: our most advanced new all-in-one VR headset. Meta. URL: https://www.meta.com/quest/products/quest-2/tech-specs/#tech-specs [accessed 2023-08-13]

18. MacPhillamy DJ, Lewinsohn PM. The pleasant events schedule: studies on reliability, validity, and scale intercorrelation. J Consult Clin Psychol 1982;50(3):363-380 [FREE Full text] [doi: 10.1037//0022-006x.50.3.363]
19. Marcoulides GA, Schumacker RE. Full information estimation in the presence of incomplete data. In: Advanced Structural Equation Modeling. London, UK: Psychology Press; 1996.
20. Wothke W. Longitudinal and multigroup modeling with missing data. In: Little TD, Schnabel KU, Baumert J, editors. Modeling longitudinal and multilevel data: practical issues, applied approaches, and specific examples. Mahwah, NJ: Lawrence Erlbaum Associates Publishers; 2000.
21. Burns DD, Spangler DL. Do changes in dysfunctional attitudes mediate changes in depression and anxiety in cognitive behavioral therapy? Behav Ther 2001;32(2):337-369 [FREE Full text] [doi: 10.1016/s0005-7894(01)80008-3]
22. Beran TN, Violato C. Structural equation modeling in medical research: a primer. BMC Res Notes 2010 Oct 22;3:267 [FREE Full text] [doi: 10.1186/1756-0500-3-267] [Medline: 20969789]
23. Barrett P. Structural equation modelling: adjudging model fit. Pers Individ Dif 2007 May;42(5):815-824 [FREE Full text] [doi: 10.1016/j.paid.2006.09.018]
24. Grimm KJ, Ram N, Estabrook R. Growth Modeling: Structural Equation and Multilevel Modeling Approaches. New York, NY: Guilford Publications; 2016.
25. Kroenke K, Spitzer RL. The PHQ-9: a new depression diagnostic and severity measure. Psychiatr Ann 2002 Sep;32(9):509-515 [FREE Full text] [doi: 10.3928/0048-5713-20020901-06]
26. Bouchard S, Dumoulin S, Robillard G, Guitard T, Klinger E, Forget H, et al. Virtual reality compared with exposure in the treatment of social anxiety disorder: a three-arm randomised controlled trial. Br J Psychiatry 2017 Apr;210(4):276-283 [FREE Full text] [doi: 10.1192/bjp.bp.116.184234] [Medline: 27979818]
27. du Sert OP, Potvin S, Lipp O, Dellazizzo L, Laurelli M, Breton R, et al. Virtual reality therapy for refractory auditory verbal hallucinations in schizophrenia: a pilot clinical trial. Schizophr Res 2018 Jul;197:176-181 [FREE Full text] [doi: 10.1016/j.schres.2018.02.031] [Medline: 29486956]
28. Christensen H, Griffiths KM, Farrer L. Adherence in internet interventions for anxiety and depression. J Med Internet Res 2009 Apr 24;11(2):e13 [FREE Full text] [doi: 10.2196/jmir.1194] [Medline: 19403466]
29. Foley LS, Prapavessis H, Osuch EA, De Pace JA, Murphy BA, Podolinsky NJ. An examination of potential mechanisms for exercise as a treatment for depression: a pilot study. Ment Health Phys Act 2008 Dec;1(2):69-73 [FREE Full text] [doi: 10.1016/j.mhpa.2008.07.001]
30. Manis KT, Choi D. The virtual reality hardware acceptance model (VR-HAM): extending and individuating the Technology Acceptance Model (TAM) for virtual reality hardware. J Bus Res 2019 Jul;100:503-513 [FREE Full text] [doi: 10.1016/j.jbusres.2018.10.021]
31. Han E, Nowak KL, Bailenson JN. Prerequisites for learning in networked immersive virtual reality. Technol Mind Behav 2022 Nov 15;3(4) [FREE Full text] [doi: 10.1037/tmb0000094]
32. Mado M, Fauville G, Jun H, Most E, Strang C, Bailenson JN. Accessibility of educational virtual reality for children during the COVID-19 pandemic. Technol Mind Behav 2022 Mar 15;3(1) [FREE Full text] [doi: 10.1037/tmb0000066]
33. IJsselsteijn WA. Elements of a multi-level theory of presence: phenomenology, mental processing and neural correlates. In: Proceedings of PRESENCE 2002. 2002 Presented at: PRESENCE 2002; October 9-11, 2002; Porto, Portugal URL: https://research.tue.nl/nl/publications/elements-of-a-multi-level-theory-of-presence-phenomenology-mental
34. Wilson JR. Effects of participating in virtual environments: a review of current knowledge. Safety Sci 1996;23(1):39-51. [doi: 10.1016/0925-7535(96)00026-4]
35. Davis S, Nesbitt K, Nalivaiko E. A systematic review of cybersickness. In: Proceedings of the 2014 Conference on Interactive Entertainment. 2014 Presented at: IE2014; December 2-3, 2014; Newcastle, Australia. [doi: 10.1145/2677758.2677780]
36. Rees CS, McEvoy P, Nathan PR. Relationship between homework completion and outcome in cognitive behaviour therapy. Cogn Behav Ther 2005;34(4):242-247. [doi: 10.1080/16506070510011548] [Medline: 16319035]
37. Rutherford BR, Wager TD, Roose SP. Expectancy and the treatment of depression: a review of experimental methodology and effects on patient outcome. Curr Psychiatry Rev 2010 Feb 01;6(1):1-10 [FREE Full text] [doi: 10.2174/157340010790596571] [Medline: 24812548]
38. Ritschel L, Cassiello-Robbins C. Hope and depression and personality disorders. Curr Opin Psychol 2023 Feb;49:101507 [FREE Full text] [doi: 10.1016/j.copsyc.2022.101507] [Medline: 36459930]
39. Stamou G, Garcia-Palacios A, Botella C. The combination of cognitive-behavioural therapy with virtual reality for the treatment of post-natal depression. In: Proceedings of the 31st Australian Conference on Human-Computer-Interaction. 2019 Presented at: OzCHI '19; December 2-5, 2019; Fremantle, Australia. [doi: 10.1145/3369457.3369541]
40. Suwanjatuporn A, Chintakovid T. Using a virtual reality system to improve quality of life of the elderly people with depression. In: Proceedings of the IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia). 2019 Presented at: ICCE-Asia; June 12-14, 2019; Bangkok, Thailand URL: https://ieeexplore.ieee.org/document/8941607 [doi: 10.1109/icce-asia46551.2019.8941607]
41. Hughes-Morley A, Young B, Waheed W, Small N, Bower P. Factors affecting recruitment into depression trials: systematic review, meta-synthesis and conceptual framework. J Affect Disord 2015 Feb 01;172:274-290 [FREE Full text] [doi: 10.1016/j.jad.2014.10.005] [Medline: 25451427]

Abbreviations

AMOS: analysis of moment structures
BA: behavioral activation
CONSORT: Consolidated Standards of Reporting Trials
EBP: evidence-based psychotherapy
IRL: in real life
ITT: intention-to-treat
MDD: major depressive disorder
PHQ-8: Patient Health Questionnaire–8
PHQ-9: Patient Health Questionnaire–9
RCT: randomized controlled trial
SEM: structural equation modeling
SSQ: Simulator Sickness Questionnaire
VR: virtual reality
VR-BA: virtual reality–enhanced behavioral activation
XR: extended reality
XR-BA: extended reality–enhanced behavioral activation

Edited by J Torous; submitted 30.08.23; peer-reviewed by W Wei, T Ong; comments to author 23.11.23; revised version received 16.02.24; accepted 01.03.24; published 15.04.24.

Please cite as:
Paul M, Bullock K, Bailenson J, Burns D
Examining the Efficacy of Extended Reality–Enhanced Behavioral Activation for Adults With Major Depressive Disorder: Randomized Controlled Trial
JMIR Ment Health 2024;11:e52326
URL: https://mental.jmir.org/2024/1/e52326
doi:10.2196/52326
PMID:38437873

Original Paper

Time-Varying Network Models for the Temporal Dynamics of Depressive Symptomatology in Patients With Depressive Disorders: Secondary Analysis of Longitudinal Observational Data

Björn Sebastian Siepe[1], MSc; Christian Sander[2,3], PhD; Martin Schultze[4], Prof Dr; Andreas Kliem[5], PhD; Sascha Ludwig[6], MSc; Ulrich Hegerl[7,8], Prof Dr; Hanna Reich[2,8], PhD

[1]Psychological Methods Lab, Department of Psychology, University of Marburg, Marburg, Germany
[2]German Depression Foundation, Leipzig, Germany
[3]Department of Psychiatry and Psychotherapy, University of Leipzig Medical Center, Leipzig, Germany
[4]Department of Psychology, Goethe University, Frankfurt, Germany
[5]Adesso SE, Dortmund, Germany
[6]Institute for Applied Informatics, University Leipzig, Leipzig, Germany
[7]Department for Psychiatry, Psychosomatics and Psychotherapy, Goethe University, Frankfurt, Germany
[8]Depression Research Center of the German Depression Foundation, Department for Psychiatry, Psychosomatics and Psychotherapy, Goethe University, Frankfurt, Germany

Corresponding Author:
Björn Sebastian Siepe, MSc
Psychological Methods Lab
Department of Psychology
University of Marburg
Gutenbergstraße 18
Marburg, 35032
Germany
Phone: 49 6421 28 23616
Email: bjoern.siepe@uni-marburg.de

Abstract

Background: As depression is highly heterogenous, an increasing number of studies investigate person-specific associations of depressive symptoms in longitudinal data. However, most studies in this area of research conceptualize symptom interrelations to be static and time invariant, which may lead to important temporal features of the disorder being missed.

Objective: To reveal the dynamic nature of depression, we aimed to use a recently developed technique to investigate whether and how associations among depressive symptoms change over time.

Methods: Using daily data (mean length 274, SD 82 d) of 20 participants with depression, we modeled idiographic associations among depressive symptoms, rumination, sleep, and quantity and quality of social contacts as dynamic networks using time-varying vector autoregressive models.

Results: The resulting models showed marked interindividual and intraindividual differences. For some participants, associations among variables changed in the span of some weeks, whereas they stayed stable over months for others. Our results further indicated nonstationarity in all participants.

Conclusions: Idiographic symptom networks can provide insights into the temporal course of mental disorders and open new avenues of research for the study of the development and stability of psychopathological processes.

(JMIR Ment Health 2024;11:e50136) doi:10.2196/50136

KEYWORDS

depression; time series analysis; network analysis; experience sampling; idiography; time varying; mobile phone

Introduction

Background

Different lines of research have established the heterogeneous nature of the etiology, clinical presentation, and treatment outcomes of depression [1-4], thus demonstrating a need for new ways to conceptualize and investigate the disorder. This is indicative of a broader issue across specific psychiatric diagnoses. The widespread evidence of substantial heterogeneity within diagnostic labels has increased the awareness of the need for more individualized research on mental disorders [5]. Although clinical psychology has a long tradition of interest in the individual, most studies in clinical psychology rely on nomothetic, cross-sectional data [5]. However, several theoretical arguments [6-8] and empirical studies [9] have shown that findings generated on a between-person basis are often not applicable to within-person processes, calling into question the extent to which cross-sectional studies are relevant for the understanding of individual clinical cases. Recently, the emergence of new theoretical approaches [10], statistical methods [11], and options for the collection of longitudinal data [12] have led to a surge in empirical studies of within-person, idiographic processes in clinical psychology [5].

Idiographic modeling of psychopathology has several possible advantages compared to group-level models. Owing to its potential to provide insights into temporal processes, an idiographic approach using longitudinal data could inform clinicians and researchers about the dynamics of psychological processes *of an individual* in a specific context [13], which is closely linked to clinical practice [14]. In combination with a focus on experiences in everyday life, idiographic models could lead to an improved understanding of mechanisms that influence the development and trajectory of mental disorders. During treatment, idiographic models could potentially be used to provide data-informed feedback to patients and therapists [13], develop personalized psychotherapy interventions [15], or design mobile interventions that are tailored to the individual [16].

To study such within-person processes, researchers use experience sampling methods to collect many observations per individual over time, also known as intensive longitudinal data [17]. Then, various forms of time series models can be used to investigate the relationships among multiple variables across time. Results of these models are often depicted as networks of variables that interact with one another. Then, these can be interpreted in accordance with the network approach of psychopathology that conceptualizes disorders as causal systems of mutually interacting symptoms [11,18]. Networks based on cross-sectional data have become very popular in the past few years, particularly in research on depression [19]. However, the so-called dynamic networks based on longitudinal data are especially promising for the network approach, as they both potentially allow insights into how disorders emerge from the interplay among individual symptoms over time and because they can reveal individual differences in symptom associations. Idiographic network models have, for example, been used to explore individual symptom patterns in different psychiatric disorders [20,21], including depression [22], to investigate psychotherapy processes [23] or to identify individualized treatment targets in eating disorders [24].

However, due to implicit assumptions of commonly used statistical methods, these patterns are typically modeled as static over time [25]. This approach restricts the investigation of change processes that can be of central interest to researchers and clinicians [14,26]. Therefore, most previous studies have not been able to investigate whether and how individual symptom networks change over time (for an early exception, refer to the study by Wichers and Groot [27]). For example, psychotherapists who are using daily diary data of their patients might be interested to examine whether the association between stress at work and subsequent depressive mood changes during therapy, as their patients might be incrementally able to handle stress better. This development would be difficult to account for when using typical models. In addition to this issue, experience sampling studies are often interested in variables that may change very fast, such as perceived stress or mood, and thus commonly follow individuals for a short time, often 1 or 2 weeks [28]. This study duration may be inappropriate for constructs such as depressive disorders, as it might miss slow changes developing over a longer time span, therefore incorrectly assuming that associations among symptoms are stable over time. We aimed to investigate the possibilities of circumventing these limitations by using recently developed methods for the estimation of time-varying models for psychological applications [25,29]. So far, this approach has not been used with time series data of multiple clinical cases.

Objective

In this study, we applied time-varying network modeling to daily self-report data of patients diagnosed with recurrent depressive disorder to explore the idiographic course of depression over several months and to gain insight into the stability or instability of individual symptom networks of depression. In addition to 2 daily depressive core symptoms, namely anhedonia and feeling down, we included daily summaries of sleep duration, rumination, and the quality and quantity of social interactions as all these aspects have been hypothesized to interact in depression. We chose these items to gain multifaceted insight into the course of depression while limiting ourselves to a few variables for the demonstration and application of the chosen modeling technique. An in-depth theoretical background discussion regarding the selection of variables is provided in Multimedia Appendix 1 [25,29-58].

To justify the necessity of time-varying modeling, which is a more complex and data-intensive approach, we first tested whether the data-generating process of each individual in the time frame of our study was time varying by using a recently proposed hypothesis test [25]. Given the nonstationarity of participants' time series, there were 2 main exploratory goals for this study. First, we aimed to construct individual networks for every patient to model the temporal associations among all variables on a day-to-day basis. Second, we wanted to explore the temporal dynamics of the individual course of depression by investigating changes in the network structure over time. To evaluate the initial indicators of trustworthiness of these models for use in the assessment of individual patients, we further aimed to assess the stability of estimates, prediction errors, and variance explained. The results of this study provide new insights into the time-varying nature of depression and highlight the usefulness and limitations of new statistical approaches to capture these temporal dynamics.

Methods

Transparency and Openness

This was, in part, a secondary analysis of the data previously analyzed by Lorenz et al [30] on the idiographic association between sleep and depression. Our analyses were preregistered after data collection and before secondary data analysis using the template for preregistration of experience sampling studies [59]. The preregistration and all code for the analyses can be accessed through Open Science Framework [60]. All deviations from the preregistered protocol are explained in detail in Multimedia Appendix 1.

Ethical Considerations

The study was approved by the ethics committee of the University of Leipzig (258/17-ek). Participants who completed the data collection process were reimbursed €250 (approximately US $280) for their efforts for each study phase, implying a maximum individual financial compensation of €750 (approximately US $840). They could also keep the mobile phone that was provided to them for the study.

Procedure

Data used in this study were collected as part of the research project, Sensor-Based System for Therapy Support and Management of Depression (STEADY). The overarching aim of the project was the creation of a sensor-based system for individuals with depression, integrating data from smartphones and wearable and stationary sensors to monitor the course of their disorder using self-assessments and physiological and behavioral markers. The data for this study were collected during a feasibility study of the STEADY system. The STEADY smartphone app was installed on a mobile phone that was provided to the participants for the completion of self-report protocols.

Data collection for the feasibility study was split into 3 consecutive study phases between 2017 and 2019, and they did not differ in their self-report protocols. An overview of all self-report questionnaires administered before, during, and after the study phases is provided in the preregistration. Participants were also given wrist-worn fitness trackers and stationary sleep sensors to collect passive sensing data. These data were not included in the analyses due to their questionable data quality. Within the smartphone app, participants were able to fill morning and evening logs. They were asked to complete them directly after waking up and shortly before going to bed, respectively. Morning logs were available from 3 AM to 3 PM, whereas evening logs were available from 3 PM to 3 AM.

Several precautionary measures were undertaken to assist in using the app and to prevent the occurrence of missing data, such as monthly visits to the study center and phone calls.

Participants

Participants were recruited in cooperation with the Department of Psychiatry and Psychotherapy (University of Leipzig Medical Center, Germany). Potential participants were informed about the study by their treating physicians. If they indicated interest, they were contacted by a staff member of the study center, who conducted the formal examination of inclusion and exclusion criteria. If the individual was found to be eligible, written consent for participation was obtained from them. During the initial diagnostic screening, invitees were inquired about sociodemographic information and their medical history in a semistructured interview. The Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders–4 [61] (in its German translation [62]) was used to assess psychiatric diagnoses. The Inventory of Depressive Symptomatology, Clinician Rated (IDS-C) [63] (in its German translation [31]) was used by trained raters to assess current depressive symptom severity.

Inclusion criteria were the following: having a diagnosis of a recurrent depressive disorder; having a current depressive symptom level of at least 14 points on the IDS-C; currently being treated professionally for depression; being aged at least 18 years; and living near the research center of the German Depression Foundation in Leipzig, Germany, to accommodate regular in-person appointments. Individuals were excluded if they had severe somatic disorders; displayed acute suicidal behavior; were pregnant or in the lactation period; had electronic implants; or experienced the following psychiatric comorbidities: borderline personality disorder, schizophrenia, alcohol or drug addiction, or schizotypal and delusional disorders. For this study, we prespecified that participants should have at least 130 days of data and <30% missingness for any variable. Of the 25 total participants, we included 20 (80%) individuals in our analyses. A detailed description of the sample is provided in the following sections.

Measures

Pre-Post Assessment of Depressive Symptom Levels

The IDS-C [63] was used to describe depressive symptom levels before and after the data collection period. It was assessed before and after daily data collection and during some of the monthly visits, but data from only 1 questionnaire each, before the start and after the end of an individual time series, was used in this paper. The IDS-C consists of 30 items that inquire about a range of depressive symptoms, of which 28 items (scored from 0-3) were included in this paper. Items of the IDS-C are weighted equally and combined into a sum score ranging from 0 to 84, where a cutoff point of 13 was originally proposed to identify individuals with symptoms [63]. The IDS-C has been evaluated psychometrically in different populations, including individuals with depression, in both its original version and its German translation [32].

Daily Diary Measures

A variety of self-report questions was used in the morning and evening protocols. As preregistered, we chose a subset of all self-report variables for the following reasons: theoretical relevance for depression, assessment on a continuous scale, sampling frequency, frequency of missingness, and a general preference for sparsity for our estimation method. In total, 6 daily items were used in this study: loss of interest or joylessness, feeling down or depressed or hopeless, rumination, quantity of social contacts, quality of social contacts, and sleep duration (refer to Multimedia Appendix 1 for the wording).

The first 2 items represent depressive symptoms resembling the items of the Patient Health Questionnaire–2 [64] and are listed as core symptoms in the *Diagnostic and Statistical Manual of Mental Disorders*, *Fifth Edition* [65]. For this study, these items were reframed to inquire about a single day. They were described as *anhedonia* for the first item and *feeling down* for the second item. The quantity and quality of social contacts were also assessed using a visual analog scale, where participants could indicate how many contacts they had and how they felt about these social contacts. All these items were assessed in the evening logs. Total sleep time was assessed in the morning logs. Participants were asked about when they went to bed and when they got up in the morning. The time spent in bed was calculated by the app, and participants were then asked how much of this time they spent sleeping. The items used in this study have not been psychometrically evaluated but were specifically created for the STEADY app. In addition to daily questions, some of the participants provided qualitative information regarding significant life events during monthly visits to the study center.

Data Preparation and Statistical Analysis

Overview

The software environment R (version 4.1.1; R Foundation for Statistical Computing) [66] was used for all analyses in the study. Version control information and the R code for all analyses can be accessed through the Open Science Framework repository connected to this project [60]. All variables in the data set were treated as continuous variables.

Missing Data

Owing to the long data collection period, several participants had many blocks of consecutive data separated by an extended period of missingness. To obtain the single longest time series for each participant, we searched for the longest phase without item-wise missing data for >7 consecutive days and discarded the remaining data. We chose a maximum window of consecutive missingness as we did not want to impute several consecutive missing data points. A detailed workflow for handling missing data is available in the preregistration. We performed item-wise missing value imputation using the Kalman filter, which has been shown to perform well in a previous simulation study of idiographic network analysis [33]. We used the Kalman filter from the R package, *imputeTS* [67], in its default setting.

Statistical Analysis

Time-Varying Vector Autoregression

Vector autoregressive (VAR) models are time series models that can be used to investigate the relationships among multiple variables at a given lag size. For example, in a VAR model of lag 1, the value of a variable at a given time point is regressed on the value of itself (known as autoregressive effect) and all other variables (known as cross-lagged effects) at the previous time point. As mentioned in the *Introduction* section, these models assume stationarity, meaning that the parameters of the model are assumed to be constant over time, which may not be appropriate for many research questions. Different techniques have been applied to explicitly account for time-varying parameters in psychological time series in the past, mostly focusing on univariate or bivariate associations [23,68-70]. Recently, a new approach for estimating time-varying VAR models based on kernel smoothing has been developed and tested in a simulation study [25,29]. Using VAR models with kernel smoothing allows the estimation of parameters that change over time and choosing between models with different degrees of flexibility to vary over time.

We used time-varying VAR models as implemented in the R package, *mgm* [29], to estimate idiographic models with a default lag size of 1 for reasons of parsimony. Further details about the method and our implementation of the model are available in Multimedia Appendix 1 and in the papers by Haslbeck et al [25] and Haslbeck and Waldorp [29]. To obtain the time-varying parameters, local VAR models are estimated at several equidistant estimation points and then combined. In these local models, observations closer to a specific estimation point are weighted more strongly than observations farther away. The kernel weighting used to achieve this is characterized by its bandwidth, which determines the number and weights of observations included in the estimation. We have described and visualized the resulting models as dynamic networks. These comprise *nodes* (representing variables) and *edges* (representing the temporal associations among variables). The networks were visualized using the R package, *qgraph* [71].

Bandwidth Selection

Bandwidth selection represents a bias-variance trade-off [72], where smaller bandwidths lead to highly local estimates and faster changes. Large bandwidths >1 lead to an estimation that is increasingly similar to the results of estimating a stationary model [25]. To select an appropriate individual bandwidth, several candidate bandwidths were compared using a time-stratified, 5-fold, cross-validation scheme. Then, we selected the bandwidth that minimized the root mean squared error (RMSE) across the test sets. Details about the bandwidth selection scheme and an exploratory analysis of the robustness of selection are available in Multimedia Appendix 1. A visual illustration of the difference between bandwidths of different sizes is provided in Multimedia Appendix 1.

Model Estimation

Using the selected bandwidth, parameter estimates for every estimation point were then obtained via regularized regression using the *least absolute shrinkage and selection operator* (lasso) [73]. The lasso is a regularization technique that shrinks parameter estimates while possibly setting some of them to 0. The choice of regularization parameters is explained in Multimedia Appendix 1. As the lasso is sensitive to different variances, we z-transformed all the variables before all the analyses. Following the simulation study by Haslbeck et al [25], our final model was estimated using 20 equally spaced estimation points. We distinguished the term estimation point from the term time point, which refers to a single daily observation.

Stability and Predictability

To gain insight into the stability of the parameter estimates, we used a block bootstrap scheme to obtain bootstrapped sampling distributions. We further computed the proportion of explained variance (R^2) and RMSE as prediction errors for each variable at each estimation point. To accomplish this, we used the weighted method of forming a prediction error as implemented in the *mgm* package. The R^2 values for all participants at all estimation points are available in Multimedia Appendix 1.

Stationarity Hypothesis Test

To test whether the data-generating process of the time series of an individual was stationary, we implemented a significance test as proposed by Haslbeck et al [25] to test the null hypothesis that the process was stationary. Details about the test are available in Multimedia Appendix 1.

Results

Sample Characteristics

The 20 included participants (n=13, 65% women) had a mean age of 44.4 (SD 11.6; range 26-67) years during screening. Data were available for 274 (SD 82.4; range 154-539) days (ie, time points) on average. In the selected time series of participants, 5.53% of the data were missing. Further information about the missing data structure is provided in Multimedia Appendix 1. Of the 20 participants, 18 (90%) were German citizens, 1 (5%) had dual citizenship, and 1 (5%) had a different nationality. Of the 20 participants, 11 (55%) had completed the general higher education entrance qualification. On average, IDS-C values decreased from 26 before the start of the time series to 22.9 afterward. Of the 20 participants, 18 (90%) took antidepressant medication and 18 (90%) currently or previously underwent psychotherapeutic treatment. More information about every participant is provided in Table 1.

Table 1. Sample characteristics.

ID[a]	Sex	Age range (y)[b]	Time series length, n	Missingness (%)[c]	IDS-C[d] score before EMA[e] data collection	IDS-C score after EMA data collection	Bandwidth
1	Female	40-45	311	12.6	24	10	0.45
2	Female	45-50	539	5.0	24	11	0.009
3	Male	50-55	308	3.7	38	30	0.12
4	Female	30-35	204	14.5	35	18	0.12
5	Female	40-45	205	3.8	19	1	0.12
6	Female	60-65	301	2.3	34	40	0.12
7	Male	40-45	309	6.9	7	2	0.23
8	Female	40-45	283	1.8	35	29	0.01
9	Female	35-40	189	2.9	19	38	0.01
10	Female	30-35	316	4.2	11	12	0.01
11	Male	40-45	204	1.0	19	9	0.12
13	Male	50-55	303	1.8	41	47	0.12
15	Male	50-55	263	8.1	28	57	0.01
16	Female	30-35	301	3.7	23	6	0.01
17	Female	30-35	154	2.9	27	21	0.12
18	Female	50-55	198	19.6	17	21	0.23
19	Female	25-30	303	2.7	20	23	0.01
20	Female	65-70	301	4.7	25	23	0.12
21	Male	25-30	194	7.2	27	25	1.00
22	Male	60-65	302	1.4	47	35	0.12

[a]Only individuals who were included in the analyses for this study.

[b]We provided age ranges to prevent the identifiability of participants.

[c]Missingness (%) reflects the individual percentage of item-wise missing data averaged over all variables.

[d]IDS-C: Inventory of Depressive Symptomatology, Clinician Rated.

[e]EMA: ecological momentary assessment.

Stationarity Tests

The test for stationarity led to the rejection of the null hypothesis for all participants (20/20, 100%), meaning that we rejected the hypothesis that the data-generating process of an individual time series was stationary. The results for all participants are available in Multimedia Appendix 1. These results provided us a first indication of the necessity of using a time-varying approach for our data.

Case Studies

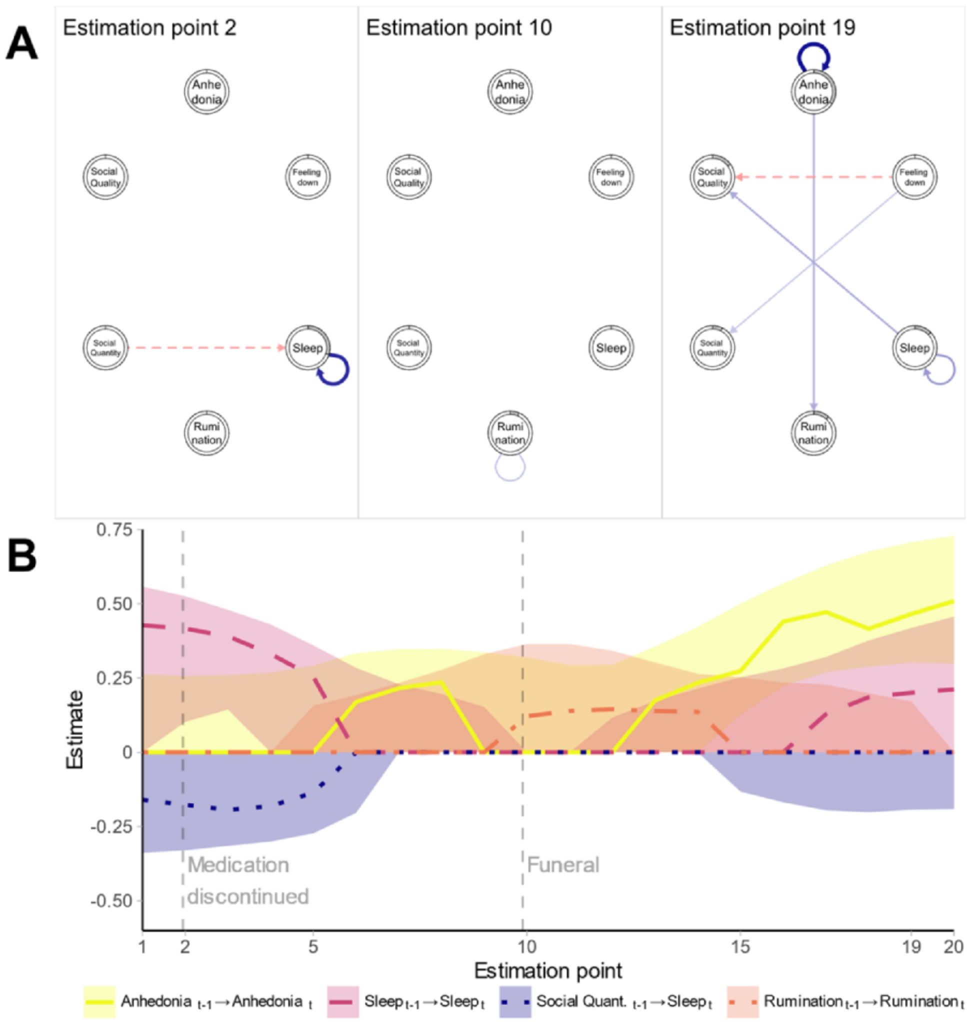
Overview

In the following sections, we have presented the individual results of 2 participants (participants 6 and 11). Participant 6 was chosen because she provided qualitative information about crucial life events during her time series, whereas the network structure of participant 11 changed alongside a reported improved depressive symptomatology over time. We have presented individual networks at estimation points 2, 10, and 19 to showcase models at the beginning, in the middle, and at the end of the time series. We chose to present the 3 edges with the highest intraindividual variability. We have included a case study of participant 2 in Multimedia Appendix 1 to showcase the potential shortcomings of the method that we used.

Participant 6

Participant 6 was a German woman in her 60s, in partial retirement, with a comorbid anxiety disorder. She started multiple antidepressant medications in the year before the study began. She also had previously undergone several psychotherapeutic treatments, and she underwent psychotherapy at the beginning of data collection. A bandwidth of 0.12 was selected for her time series with a final length of 301 days. Her results are visualized in Figure 1.

Figure 1. Results for participant 6: (A) networks at estimation points 2, 10, and 19; (B) time-varying parameters.



The networks at estimation points 2, 10, and 19 (corresponding to days 17, 143, and 285) are displayed in Figure 1A. Width and saturation of edges in the networks were scaled with respect to an arbitrary maximum of 0.5 for all participants, as approximately 80% of all absolute non-0 edge weights were below this threshold. Time-varying parameters are shown in Figure 1B. Estimates at the end of the time series need to be interpreted with caution as fewer data are available for that period. Effects are sorted based on the extent of variability over time in all plots—the effect with the highest variability is plotted in yellow, the effect with the second-highest variability is plotted in violet, and the effect with the third-highest variability is plotted in blue. The autoregressive effect of rumination in orange was added irrespective of its extent of variability over time.

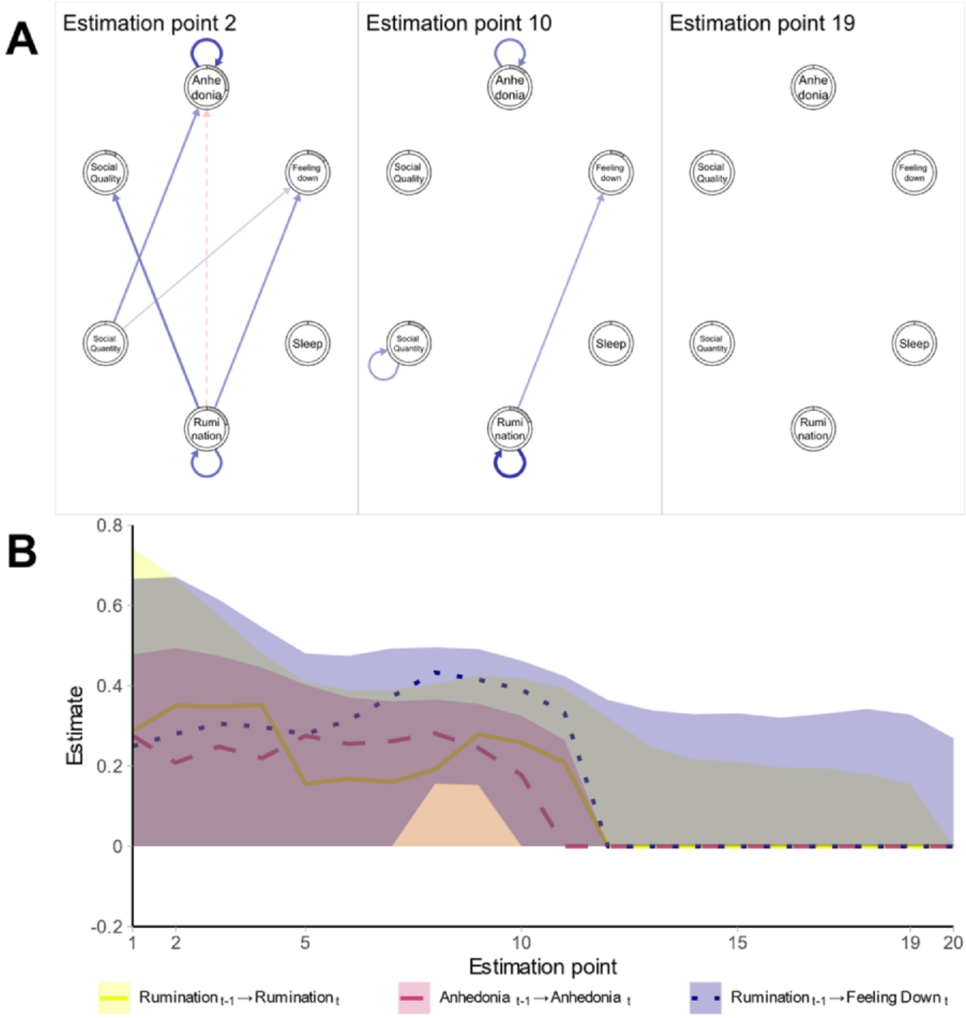
Immediately before the start of the time series, the participant experienced panic attacks, and the dosage of her medication was increased. A few days before the second estimation point (approximately 32 days after the beginning of the time series), she started to phase out her medication, in other words, she continuously reduced the dosage. Both the negative cross-lagged effect from the quantity of social contacts on sleep and the positive autoregressive effect of sleep subsequently disappeared. As self-report sleep involves a different time interval (night) than the remaining measures that summarize the day, the cross-lagged effects including sleep are only interpretable in an asymmetric fashion. In this example, social contacts had an effect on sleep during the subsequent night. Shortly before the 10th estimation point, the participant reported that she visited a funeral and worried about a friend of hers. After that, an autoregressive effect of rumination was detected, which can be interpreted as a prolonged tendency to stay in a ruminative thought process. No more qualitative information about life events was available after this time point. As with many other participants, the bootstrapping results, which we did not include in this paper to keep the plot from being very visually cluttered, indicated a substantial uncertainty of point estimates. This implies that the specific numerical results of an association should be interpreted with caution, as they are likely unstable.

At the end of the time series, the network of participant 6 was relatively strongly connected, with a strong autoregressive effect of anhedonia. Feeling down was positively associated with the quantity of next-day social contacts but negatively associated with their perceived quality. Her IDS-C sum score increased from 34 to 40 from the beginning to the end of the time series, indicating a worsening of her symptomatology. The average R^2 value averaged over all items and estimation points was 0.113, reflecting many estimation points with almost no associations, which we observed for many participants with smaller bandwidths. R^2 was lowest in the middle of the time series and highest at the end. While the networks of these participants were estimated to be empty at many estimation points, the remaining networks changed very fast over time such that most estimated edges were present only for a short time. In some cases, edges even became inverted within a few estimation points, such that the association between 2 variables was positive at one estimation point and negative at another.

Participant 11

Participant 11 was a self-employed German man in his 40s, who was prescribed an antidepressant medication to treat his depression. He did not previously access any psychotherapeutic services. His time series spanned 204 days, and a bandwidth of 0.12 was selected for his data. Again, the estimated networks and visualization of 3 parameters over time are presented in Figure 2.

Figure 2. Results for participant 11: (A) networks at estimation points 2, 10, and 19; (B) time-varying parameters.



The networks at estimation points 2, 10, and 19 (corresponding to days 12, 97, and 193) are displayed in Figure 2A. Time-varying parameters are shown in Figure 2B. The y-axis in this plot is different from that in Figure 1 to accommodate the large width of the bootstrapped sampling distribution.

At the start of the time series, the network of participant 11 was strongly connected. The effect of rumination on depressive symptoms diverged in the beginning, with a negative effect on next-day anhedonia and a positive effect on feeling down. Both anhedonia and rumination showed strong autoregressive effects at the beginning of the time series. This indicates a resistance to change for these variables, meaning that if a larger deviation from their expected value occurs, it takes longer for these variables to return to their "normal" values. Over time, most of these effects became weaker and ultimately disappeared, whereas only autoregressive effects and a positive association of rumination with next-day feeling down remained. At some estimation points, a weak, positive autoregressive effect of both social variables and a positive effect of quality on the quantity of next-day social contacts emerged (not depicted in this paper). In the end, the network of participant 11 became empty, which was also observed in other participants (such as participants 1, 5, and 18). In addition, depressive symptomatology, as measured using the IDS-C, decreased from 19 to 9, which is an improvement to what typically would be judged as a subclinical symptom level. The average R^2 value over time points and variables was 0.228. It decreased toward the end, reflecting the empty networks estimated at the end of the time series. All the bootstrapped sampling distributions around the point estimates pointed toward a strong instability of the point estimates, showing that the interpretation of point estimates warrants caution.

Model Quality

In addition to focusing on the individual associations among variables and their change over time, we investigated model fit indices and the raw distribution of variables to check the overall quality of our models and possible violations of assumptions. The RMSE and R^2 values were computed for every variable of each participant at all time points. Averaged over all participants, the mean RMSE was 0.865 (SD 0.111), and the mean R^2 value was 0.235 (SD 0.181). Participant 9 showed the best fit, with an RMSE of 0.714 and an R^2 value of 0.486, whereas participant 3 showed the worst fit, with an RMSE of 0.972 and an R^2 value of 0.054. Overall, this indicates large differences in model fit among participants, with some models showing very poor fit, whereas others had a relatively good fit to the data. There are many potential reasons for these differences, such as overfitting or underfitting of the models or characteristics of the data that violate the assumptions of the model such as strong nonnormality or abrupt changes.

For 30% (6/20) of the participants, we observed strong floor or ceiling effects in at least 1 item, commonly in either core depressive or social contact items, with the relative frequencies of maximum or minimum scale values exceeding 50%. As these highly skewed data violated the model assumptions, we chose not to interpret the resulting networks of these participants further. However, participant 6 also showed a ceiling effect for the first depression item, as she answered “all the time” for approximately 54.8% (165/301) of the days. Nevertheless, we chose to present her results, as none of the other items were affected, and she provided more relevant qualitative information than any other participant.

Discussion

Principal Findings

The goal of this study was to model the idiographic temporal dynamics of depressive symptoms and other variables associated with depression using daily diary data to gain insight into the temporal dynamics of the disorder. Therefore, we used a recently developed modeling technique that allows for the estimation of time-varying parameters. Both the results of our hypothesis tests and the bandwidth selection procedure provided evidence of substantial changes over time for most participants and thus supported the use of a technique that accounts for these dynamics. We described our results as networks of mutually influencing variables and highlighted the changes in the connections among them over time for exemplar participants. Our results showed extensive variation over time for some participants and marked variability among the networks of different individuals, whereas the bootstrap results suggested the general instability of point estimates.

Individual networks showed temporal associations that might be useful for clinical interpretation and use in self-monitoring contexts. A positive autoregressive effect of rumination, which was present for some estimation points for participants 6 and 11, is sometimes termed as ruminative inertia [74]. Becoming stuck in rumination might be a relevant cognitive mechanism that explains the negative influences of rumination on depression [74]. The contrasting effect of rumination on both depressive symptoms for participant 11 at some estimation points highlights the notion that certain aspects of rumination could also be adaptive for this individual at some times and could therefore differentially impact depressive symptoms. Regarding participant 6, observing a positive effect of feeling down on next-day social quantity at the same time as a negative effect on next-day social quality at the end of her time series could be interpreted as seeking more social contacts after days when she felt depressed, possibly as a remedy or coping strategy, but still being less able to enjoy them or shape them positively.

These person-specific relationships could then be translated into a treatment context by discussing them with the patient or by using them as hypothesis-generating models for potential intervention targets [75]. In addition to specific temporal associations, node-wise summary measures such as node centrality or predictability have been discussed as potential indicators of the relevance of a certain symptom and, subsequently, as potential guides for intervention targets [24,26]. Time-varying networks provide a potential advantage over time-invariant networks in that they could be leveraged for just-in-time adaptive interventions [76]. For example, if the increase of a certain symptom or behavior strongly predicts increased depressive symptomatology, such information could be used to generate personalized interventions. We have discussed the statistical issues and potential solutions related to this topic in the following sections.

As evident in the provided examples, the results of bandwidth selection and in our hypothesis tests, we observed a strong variation in parameter estimates over time, which highlights the substantial variability of symptom interrelations within a person over the course of their depression. This is consistent with the results obtained by Howe et al [70], who found strong intraindividual variation in the networks among different mood states of participants over time. As they discussed, this finding indicates that when associations among symptoms are not time invariant, interventions based on an analysis of symptom-level associations that do not take this variation over time into account might be suboptimal. This could be especially relevant for the development of personalized, just-in-time adaptive interventions, which aim to provide personalized interventions at the right time. For example, if the quantity of social contacts showed a negative predictive association with depressive mood at some point but a positive one at another, an intervention to increase social activity might not always be beneficial. Thus, our results again reinforce recent calls for the use of network analysis methods that are equipped to detect variation over time [26] to better understand the dynamic nature of mental disorders. Additional qualitative information, as we presented for participant 6, could provide important information to interpret the changes in parameters over time and thus increase the clinical utility of the method.

The general decrease in the number and strength of next-day associations alongside an improvement in depressive symptomatology, which was evident for both participant 11 and other participants not shown in this paper, lends itself to an interpretation from a dynamical systems perspective about depression. The idea that individuals with more strongly connected depressive symptoms are more susceptible to ending up in a depressive state (put forth by Cramer et al [77]) has been investigated in various populations and contexts [78]. The time-varying approach used in this study could provide an interesting, new perspective on these issues [25]. The exemplars presented in this paper are not meant to provide any substantial evidence on the general question of the role of network connectivity in mental disorders. Decreased connectivity was, for example, also observed for participant 18, whose symptoms worsened slightly.

Although we can draw interesting insights from individual networks and their change over time, results for participants with a small bandwidth (refer to Multimedia Appendix 1 for a detailed case study) stand out because parameters changed fast and many networks were empty. While it is possible that the symptomatology of these individuals changes quickly and that no days, weeks, or months are alike, there are various other, at least equally plausible reasons for these results. These include the possibility of fundamental issues with assessment, such as the inappropriateness of our measures or inaccurate responses [79]. Irrespective of its root cause, the nature of these results can hamper the usefulness of this method. We have discussed the possible technical solutions to this issue in Multimedia Appendix 1, but general issues with power and interpretability remain notwithstanding.

These considerations point to a more general question: in which contexts can time-varying models be useful? Simulation results have shown that time-varying models can outperform stationary ones even at a low number of observations of approximately 50, under certain conditions [25]. Thus, in principle, these models are both applicable to research, where one may be interested in finding specific time-varying phenomena and clinical contexts, where change over time may be interesting information as feedback for clinicians or patients. To make them useful in the latter case, choosing an appropriate context where gradual change is to be expected is important. In the case of changes due to major life events reported by participants, time-varying models that can accommodate abrupt changes might be an appropriate choice [80]. As time-varying techniques need a large amount of data, the proper selection of a limited number of variables is crucial. The need for a large number of observations per individual can be easier to achieve with passively collected data, such as from fitness trackers or smartphone data. When a large amount of data are available and there is a lack of theoretical knowledge about the form of relationships among variables, time-varying models might prove to be especially useful. In summary, when sufficient data can be collected and some gradual change among variables is to be expected and is of interest, time-varying models can shine.

However, although idiographic network models and time-varying subtypes are promising approaches, their clinical utility has not yet been established [81]. In general, idiographic network models have only been applied to clinical practice in small pilot studies [82]. Models that use purely data-driven approaches based on data from a single individual have several limitations, as they can ignore clinical judgment and can be difficult to estimate and interpret properly [34,83]. These issues can be counteracted by integrating clinical knowledge [83] or information from other individuals into individual networks [33,84]. These potential drawbacks of a purely idiographic approach may seem contradictory to the information in the Introduction section, where the nomothetic-idiographic divide and the advantages of the latter were emphasized. Instead of adopting an either-or perspective, highly person-specific approaches with intensive data collection such as those presented in this paper can still be crucial to provide individual feedback and to detect phenomena that would be obscured with less granular methods, thus serving as building blocks for nomothetic studies that aim to generalize these results.

Strengths and Limitations

Our study had several considerable strengths. We used an innovative modeling technique to explicitly model time-varying parameters in symptom networks. The combination of long individual time series with relatively few missing data and the low number of variables, which is desirable for the performance of the presented method, was conducive to the quality of estimation. Furthermore, our detailed preregistration and open code provide transparency for other researchers.

However, the psychological processes that we are interested in are complex and can occur on a variety of timescales, and our assessments of those processes are affected by measurement error [85,86]. Therefore, our models should be interpreted with several caveats in mind. In general, estimated effects are strongly dependent on which variables are present or absent in the network [87], implying that the inclusion of more depressive symptoms would possibly change our results. Furthermore, the results of parameter estimation crucially depend on the chosen sampling frequency [88] and the subsequent lag choice [17]. Regarding measurement, the use of single items to assess psychological constructs has psychometric disadvantages [14] and may obfuscate the inherent heterogeneity of what likely should not be considered as homogenous constructs (refer to the paper by Bernstein et al [35] for the example of rumination). Moreover, the specific items used in this study were not previously assessed for validity or reliability.

A large number of empty or very sparse networks point to two further limitations of our method. In general, idiographic network models often experience power problems [33], which are further exacerbated in the power-hungry estimation of time-varying networks. Relatedly, regularization techniques decrease sensitivity and prohibit the construction of conventional CIs (refer to the paper by Williams et al [89] and Williams [90] for a discussion regarding lasso in network estimation).

Future Directions

There are multiple avenues for further studies. First, applied researchers can use time-varying VAR models in scenarios where substantial change can be expected and is of core interest, and repeated intensive assessment is feasible to better understand the temporal development of mental disorders. Second, building on its potential clinical use, further methodological research into the estimation method used in this study could provide more information on best practices regarding modeling options. Third, we observed the potential utility of qualitative information as context for time series data. Thus, further studies could investigate best practices regarding the collection and integration of qualitative information into intensive, longitudinal designs and analyses.

Conclusions

Attempts to develop personalized models of psychopathology have become increasingly refined in recent years. While there have been large advances regarding the modeling of interindividual variation, studies of variation within individuals with high temporal solution have lagged. We have made a step forward in this direction by explicitly modeling individual changes in the associations among depressive symptoms over time. Pronounced within-person variation in our results highlights the importance of investigating the temporal dynamics of symptoms and their interplay over time.

Acknowledgments

This work was supported by the DYNAMIC center, which is funded by the LOEWE program of the Hessian Ministry of Science and Arts (grant LOEWE1/16/519/03/09.001(0009)/98). The Sensor-Based System for Therapy Support and Management of Depression project was supported by a grant from the German Federal Ministry of Education and Research (project 13GW0162B/13GW0162A). The Sensor-Based System for Therapy Support and Management of Depression Consortium consisted of the German Depression Foundation; Institute for Applied Informatics, University Leipzig; Institute for Applied Informatics Management GmbH; adesso SE; and Information Systems Institute, University Leipzig. BSS was supported with a scholarship of the "Deutschlandstipendium" by the Goethe University Frankfurt during the creation of this paper (originally as a thesis in the psychology master's program).

Authors' Contributions

BSS, CS, AK, UH, and HR were involved in conceptualization. BSS and SL were responsible for the software. BSS, MS, and SL were involved in formal analysis. CS was involved in data curation. BSS, MS, and HR were involved in writing the original draft. BSS, CS, MS, AK, SL, UH, and HR were involved in reviewing and editing**.** BSS was responsible for the visualizations. CS, MS, UH, and HR were involved in supervision. CS, AK, UH, and HR were involved in project administration. CS, AK, UH were responsible for funding acquisition.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary materials.
[DOCX File , 1794 KB - mental_v11i1e50136_app1.docx]

References

1. Cai N, Choi KW, Fried EI. Reviewing the genetics of heterogeneity in depression: operationalizations, manifestations and etiologies. Hum Mol Genet 2020 Sep 30;29(R1):R10-R18 [FREE Full text] [doi: 10.1093/hmg/ddaa115] [Medline: 32568380]
2. Fried EI. Problematic assumptions have slowed down depression research: why symptoms, not syndromes are the way forward. Front Psychol 2015 Mar 23;6:309 [FREE Full text] [doi: 10.3389/fpsyg.2015.00309] [Medline: 25852621]
3. Kessler RC, van Loo HM, Wardenaar KJ, Bossarte RM, Brenner LA, Ebert DD, et al. Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. Epidemiol Psychiatr Sci 2017 Feb;26(1):22-36 [FREE Full text] [doi: 10.1017/S2045796016000020] [Medline: 26810628]
4. Maj M, Stein DJ, Parker G, Zimmerman M, Fava GA, de Hert M, et al. The clinical characterization of the adult patient with depression aimed at personalization of management. World Psychiatry 2020 Oct;19(3):269-293 [FREE Full text] [doi: 10.1002/wps.20771] [Medline: 32931110]
5. Wright AG, Woods WC. Personalized models of psychopathology. Annu Rev Clin Psychol 2020 May 07;16:49-74 [FREE Full text] [doi: 10.1146/annurev-clinpsy-102419-125032] [Medline: 32070120]
6. Borsboom D, Mellenbergh GJ, van Heerden J. The theoretical status of latent variables. Psychol Rev 2003 Apr;110(2):203-219. [doi: 10.1037/0033-295X.110.2.203] [Medline: 12747522]
7. Hamaker EL. Why researchers should think "within-person": a paradigmatic rationale. In: Mehl MR, Conner TS, editors. Handbook of Research Methods for Studying Daily Life. New York, NY: The Guilford Press; 2012:43-61.
8. Molenaar PC. A manifesto on psychology as idiographic science: bringing the person back into scientific psychology, this time forever. Meas Interdiscip Res Perspect 2004 Oct;2(4):201-218. [doi: 10.1207/s15366359mea0204_1]
9. Fisher AJ, Medaglia JD, Jeronimus BF. Lack of group-to-individual generalizability is a threat to human subjects research. Proc Natl Acad Sci U S A 2018 Jul 03;115(27):E6106-E6115 [FREE Full text] [doi: 10.1073/pnas.1711978115] [Medline: 29915059]
10. Borsboom D. A network theory of mental disorders. World Psychiatry 2017 Feb 26;16(1):5-13 [FREE Full text] [doi: 10.1002/wps.20375] [Medline: 28127906]
11. Bringmann LF, Vissers N, Wichers M, Geschwind N, Kuppens P, Peeters F, et al. A network approach to psychopathology: new insights into clinical longitudinal data. PLoS One 2013 Apr 4;8(4):e60188 [FREE Full text] [doi: 10.1371/journal.pone.0060188] [Medline: 23593171]
12. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008 Apr;4(1):1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
13. Hayes SC, Hofmann SG, Stanton CE, Carpenter JK, Sanford BT, Curtiss JE, et al. The role of the individual in the coming era of process-based therapy. Behav Res Ther 2019 Jun;117:40-53. [doi: 10.1016/j.brat.2018.10.005] [Medline: 30348451]

14. Piccirillo ML, Beck ED, Rodebaugh TL. A clinician's primer for idiographic research: considerations and recommendations. Behav Ther 2019 Sep;50(5):938-951. [doi: 10.1016/j.beth.2019.02.002] [Medline: 31422849]
15. Rubel JA, Fisher AJ, Husen K, Lutz W. Translating person-specific network models into personalized treatments: development and demonstration of the dynamic assessment treatment algorithm for individual networks (DATA-IN). Psychother Psychosom 2018;87(4):249-251 [FREE Full text] [doi: 10.1159/000487769] [Medline: 29680835]
16. Wang L, Miller LC. Just-in-the-moment adaptive interventions (JITAI): a meta-analytical review. Health Commun 2020 Nov 05;35(12):1531-1544. [doi: 10.1080/10410236.2019.1652388] [Medline: 31488002]
17. Hamaker EL, Wichers M. No time like the present: discovering the hidden dynamics in intensive longitudinal data. Curr Dir Psychol Sci 2017 Feb 08;26(1):10-15. [doi: 10.1177/0963721416666518]
18. Borsboom D, Cramer AO. Network analysis: an integrative approach to the structure of psychopathology. Annu Rev Clin Psychol 2013 Mar 28;9(1):91-121. [doi: 10.1146/annurev-clinpsy-050212-185608] [Medline: 23537483]
19. Wichers M, Riese H, Hodges TM, Snippe E, Bos FM. A narrative review of network studies in depression: what different methodological approaches tell us about depression. Front Psychiatry 2021 Oct 28;12:719490 [FREE Full text] [doi: 10.3389/fpsyt.2021.719490] [Medline: 34777038]
20. David SJ, Marshall AJ, Evanovich EK, Mumma GH. Intraindividual dynamic network analysis - implications for clinical assessment. J Psychopathol Behav Assess 2018 Dec 1;40(2):235-248 [FREE Full text] [doi: 10.1007/s10862-017-9632-8] [Medline: 29937621]
21. Fisher AJ, Reeves JW, Lawyer G, Medaglia JD, Rubel JA. Exploring the idiographic dynamics of mood and anxiety via network analysis. J Abnorm Psychol 2017 Nov;126(8):1044-1056 [FREE Full text] [doi: 10.1037/abn0000311] [Medline: 29154565]
22. Rodriguez M, Aalbers G, McNally RJ. Idiographic network models of social media use and depression symptoms. Cogn Ther Res 2021 May 22;46(1):124-132. [doi: 10.1007/s10608-021-10236-2]
23. Kaiser T, Laireiter AR. Process-symptom-bridges in psychotherapy: an idiographic network approach. J Pers Oriented Res 2018 Dec 27;4(2):49-62 [FREE Full text] [doi: 10.17505/jpor.2018.06] [Medline: 33569133]
24. Levinson CA, Hunt RA, Keshishian AC, Brown ML, Vanzhula I, Christian C, et al. Using individual networks to identify treatment targets for eating disorder treatment: a proof-of-concept study and initial data. J Eat Disord 2021 Nov 04;9(1):147 [FREE Full text] [doi: 10.1186/s40337-021-00504-7] [Medline: 34736538]
25. Haslbeck JM, Bringmann LF, Waldorp LJ. A tutorial on estimating time-varying vector autoregressive models. Multivariate Behav Res 2021 Apr 23;56(1):120-149. [doi: 10.1080/00273171.2020.1743630] [Medline: 32324066]
26. Bringmann LF, Albers C, Bockting C, Borsboom D, Ceulemans E, Cramer A, et al. Psychopathological networks: theory, methods and practice. Behav Res Ther 2022 Feb;149:104011 [FREE Full text] [doi: 10.1016/j.brat.2021.104011] [Medline: 34998034]
27. Wichers M, Groot PC, Psychosystems, ESM Group, EWS Group. Critical slowing down as a personalized early warning signal for depression. Psychother Psychosom 2016 Jan 26;85(2):114-116 [FREE Full text] [doi: 10.1159/000441458] [Medline: 26821231]
28. Wrzus C, Neubauer A. Ecological momentary assessment: a meta-analysis on designs, samples, and compliance across research fields. PsyArXiv Preprints posted online on October 13, 2021 [FREE Full text] [doi: 10.31234/osf.io/vmr8q]
29. Haslbeck JM, Waldorp LJ. mgm: estimating time-varying mixed graphical models in high-dimensional data. J Stat Softw 2020;93(8):1-46 [FREE Full text] [doi: 10.18637/jss.v093.i08]
30. Lorenz N, Sander C, Ivanova G, Hegerl U. Temporal associations of daily changes in sleep and depression core symptoms in patients suffering from major depressive disorder: idiographic time-series analysis. JMIR Ment Health 2020 Apr 23;7(4):e17071 [FREE Full text] [doi: 10.2196/17071] [Medline: 32324147]
31. Drieling T, Schärer LO, Langosch JM. The inventory of depressive symptomatology: German translation and psychometric validation. Int J Methods Psychiatr Res 2007 Dec;16(4):230-236 [FREE Full text] [doi: 10.1002/mpr.226] [Medline: 18200596]
32. Boden S. Diagnostik von Depressivität: Validierung des Inventars depressiver Symptome (IDS). Eberhard Karl University of Tübingen. 2018. URL: https://publikationen.uni-tuebingen.de/xmlui/handle/10900/87535 [accessed 2024-03-23]
33. Mansueto AC, Wiers RW, van Weert JC, Schouten BC, Epskamp S. Investigating the feasibility of idiographic network models. Psychol Methods 2023 Oct;28(5):1052-1068. [doi: 10.1037/met0000466] [Medline: 34990189]
34. Klintwall L, Bellander M, Cervin M. Perceived causal problem networks: reliability, central problems, and clinical utility for depression. Assessment 2023 Jan 01;30(1):73-83 [FREE Full text] [doi: 10.1177/10731911211039281] [Medline: 34467772]
35. Bernstein EE, Heeren A, McNally RJ. Reexamining trait rumination as a system of repetitive negative thoughts: a network analysis. J Behav Ther Exp Psychiatry 2019 Jun;63:21-27. [doi: 10.1016/j.jbtep.2018.12.005] [Medline: 30590225]
36. Nolen-Hoeksema S, Wisco BE, Lyubomirsky S. Rethinking rumination. Perspect Psychol Sci 2008 Sep;3(5):400-424. [doi: 10.1111/j.1745-6924.2008.00088.x] [Medline: 26158958]
37. Olatunji BO, Naragon-Gainey K, Wolitzky-Taylor KB. Specificity of rumination in anxiety and depression: a multimodal meta‐analysis. Clin Psychol Sci Pract 2013 Sep;20(3):225-257. [doi: 10.1037/h0101719]

38. Watkins E. Psychological treatment of depressive rumination. Curr Opin Psychol 2015 Aug;4:32-36. [doi: 10.1016/j.copsyc.2015.01.020]
39. Ebrahimi OV, Burger J, Hoffart A, Johnson SU. Within- and across-day patterns of interplay between depressive symptoms and related psychopathological processes: a dynamic network approach during the COVID-19 pandemic. BMC Med 2021 Nov 30;19(1):317 [FREE Full text] [doi: 10.1186/s12916-021-02179-y] [Medline: 34844588]
40. Lunansky G, van Borkulo CD, Haslbeck JM, van der Linden MA, Garay CJ, Etchevers MJ, et al. The mental health ecosystem: extending symptom networks with risk and protective factors. Front Psychiatry 2021 Mar 18;12:640658 [FREE Full text] [doi: 10.3389/fpsyt.2021.640658] [Medline: 33815173]
41. Hirschfeld RM, Montgomery SA, Keller MB, Kasper S, Schatzberg AF, Möller HJ, et al. Social functioning in depression: a review. J Clin Psychiatry 2000 Apr 15;61(4):268-275. [doi: 10.4088/jcp.v61n0405] [Medline: 10830147]
42. Gariépy G, Honkaniemi H, Quesnel-Vallée A. Social support and protection from depression: systematic review of current findings in Western countries. Br J Psychiatry 2016 Oct 02;209(4):284-293. [doi: 10.1192/bjp.bp.115.169094] [Medline: 27445355]
43. Brown LH, Strauman T, Barrantes-Vidal N, Silvia PJ, Kwapil TR. An experience-sampling study of depressive symptoms and their social context. J Nerv Ment Dis 2011 Jun;199(6):403-409. [doi: 10.1097/NMD.0b013e31821cd24b] [Medline: 21629020]
44. Elmer T, Stadtfeld C. Depressive symptoms are associated with social isolation in face-to-face interaction networks. Sci Rep 2020 Jan 29;10(1):1444 [FREE Full text] [doi: 10.1038/s41598-020-58297-9] [Medline: 31996728]
45. Chevance A, Ravaud P, Tomlinson A, Le Berre C, Teufer B, Touboul S, et al. Identifying outcomes for depression that matter to patients, informal caregivers, and health-care professionals: qualitative content analysis of a large international online survey. Lancet Psychiatry 2020 Aug;7(8):692-702. [doi: 10.1016/S2215-0366(20)30191-7] [Medline: 32711710]
46. Elmer T, Geschwind N, Peeters F, Wichers M, Bringmann L. Getting stuck in social isolation: solitude inertia and depressive symptoms. J Abnorm Psychol 2020 Oct;129(7):713-723. [doi: 10.1037/abn0000588] [Medline: 32672987]
47. Fried EI. The 52 symptoms of major depression: lack of content overlap among seven common depression scales. J Affect Disord 2017 Jan 15;208:191-197. [doi: 10.1016/j.jad.2016.10.019] [Medline: 27792962]
48. Baglioni C, Nanovska S, Regen W, Spiegelhalder K, Feige B, Nissen C, et al. Sleep and mental disorders: a meta-analysis of polysomnographic research. Psychol Bull 2016 Sep;142(9):969-990 [FREE Full text] [doi: 10.1037/bul0000053] [Medline: 27416139]
49. Nutt D, Wilson S, Paterson L. Sleep disorders as core symptoms of depression. Dialogues Clin Neurosci 2022 Apr 01;10(3):329-336. [doi: 10.31887/dcns.2008.10.3/dnutt]
50. Alvaro PK, Roberts RM, Harris JK. A systematic review assessing bidirectionality between sleep disturbances, anxiety, and depression. Sleep 2013 Jul 01;36(7):1059-1068 [FREE Full text] [doi: 10.5665/sleep.2810] [Medline: 23814343]
51. Cunningham JE, Shapiro CM. Cognitive behavioural therapy for insomnia (CBT-I) to treat depression: a systematic review. J Psychosom Res 2018 Mar;106:1-12. [doi: 10.1016/j.jpsychores.2017.12.012] [Medline: 29455893]
52. Pillai V, Steenburg LA, Ciesla JA, Roth T, Drake CL. A seven day actigraphy-based study of rumination and sleep disturbance among young adults with depressive symptoms. J Psychosom Res 2014 Jul;77(1):70-75. [doi: 10.1016/j.jpsychores.2014.05.004] [Medline: 24913345]
53. Manea L, Gilbody S, Hewitt C, North A, Plummer F, Richardson R, et al. Identifying depression with the PHQ-2: a diagnostic meta-analysis. J Affect Disord 2016 Oct;203:382-395. [doi: 10.1016/j.jad.2016.06.003] [Medline: 27371907]
54. Staples LG, Dear BF, Gandy M, Fogliati V, Fogliati R, Karin E, et al. Psychometric properties and clinical utility of brief measures of depression, anxiety, and general distress: the PHQ-2, GAD-2, and K-6. Gen Hosp Psychiatry 2019 Jan;56:13-18 [FREE Full text] [doi: 10.1016/j.genhosppsych.2018.11.003] [Medline: 30508772]
55. Chen J, Chen Z. Extended Bayesian information criteria for model selection with large model spaces. Biometrika 2008 Sep 01;95(3):759-771 [FREE Full text] [doi: 10.1093/biomet/asn034]
56. Epskamp S, Fried EI. A tutorial on regularized partial correlation networks. Psychol Methods 2018 Dec;23(4):617-634. [doi: 10.1037/met0000167] [Medline: 29595293]
57. Hastie T, Friedman JH, Tibshirani R. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd edition. Cham, Switzerland: Springer; 2017.
58. Dablander F, Ryan O, Haslbeck JM. Choosing between AR(1) and VAR(1) models in typical psychological applications. PLoS One 2020 Oct 29;15(10):e0240730 [FREE Full text] [doi: 10.1371/journal.pone.0240730] [Medline: 33119716]
59. Kirtley OJ, Lafit G, Achterhof R, Hiekkaranta AP, Myin-Germeys I. Making the black box transparent: a template and tutorial for registration of studies using experience-sampling methods. Adv Methods Pract Psychol Sci 2021 Mar 02;4(1):251524592092468. [doi: 10.1177/2515245920924686]
60. Siepe BS, Sander C, Schultze M, Kliem A, Ludwig S, Hegerl U, et al. Supplementary materials for temporal dynamics of depressive symptomatology: an idiographic time series analysis applying network models to patients with depressive disorders. PsyArXiv Preprints posted online on July 13, 2022 [FREE Full text] [doi: 10.31234/osf.io/hnw69]
61. First MB, Gibbon M. The structured clinical interview for DSM-IV Axis I Disorders (SCID-I) and the structured clinical interview for DSM-IV Axis II Disorders (SCID-II). In: Hilsenroth MJ, Segal DL, Hersen M, editors. Comprehensive Handbook of Psychological Assessment. Volume 2. Hoboken, NJ: John Wiley & Sons; 2004:134-143.

62. Wittchen HU, Zaudig M, Fydrich T. Strukturiertes Klinisches Interview Für DSM-IV (SKID): Achse I Und II. Göttingen, Germany: Hogrefe Publishing; 1997.
63. Rush AJ, Gullion CM, Basco MR, Jarrett RB, Trivedi MH. The Inventory of Depressive Symptomatology (IDS): psychometric properties. Psychol Med 1996 May;26(3):477-486 [FREE Full text] [doi: 10.1017/s0033291700035558] [Medline: 8733206]
64. Kroenke K, Spitzer RL, Williams JB. The patient health questionnaire-2: validity of a two-item depression screener. Med Care 2003;4(11):1284-1292 [FREE Full text] [doi: 10.1097/01.mlr.0000093487.78664.3c]
65. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). New York, NY: American Psychiatric Association; 2013.
66. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. URL: https://www.R-project.org/ [accessed 2024-03-23]
67. Moritz S, Bartz-Beielstein T. imputeTS: time series missing value imputation in R. R J 2017;9(1):207-218 [FREE Full text] [doi: 10.32614/rj-2017-009]
68. Bringmann LF, Ferrer E, Hamaker EL, Borsboom D, Tuerlinckx F. Modeling nonstationary emotion dynamics in dyads using a time-varying vector-autoregressive model. Multivariate Behav Res 2018 Mar 05;53(3):293-314 [FREE Full text] [doi: 10.1080/00273171.2018.1439722] [Medline: 29505311]
69. Chen M, Chow SM, Hammal Z, Messinger DS, Cohn JF. A person- and time-varying vector autoregressive model to capture interactive infant-mother head movement dynamics. Multivariate Behav Res 2021 Jun 12;56(5):739-767 [FREE Full text] [doi: 10.1080/00273171.2020.1762065] [Medline: 32530313]
70. Howe E, Bosley HG, Fisher AJ. Idiographic network analysis of discrete mood states prior to treatment. Couns Psychother Res 2020 Feb 14;20(3):470-478 [FREE Full text] [doi: 10.1002/capr.12295]
71. Epskamp S, Cramer AOJ, Waldorp LJ, Schmittmann VD, Borsboom D. qgraph: network visualizations of relationships in psychometric data. J Stat Softw 2012;48(4):1-18 [FREE Full text] [doi: 10.18637/jss.v048.i04]
72. Ghosh S. Kernel Smoothing: Principles, Methods and Applications. Hoboken, NJ: John Wiley & Sons; 2018.
73. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Series B Stat Methodol 2018 Dec 05;58(1):267-288 [FREE Full text] [doi: 10.1111/j.2517-6161.1996.tb02080.x]
74. Bean CA, Heggeness LF, Kalmbach DA, Ciesla JA. Ruminative inertia and its association with current severity and lifetime course of depression. Clin Psychol Sci 2020 Sep 15;8(6):1007-1016. [doi: 10.1177/2167702620949174]
75. Scholten S, Rubel J, Glombiewski J, Milde C. What time-varying network models based on functional analysis tell us about the course of a patient's problem. OSF Preprint posted online on May 16, 2023 [FREE Full text] [doi: 10.31219/osf.io/gz3c2]
76. Nemesure MD, Collins AC, Price GD, Griffin TZ, Pillai A, Nepal S, et al. Depressive symptoms as a heterogeneous and constantly evolving dynamical system: idiographic depressive symptom networks of rapid symptom changes among persons with major depressive disorder. J Psychopathol Clin Sci 2024 Feb;133(2):155-166. [doi: 10.1037/abn0000884] [Medline: 38271054]
77. Cramer AO, van Borkulo CD, Giltay EJ, van der Maas HL, Kendler KS, Scheffer M, et al. Major depression as a complex dynamic system. PLoS One 2016 Dec 8;11(12):e0167490 [FREE Full text] [doi: 10.1371/journal.pone.0167490] [Medline: 27930698]
78. Robinaugh DJ, Hoekstra RH, Toner ER, Borsboom D. The network approach to psychopathology: a review of the literature 2008-2018 and an agenda for future research. Psychol Med 2020 Feb 26;50(3):353-366 [FREE Full text] [doi: 10.1017/S0033291719003404] [Medline: 31875792]
79. Jaso BA, Kraus NI, Heller AS. Identification of careless responding in ecological momentary assessment research: from posthoc analyses to real-time data monitoring. Psychol Methods 2022 Dec 16;27(6):958-981. [doi: 10.1037/met0000312] [Medline: 34582244]
80. Albers CJ, Bringmann LF. Inspecting gradual and abrupt changes in emotion dynamics with the time-varying change point autoregressive model. Eur J Psychol Assess 2020 May;36(3):492-499. [doi: 10.1027/1015-5759/a000589]
81. Bringmann LF. Person-specific networks in psychopathology: past, present, and future. Curr Opin Psychol 2021 Oct;41:59-64 [FREE Full text] [doi: 10.1016/j.copsyc.2021.03.004] [Medline: 33862345]
82. Frumkin MR, Piccirillo ML, Beck ED, Grossman JT, Rodebaugh TL. Feasibility and utility of idiographic models in the clinic: a pilot study. Psychother Res 2021 Apr 24;31(4):520-534 [FREE Full text] [doi: 10.1080/10503307.2020.1805133] [Medline: 32838671]
83. Burger J, Epskamp S, van der Veen DC, Dablander F, Schoevers RA, Fried EI, et al. A clinical PREMISE for personalized models: toward a formal integration of case formulations and statistical networks. J Psychopathol Clin Sci 2022 Nov;131(8):906-916. [doi: 10.1037/abn0000779] [Medline: 36326631]
84. Beltz AM, Wright AG, Sprague BN, Molenaar PC. Bridging the nomothetic and idiographic approaches to the analysis of clinical data. Assessment 2016 Aug;23(4):447-458 [FREE Full text] [doi: 10.1177/1073191116648209] [Medline: 27165092]
85. Fisher ZF, Chow SM, Molenaar PC, Fredrickson BL, Pipiras V, Gates KM. A square-root second-order extended Kalman filtering approach for estimating smoothly time-varying parameters. Multivariate Behav Res 2022 Oct 07;57(1):134-152 [FREE Full text] [doi: 10.1080/00273171.2020.1815513] [Medline: 33025834]
86. Schuurman NK, Houtveen JH, Hamaker EL. Incorporating measurement error in n = 1 psychological autoregressive modeling. Front Psychol 2015 Jul 28;6:1038 [FREE Full text] [doi: 10.3389/fpsyg.2015.01038] [Medline: 26283988]

87. Fried EI, Cramer AO. Moving forward: challenges and directions for psychopathological network theory and methodology. Perspect Psychol Sci 2017 Nov 05;12(6):999-1020. [doi: 10.1177/1745691617705892] [Medline: 28873325]
88. Haslbeck JM, Ryan O. Recovering within-person dynamics from psychological time series. Multivariate Behav Res 2022 Jun 21;57(5):735-766. [doi: 10.1080/00273171.2021.1896353] [Medline: 34154483]
89. Williams DR, Rhemtulla M, Wysocki AC, Rast P. On nonregularized estimation of psychological networks. Multivariate Behav Res 2019;54(5):719-750 [FREE Full text] [doi: 10.1080/00273171.2019.1575716] [Medline: 30957629]
90. Williams DR. The confidence interval that wasn't: bootstrapped "confidence intervals" in L1-regularized partial correlation networks. PsyArXiv Preprints posted online on May 5, 2021 [FREE Full text] [doi: 10.31234/osf.io/kjh2f]

Abbreviations

IDS-C: Inventory of Depressive Symptomatology, Clinician Rated
lasso: least absolute shrinkage and selection operator
RMSE: root mean squared error
STEADY: Sensor-Based System for Therapy Support and Management of Depression
VAR: vector autoregressive

Edited by J Torous; submitted 20.06.23; peer-reviewed by L Bringmann, M Rubin; comments to author 04.01.24; revised version received 27.01.24; accepted 14.02.24; published 18.04.24.

Please cite as:
Siepe BS, Sander C, Schultze M, Kliem A, Ludwig S, Hegerl U, Reich H
Time-Varying Network Models for the Temporal Dynamics of Depressive Symptomatology in Patients With Depressive Disorders: Secondary Analysis of Longitudinal Observational Data
JMIR Ment Health 2024;11:e50136
URL: https://mental.jmir.org/2024/1/e50136
doi:10.2196/50136
PMID:38635978

Original Paper

Feasibility and Acceptability of a Mobile App–Based TEAM-CBT (Testing Empathy Assessment Methods–Cognitive Behavioral Therapy) Intervention (Feeling Good) for Depression: Secondary Data Analysis

Nicholas Bisconti[1*], BA, MS; Mackenzie Odier[1*], BA, MS; Matthew Becker[1*], BA, MS; Kim Bullock[2*], MD

[1]PGSP-Stanford PsyD. Consortium, Palo Alto, CA, United States

[2]Stanford School of Medicine Department of Psychiatry and Behavioral Sciences, Stanford University, Palo Alto, CA, United States

[*]all authors contributed equally

Corresponding Author:
Nicholas Bisconti, BA, MS
PGSP-Stanford PsyD. Consortium
401 Quarry Rd
Palo Alto, CA, 94304
United States
Phone: 1 6506449946
Email: nbisco@stanford.edu

Abstract

Background: The Feeling Good App is an automated stand-alone digital mobile mental health tool currently undergoing beta testing with the goal of providing evidence-informed self-help lessons and exercises to help individuals reduce depressive symptoms without guidance from a mental health provider. Users work through intensive basic training (IBT) and ongoing training models that provide education regarding cognitive behavioral therapy principles from a smartphone.

Objective: The key objective of this study was to perform a nonsponsored third-party academic assessment of an industry-generated data set; this data set focused on the safety, feasibility, and accessibility of a commercial automated digital mobile mental health app that was developed to reduce feelings associated with depression.

Methods: The Feeling Good App development team created a waitlist cohort crossover design and measured symptoms of depression and anxiety using the Patient Health Questionnaire-9, Generalized Anxiety Disorder-7, and an app-specific measure of negative feelings called the 7 Dimension Emotion Slider (7-DES). The waitlist cohort crossover design divided the participants into 2 groups, where 48.6% (141/290) of the participants were given immediate access to the apps, while 51.4% (149/290) were placed on a 2-week waitlist before being given access to the app. Data collected by the Feeling Good App development team were deidentified and provided to the authors of this paper for analysis through a nonsponsored university data use agreement. All quantitative data were analyzed using SPSS Statistics (version 28.0; IBM Corp). Descriptive statistics were calculated for demographic variables. Feasibility and acceptability were descriptively assessed. All participants included in the quantitative data were given access to the Feeling Good App; this study did not include a control group.

Results: In terms of safety, there was no statistically significant change in suicidality from preintervention to postintervention time points (t_{288}=0.0; P>.99), and there was a statistically significant decrease in hopelessness from preintervention to postintervention time points (F_{289}=30.16; P<.01). In terms of acceptability, 72.2% (166/230) of the users who started the initial 2-day IBT went on to complete it, while 34.8% (80/230) of the users who started IBT completed the entirety of the apps' 4-week protocol (150/230, 65.22% dropout rate over 4 weeks).

Conclusions: This study is the first reported proof-of-concept evaluation of the Feeling Good App in terms of safety, feasibility, and statistical trends within the data set. It demonstrates a feasible and novel approach to industry and academic collaboration in the process of developing a digital mental health technology translated from an existing evidence-informed treatment. The results support the prototype app as safe for a select nonclinical population. The app had acceptable levels of engagement and dropouts throughout the intervention. Those who stay engaged showed reductions in symptom severity of depression warranting further investigation of the app's efficacy.

(JMIR Ment Health 2024;11:e52369) doi:10.2196/52369

KEYWORDS

depression; mobile health; mHealth; cognitive behavioral therapy; mobile phone

Introduction

Background

Depression is a highly prevalent illness that affects approximately 280 million people worldwide [1]. Depression leads to varying degrees of functional impairment and distress [2]. Depression is most often treated through pharmacological intervention, psychotherapy, or neuromodulation with varying efficacy [3]. Suicide is the second leading cause of death in the United States among individuals aged between 10 and 14 years and 25 and 34 years, the third leading cause of death among individuals aged between 15 and 24 years, and the fourth leading cause of death among individuals aged between 35 and 44 years [4,5]. There are approximately 2 times as many suicides (45,979) in the United States as there are homicides (n=24,576) [4]. In pre–COVID-19 times, only 56.8% of people diagnosed with major depressive disorder obtained care. For those who reach out for help, it is estimated that only 37.5% receive minimally adequate treatment or evidence-based psychotherapies [4]. Untreated mental health issues have immense negative effects, from the psychological and physical impacts on the individual to increased health care use costs and loss of productivity that impact communities, societies, and economies [5,6]. Depression is the world's leading cause of disability, and despite various known and effective treatments, >75% of people in low- and middle-income countries receive no treatment [7]. Many individuals with depression are unable to receive effective treatment for a variety of reasons including due to a lack of resources [2], barriers to access such as systems without enough trained providers, and concerns about privacy and stigma. Providing accessible scalable services at low cost is imperative in combating the global mental health crisis.

Mobile smartphone apps and internet-based treatments present a tremendous opportunity to increase the accessibility of mental health care due to their ability to scale and reduce other barriers to entry [8,9]. A meta-analysis from Serrano-Ripoll et al [10] examined existing app-based interventions for reducing depressive symptoms and found that most of these apps significantly reduced symptoms, with moderate effect sizes (0.51) and significantly larger efficacy in more severe cases. There are a variety of publicly available mobile apps for depression that demonstrate clinically significant decreases in depression. Many interventions use a mixed model of automated multimedia and guided interventions to treat symptoms of depression. These interventions, which use both user-guided material as well as clinician support (nurse-moderated and therapist-guided), are shown to be effective, yet they are limited in their practicality and scale, as they rely heavily on clinician support [11-15]. There is also substantial literature into the efficacy of self-guided apps to treat symptoms of depression and anxiety, with many of the publicly available apps based in basic psychoeducation, mindfulness, and cognitive behavioral therapy (CBT). CBT and mindfulness-based apps, interventions that typically range from 4 to 12 weeks, show significant reductions in symptoms and symptom severity from baseline [16-23].

With >10,000 mental health apps in the market, the choices are overwhelming, and knowing how to evaluate and choose a high-quality app is important. One potential danger in the development and innovations surrounding mental health mobile app–based solutions for depression is that they can be distributed without scientific evidence supporting their treatment techniques [10]. A recent study by Larsen et al [24] showed that in the description of the top 73 mental health apps, 44% used general scientific language such as *evidence-based treatment* while only 2 apps cited low-quality primary evidence and 1 app included a citation to published literature [24]. Differentiating between an app being based in evidence (ie, informed and including principles of evidence-based treatments) and being evidence based (ie, research demonstrating the app is efficacious) is crucial for both app developers and consumers alike [25]. Often, obtaining high-quality evidence is challenging given the cost of industry and academia collaboration in the form of sponsored projects. These projects often occur after the app has been developed and do not leave room for academic or clinical input into industry design during the translational process.

A challenge for mobile mental health solutions is whether an effective app is actually being used. In assessing the usability of a mobile mental health app, one of the primary criteria is adherence. Baumel et al [8] showed that in a study of 100,000 downloads of mental health apps, the median retention rate after 15 days was only 3.9%. In a review of randomized controlled trials of smartphones apps targeting depressive symptoms, Baumel et al [8] found a mean dropout rate of 26.2% that increased to 47.8% when accounting for publication bias. While digital tools with no human interaction have shown to be similarly effective compared with traditional face-to-face treatment, dropout rates are more than twice as high (33% dropout with no interaction) when compared with 11% dropout rates in apps with human interaction [8].

As many conventional evidence-based therapies make the leap to incorporating technology and mobile app use into their dissemination processes, evaluating the efficacy and use of these innovations by independent clinician researchers is crucial. Our group set out to evaluate such a mobile app based on a very popular and common therapy from Dr David Burns' *Feeling Great*, a self-help CBT book for depression by the Feeling Good Corporation. The Feeling Good App being developed is a smartphone app based upon the principles of Testing Empathy Assessment Methods CBT (TEAM-CBT) [26-29]. TEAM-CBT differs from traditional CBT in four key ways: (1) T=testing, frequent ecological momentary assessment of how users feel in the here-and-now (as opposed to the past 2 weeks) is used to track progress as well as relapses between classes; (2) E=empathy, the use of a variety of validation techniques to make the user feel cared about and understood; (3) A=address resistance, the identification and reduction of motivational process and outcome resistance that nearly always interferes with recovery from depression and anxiety; and (4) M=methods, the use of dozens of techniques that draw from multiple schools of therapy. The techniques are chosen based on the type and categories of problem the user is struggling with. Although TEAM-CBT therapy itself has never been clinically tested, it is considered an evidence-informed therapy incorporating a myriad of evidence-based techniques into its practices.

The current Feeling Good App prototype does not require human intervention and is entirely automated. The current digital TEAM-CBT–based intervention consists of (1) intensive basic training (IBT), which spans 1 to 2 days and interactively teaches principles of cognitive therapy, and (2) ongoing training modules (OTMs) designed to integrate prior lessons and challenge automatic thoughts and core beliefs over 3 to 4 weeks. These training modules include a variety of TEAM-CBT techniques and interactive quizzes to reinforce learning.

Objective

To begin the process of rigorously testing the efficacy, usability, and scalability of this digital mental health intervention, we engaged in a third-party, nonsponsored proof-of-concept analysis of this company's data set. The electronic data were reportedly generated from an internally blinded randomized waitlist crossover trial. This trial's goal was to inform both the app development and help determine whether the technology warranted more resource-intensive sponsored research by academic clinical scientists. This form of research may serve as a model for initial low-risk engagement by academia and industry with minimal conflicts of interest.

The data set was collected during several days of individual digitally delivered intensive basic TEAM-CBT training followed by 4 weeks of additional content to help users challenge distorted thoughts and core beliefs. Using a randomized waitlist crossover data set of a beta test run by and provided to us by the Feeling Good App development team, we examined the safety, feasibility, and acceptability of this digital intervention. In addition, we performed exploratory analyses, examining the evidence for efficacy of this digital app for improving mood symptoms.

Methods

Recruitment

Data were collected by the Feeling Good App Corporation from users who were recruited from a waitlist of several thousand individuals interested in participating in a beta test of a TEAM-CBT mobile app mentioned in one of the Feeling Good podcast episodes hosted by David Burns. A deidentified data set was provided to the Department of Psychiatry and Behavioral Sciences, Standford School of Medicine, via a nonsponsored data use agreement with Feeling Good App Corporation. The Feeling Good App development team retained ownership of the original data and allowed permission through the data use agreement for the authors to analyze and publish the findings independently. All data for this study were collected by the Feeling Good App development team and not by the authors of this study. During data collection, authors consulted with the Feeling Good App development team and shared insight on best practices for creating a waitlist crossover cohort design and types of data to collect from users (eg, demographic information, exclusionary criteria, and measures); however, the authors did not have any direct oversight over its collection or means to assess accuracy.

The Feeling Good App was still in development and not commercially available during the study period. All participants were recruited from a waitlist of individuals who had heard of the app from Burns' social media sites and podcast and requested access. An email invitation was sent to waitlist members, and to those who responded to the email invitation, a consent for use of deidentified data in research and a screening form were emailed. Screening information was collected using a Google Survey. Screened respondents were included if they were aged >18 years, were fluent in English, had regular access to an iPhone, and had no previous experience with the Feeling Good App. They also indicated a willingness to spend at least 4 hours per day on IBT and 30 minutes per day on OTMs. Participants were excluded if they admitted to ever having a suicide attempt, suicidal urges, or plans, or self-harm behavior engagement or urges to harm others. Participants who indicated "yes" to any of these risk and safety questions were excluded from the study at enrollment and encouraged to seek help from a mental health professional and given resources to do so.

Ethical Considerations

Use and analysis of the data set was approved by the Stanford institutional review board (protocol ID 67718; Palo Alto, CA).

Intervention

After recruitment and screening, participants obtained access to the contents of the app by invitation to download the iOS app onto their Apple device by the app developers. Participants were then given access to the Feeling Good App, version title "Basic Training Release 2 RCS," a CBT self-help smartphone app designed to reduce a variety of negative feelings. It was not described as a treatment for any mental disorder or as a substitute for professional treatment, and at sign-up, participants were given resources to pursue professional treatment if desired. The app focused on identifying and modifying negative thoughts and beliefs associated with depression and anxiety and was entirely automated with no human intervention. The app used a 2-phased intervention model, with IBT spanning from 1 to 2 days and then OTMs that required 15 to 30 minutes a day for 4 weeks. The IBT contained modules on cognitive distortions, positive refraining, and relapse prevention and provided psychoeducation in the cognitive behavioral model of mental health disorders. The OTMs taught user strategies to challenge and change distorted thoughts and beliefs.

Participants were randomly assigned to an immediate start (IS) group or a waitlist control (WC) group in a crossover design. This study did not include a true control group, as all participants used the Feeling Good App and were included in the data set. Randomization was conducted using the percentage rollouts randomization feature in a software called LaunchDarkly [30]. The IS group was asked to select a day when they could devote

4 hours to the app for 2 consecutive days to start IBT and were given access to the app. The WC group was also asked to select a day to start IBT after a 4-week waiting period. Each week, participants completed self-report measures. Participants were assessed frequently throughout the beta test, but for this analysis, only four time points were reported: (1) *preintervention time point*, baseline at recruitment from the app waitlist; (2) *start of basic training (SBT)*, at the start of IBT; (3) *end of basic training (EBT)*, completion of IBT; and (4) *postintervention time point*, at the end of ongoing OBT and 4 weeks after SBT.

Measures

Overview

At the preintervention time point, all participants completed a survey that provided information about the app as well as complete demographic information. The preintervention information was gathered using a Google Survey that was securely collected and monitored by the Feeling Good App development team by their report. All other measures were collected electronically through the digital app. When participants were recruited from the waitlist at SBT, they completed a variety of self-report measures, including the Patient Health Questionnaire-9 (PHQ-9), the Generalized Anxiety Disorder-7 (GAD-7), and 7-Dimension Emotion Sliders (7DES). At EBT, participants completed the 7DES. At the postintervention time point, participants completed the PHQ-9, GAD-7, and 7DES. According to the Feeling Good App development team, users interacted with the service via browser at the time they applied to be in the beta test and in the app during the duration of the beta test. During that time, the information the users sent to the Feeling Good App servers was encrypted over HTTPS. The Feeling Good App servers stored information in a Postgres database. The connection between the servers and the database was encrypted. The database application was hosted in a private stack that did not allow connections except from application servers. The database used encryption at rest so that if an attacker was able to obtain the database, they would not be able to access the underlying information. In addition, all identifying information, such as names and email addresses, was encrypted by the application server, so that in the event an attacker was able to access the database they would not be able to match any data to user identities. All data were deidentified before being transferred securely via Box to the authors.

Demographic Measures

Demographic information gathered at sign-up included age, gender, income, marital status, education, and race. Clinical information was gathered from all participants, which included current engagement in psychotherapy or psychotropic medication (at sign-up). All demographic data were collected using a Google Survey at sign-up.

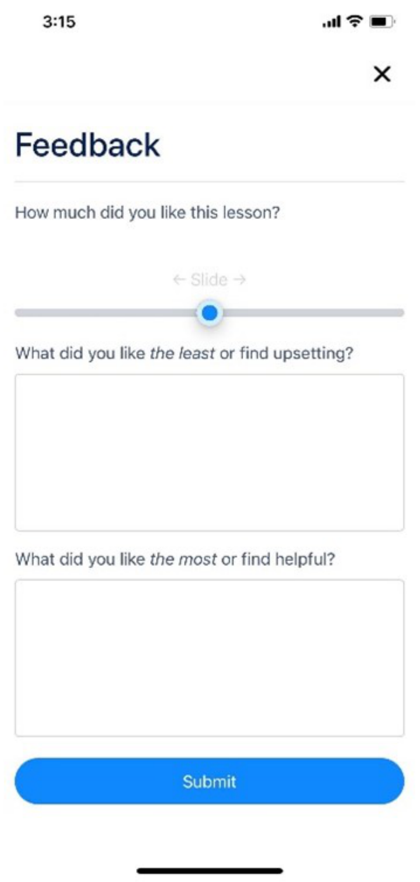
Safety

During the preintervention screening, any participants who indicated a “1” or higher on item 9 of the PHQ-9 (“Thoughts that you would be better off dead or of hurting yourself in some way”) were excluded from the beta test and given resources for support. The preintervention screening also included 4 questions: “Have you ever made a suicide attempt in the past?” “Have you ever struggled with suicidal plans or urges?” “Have you ever engaged in self-harm behaviors, like burning yourself, cutting yourself, and so forth?” “Do you sometimes have urges or plans to harm others?” Any individuals who responded positively to these questions were excluded and given resources for support. No other safety data collection or monitoring was performed.

Qualitative Feedback

Throughout the beta test, qualitative feedback was collected at the end of every exercise (58 times in total for each individual who completed the full intervention) where participants were asked (1) “What did you like least or find upsetting?” and (2) “What did you like the best?” (Figure 1). Qualitative feedback was not reviewed by the development team until after the conclusion of the study.

Figure 1. Screenshot of qualitative feedback prompts within the app.



PHQ-9 Scale

The PHQ-9 [31] is a self-assessment tool that evaluates the degree to which one is experiencing each of the symptoms of depression. This tool uses a Likert scale ranging from 0 to 3 (*not at all* to *nearly every day*) to measure the severity of depression [31]. Summed scores range from 0 to 27, with higher scores indicating greater symptom severity. The PHQ-9 has strong construct validity and reliability as a measure of depression symptom severity [31,32]. Responses were collected via Google Survey during sign-up, and all subsequent responses were collected within the app.

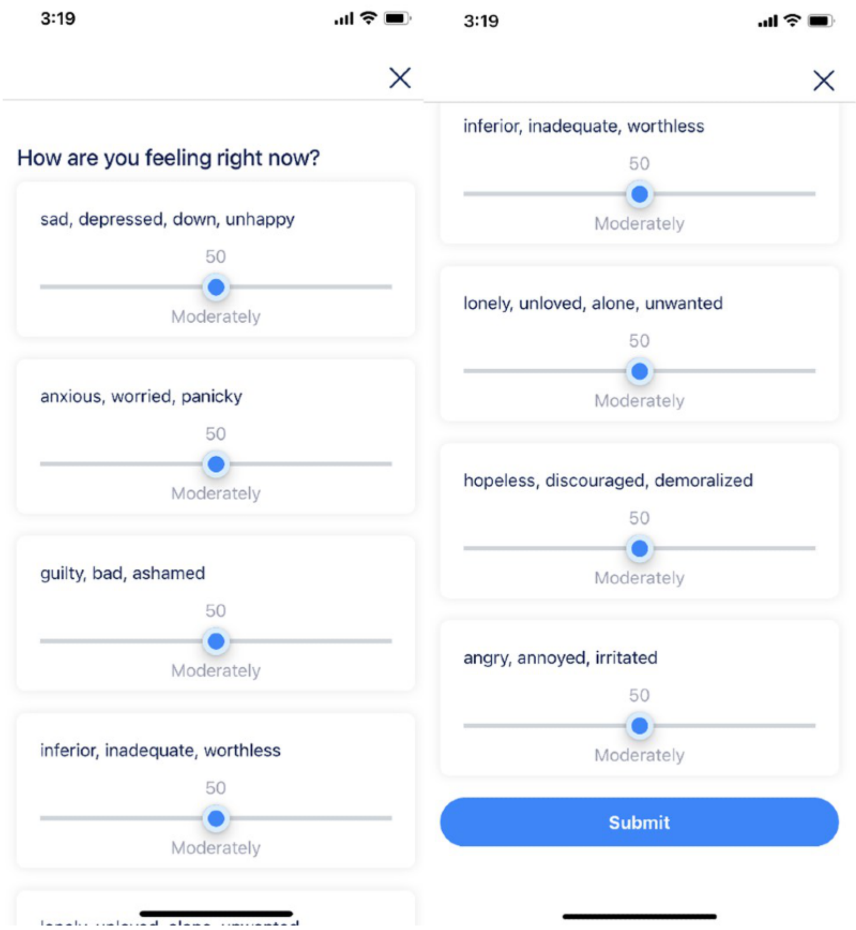
GAD-7 Scale

The GAD-7 [33] is a widely used, freely available self-assessment tool that evaluates the severity symptoms of GAD. This tool uses a Likert scale ranging from 0 to 3 (not at all to nearly every day) to measure the severity of anxiety symptoms. Summed scores range from 0 to 21, with higher scores indicating greater symptom severity. The GAD-7 has strong validity [33,34], internal consistency and convergent validity [34,35], and validity as a screening measure for symptom severity [36]. Responses were collected via Google Survey during sign-up, and all subsequent responses were collected within the app.

7DES Scale

The Feeling Good App was designed to produce and track extremely rapid changes in mood. As a result, conventional scales such as the PHQ-9 are not able to measure these types of rapid changes since they ask about the last 2 weeks. Developed for the purpose of the app, the 7DES scale is an ecological momentary assessment [37], which asks users to evaluate on a scale from 0 to 100 the extent to which they are experiencing the indicated negative feelings (Figure 2). These sliders allow for specific report of symptoms, as they change throughout the use of the app. Participants were prompted with the question of "How are you feeling right now?" and asked to move a slider (scale of 0 to 100) to reflect their experience of the following emotions: anger, anxiety, depression, guilt, hopelessness, inferiority, and loneliness. Participants were also asked to complete a single prompt slider rating their current happiness (happiness slider). Participants were prompted with "How happy are you right now?" and provided a 0 to 100 slider to respond (Figure 2). This study examined the data of the full-scaled 7DES-depression (7DES-d) of the 7DES; the anxiety scale (7DES-anxiety) of the 7DES; as well as data from the 5 emotion scales most commonly associated with depression (7DES-composite; depression, guilt, hopelessness, inferiority, and loneliness). Before our analysis, this data set showed that the reliability of this scale at the start of the app was 0.94. The 7DES scale is highly correlated to the PHQ-9 measure (r=0.67; P<.001). All 7DES responses were collected within the app.

Figure 2. Screenshot of 7-Dimension Emotion Sliders within the app.



Data Analysis

All quantitative data were analyzed using SPSS Statistics (version 28.0; IBM Corp). Descriptive statistics were calculated for demographic variables. Feasibility and acceptability were descriptively assessed using the engagement diagram and dropout tables.

Safety was assessed via qualitative and numerical reports provided by Feeling Good App Corporation team developers. Safety was measured quantitatively in 2 ways. First, we directly measured suicidal ideation using item 9 from the PHQ-9, which asks about the frequency of suicidal thoughts over the past 2 weeks. In addition, we measured suicidal ideation indirectly by using the hopelessness slider from the 7DES, which asks about the user's current feelings of hopelessness. Past research has shown that hopelessness independently predicts suicidal ideation, even when controlling for depression [38,39]. Therefore, we were able to assess safety both directly and indirectly, analyzing changes over time using repeated measures ANOVA.

Efficacy was evaluated using repeated measures ANOVA, which identified changes over time (SBT to postintervention) and by group (IS and WC). The PHQ-9, 7DES-d, and 7DES-composite scores were used to measure depression, while the GAD-7 and 7DES-anxiety scores were used to measure anxiety. Finally, we tested whether baseline characteristics of participant demographics predicted intervention response for both the IS and WC groups by running univariate ANOVA, where the demographic variables were moderators for changes in 7DES scores.

Results

Overview

The demographic variables of age, gender, race and ethnicity, highest level of education, income, marital status, current engagement with therapy, and current engagement with psychopharmacology were used. Descriptive statistics were calculated for each group using chi-square test. As shown in Table 1, there were no statistically significant differences between groups for any demographic variable.

Table 1. Demographics by group.

Characteristics	Immediate start group (n=141)	Waitlist control group (n=149)	Chi-square (*df*)	*P* value
Age (years), mean (SD)	45.4 (11.91)	44.05 (12.96)	46.15 (1)	.82
Gender, n (%)			0.01 (2)	.95
Male	57 (40.4)	61 (40.9)		
Female	80 (56.7)	87 (58.4)		
No reply	4 (2.8)	1 (0.7)		
Race, n (%)			6.02 (9)	.65
African American	3 (2.1)	6 (4)		
Asian (eastern)	5 (3.5)	8 (5.4)		
Asian (Indian)	3 (2.1)	3 (2)		
Hispanic	8 (5.7)	6 (4)		
Mixed race	3 (2.1)	5 (3.3)		
Native American	1 (0.7)	0 (0)		
White	105 (74.5)	114 (76.5)		
Other	8 (5.7)	3 (2)		
No reply	5 (3.5)	4 (2.7)		
Highest education level, n (%)			1.89 (5)	.87
Grammar or middle school	0 (0)	1 (0.7)		
High school	2 (1.4)	4 (2.7)		
Some college or technical training	13 (9.2)	13 (8.7)		
College degree	26 (18.4)	25 (16.8)		
Some graduate school	5 (3.5)	7 (4.7)		
Graduate degree	85 (60.3)	89 (59.7)		
Income (US $), n (%)			3.91 (8)	.87
0-25,000	9 (6.4)	9 (6)		
25,000-50,000	21 (14.9)	21 (14.1)		
50,000-75,000	16 (11.3)	21 (14.1)		
75,000-100,000	24 (17)	25 (16.8)		
100,000-125,000	14 (9.9)	19 (12.7)		
125,000-150,000	14 (9.9)	18 (12.1)		
150,000-175,000	10 (7.1)	7 (4.7)		
175,000-200,000	6 (4.2)	9 (6)		
>200,000	27 (19.4)	20 (13.4)		
Marital status, n (%)			9.24 (7)	.24
Divorced	6 (4.2)	10 (6.7)		
Living together	4 (2.8)	12 (8)		
Married	77 (54.6)	86 (57.7)		
Other	2 (1.4)	3 (2)		
Separated	1 (0.7)	2 (1.3)		
Single (not partnered)	39 (27.7)	30 (20.1)		
Single (partnered)	11 (7.8)	5 (3.3)		
Widowed	1 (0.7)	1 (0.7)		

Characteristics	Immediate start group (n=141)	Waitlist control group (n=149)	Chi-square (*df*)	*P* value
Currently in therapy, n (%)			0.12 (1)	.73
No	101 (71.6)	104 (70)		
Yes	40 (28.4)	45 (30)		
Taking medication, n (%)			0.66 (1)	.42
No	104 (73.8)	116 (78)		
Yes	37 (26.2)	33 (22)		

Safety

Overall, 207 (40.2%) out of 515 individuals were excluded at screening for suicidality endorsing suicidal urges, suicidal thoughts, and past suicide attempts. For participants of the intervention, on item 9 of the PHQ-9 ("Thoughts that you would be better off dead or of hurting yourself in some way"), there was not a statistically significant change in scores from the SBT to postintervention time point (t=0.0, P>.99). Although users with suicidal thoughts were excluded from the study at the initial evaluation, at the start of app use, 10 (3.4%) of the 290 users endorsed a 1 (several days over the past 2 weeks) on item 9, while 2 (0.7%) of the 290 users endorsed a 2 (more than half the days). At the end of the intervention, 1 (0.3%) user endorsed a 1 (several days) and 1 (0.3%) user endorsed a 2 (more than half the days).

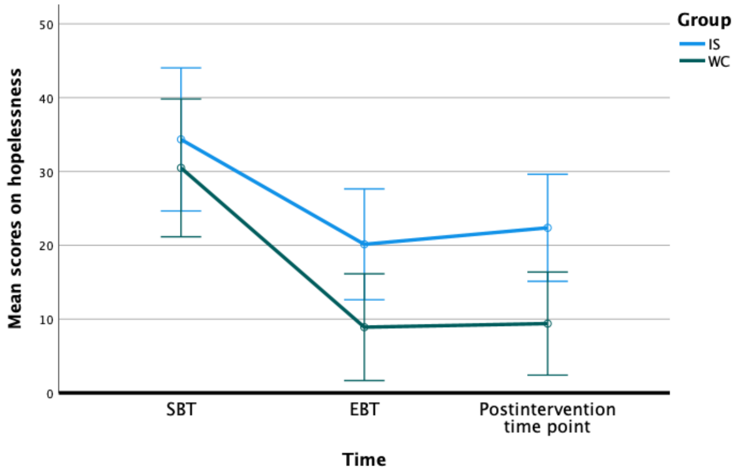
Throughout the beta test, qualitative feedback was collected at the end of every exercise (58 times in total) where participants were asked "What did you like least or find upsetting?" Qualitative feedback was monitored, and no participants reported endorsed safety concerns throughout the entirety of the beta test. In addition, we measured changes in hopelessness from the 7DES. There was a statistically significant decrease in hopelessness from the SBT to the postintervention time point in both groups (F_{289}=30.16, P<.001), and there was not a statistically significant difference in the hopelessness change between the IS and WC groups (F_1=1.78, P=.18; see Table 2 for engagement rates and Figure 3 for changes in hopelessness).

Table 2. Engagement rates from preintervention to postintervention time point.

Time point	Total, n	SBT[a], total (%)	IS[b], n	IS % of SBT	WC[c], n	WC % of SBT
Invited to join beta test	4466	N/A[d]	N/A	N/A	N/A	N/A
Started preintervention screening (preintervention time point)	612	N/A	N/A	N/A	N/A	N/A
Completed preintervention screening—enrolled	515	N/A	N/A	N/A	N/A	N/A
Accepted to beta test—randomized	290	N/A	141	N/A	149	N/A
Started intensive basic training (SBT)	230	N/A	113	N/A	117	N/A
Completed intensive basic training (EBT[e])	166	72.2%	82	72.6%	84	71.8%
Completed ongoing training module (postintervention time point)	80	34.8%	38	33.6%	42	35.9%
Completed follow-up survey	63	27.4%	27	23.9%	36	30.8%

[a]SBT: start of basic training.
[b]IS: immediate start.
[c]WC: waitlist control.
[d]N/A: not applicable.
[e]EBT: end of basic training.

Figure 3. Changes on hopelessness scores from the start of basic training (SBT) to postintervention time point by group. EBT: end of basic training; IS: immediate start; WC: waitlist control.



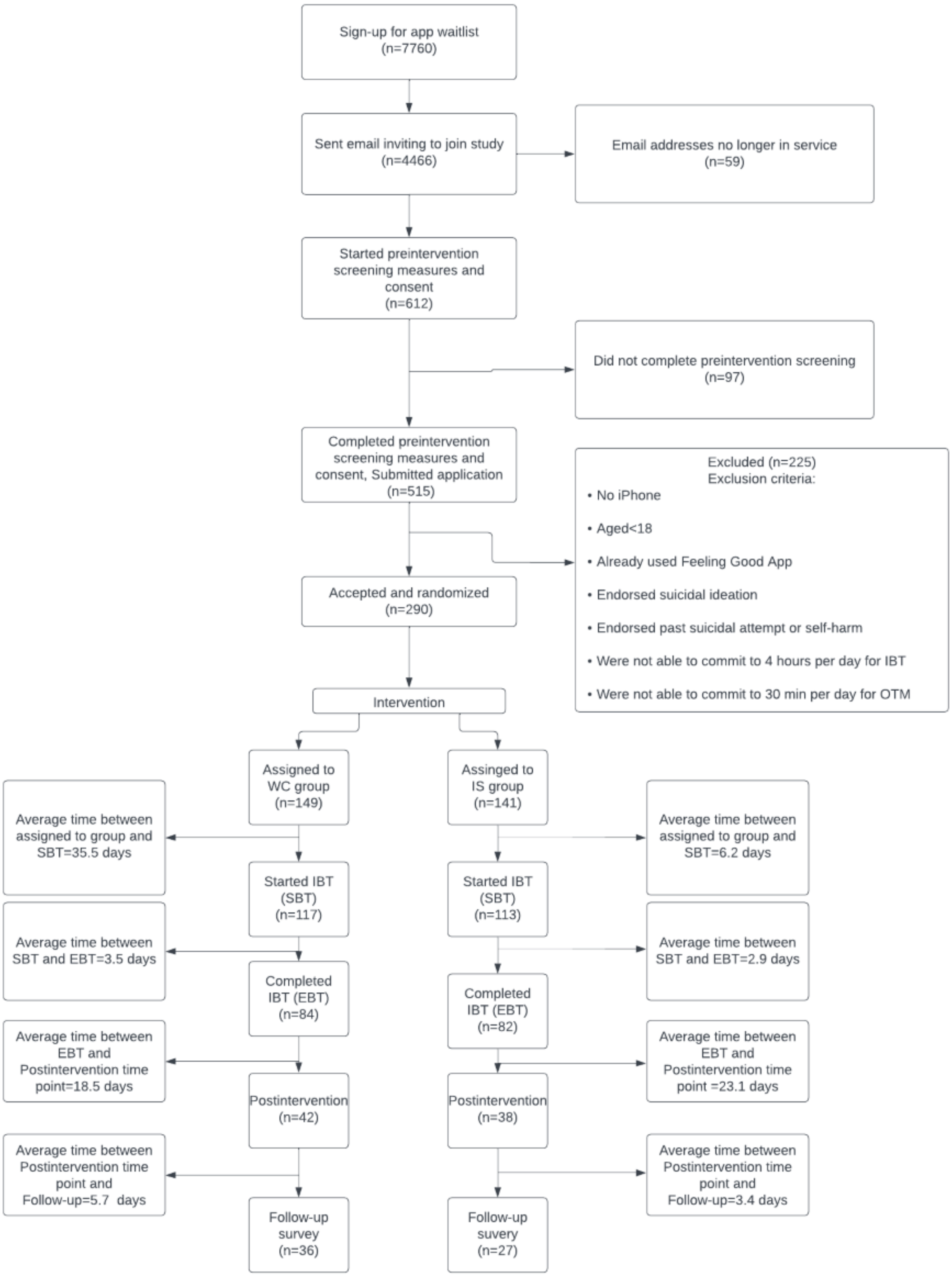
Feasibility and Acceptability

Overview

Invitations to join the study were sent to 4466 individuals who were randomly chosen from beta test waitlist using the percentage rollouts feature on LaunchDarkly. Of those who were invited to join the study, 612 (13.7%) individuals accepted the preintervention screening measures. Overall, 515 individuals completed the preintervention screening process, and 225 were excluded from participation.

An initial group of 290 individuals were enrolled to participate in the study and completed preintervention screening measures and consent forms. Of them, 141 (48.6%) participants were randomly assigned to the IS group and 149 (51.4%) participants were assigned to the WC group. Furthermore, 113 (80.1%) participants who were assigned to the IS group and 117 (78.5%) participants in the WC group started IBT. Initial enrollment dropout rates for the IS group were 19.9% (28/141) and 21.5% (32/149) for the WC group. Data were not collected about the reason for dropout. Moreover, 27% (38/141) of participants in the IS group completed the postintervention survey, while 28.2% (42/149) of participants in the WC group completed the postintervention survey. A total of 223 participants started IBT, 166 (74.4%) participants completed IBT, and 80 (35.8%) participants completed OTMs. Dropout rates from SBT to EBT were 27.4% (31/113) for the IS group and 28.2% (33/117) for the WC group. From the EBT to postintervention time point, the IS group had a dropout rate of 54% (44/82), while the WC group had a dropout rate of 50% (42/84). Over the 4-week period that users engaged with the app, there was a total dropout rate of 66.4% (75/113) for the IS group and 64.1% (75/117) for the WC group. In total, of the 290 participants who were randomly assigned to groups, 166 (57.2%) users completed IBT, while 80 (27.6%) users completed the 4-week OTMs (see Table 2 for engagement rates and Figure 4 for CONSORT [Consolidated Standards of Reporting Trials] diagram).

Figure 4. Engagement diagram. IBT: intensive basic training; IS: immediate start; OTM: ongoing training module; WC: waitlist control.



Qualitative Feedback

Qualitative feedback, gathered at the end of each exercise throughout IBT and OTMs, was reviewed to identify content-based themes. One prominent theme that emerged pertains to users' expectations and engagement with the app's structure and pacing. Some users expressed satisfaction with its user-friendly interface, while others raised concerns about its usefulness in addressing intricate psychological challenges. This theme underscores the significance of aligning app design with users' diverse needs and preferences. Many users expressed appreciation for the app's educational value, finding tremendous benefit from the lessons, tools, and techniques offered. However, the theme of personalization resonates strongly in users' comments, highlighting their desire for content tailored to their unique experiences.

Technical glitches and functional concerns represent another notable theme. Users' reports of technical issues underscore the need for rigorous quality assurance to ensure smooth app functionality. This theme also emphasizes the critical role of seamless user experiences in maintaining engagement and fostering positive change in how they are feeling. Furthermore, users' reflections on the app's engagement dynamics revealed its potential to evoke hope and encourage proactive self-care behaviors. Users shared hopeful sentiments about their ability to apply learned techniques from the app into their daily life, and many users looked forward to a commercial release of the app to share with others.

Some users highlighted areas where the app's pacing and content may not fully align with their preferences, suggesting the need for dynamic content delivery strategies that accommodate varying user needs. One area of concern for many users was the time commitment expected for the completion of the IBT portion of the modules. Results from the data analysis showed that the IBT brings forth some of the most foundational and persistent decreases in symptoms although users report difficulty in committing to the app's daily requirements of multiple hours.

Upon completing the beta test, many users reflected on the personal growth and positive changes that they experienced throughout both IBT and OTMs. Users expressed a sense of pride and accomplishment and shared how they found the app to be incredibly effective in solidifying their understanding of CBT concepts and skills. Many users reported anticipating using the app again in the future, signaling a desire to continued engagement among users.

Efficacy

PHQ-9 Scale

We examined the possible effect of the Feeling Good App intervention on depression scores as measured by the PHQ-9, a standardized measure of depression. Using a repeated measures ANOVA, we found that there was a statistically significant within-groups effect of time, where mean depression scores measured by the PHQ-9 decreased from sign-up to 4 weeks later (F_{197}=31.30, P<.001). In addition, there was a between-group effect (F_1=19.88, P<.001). As shown in Figure 5, the IS group saw significant changes (P<.001) on the PHQ-9 scores over 4 weeks, while the WC group did not (P=.20).

We then used repeated measures ANOVA to assess changes in PHQ-9 scores from the SBT to the end of the app intervention for both groups. There was a statistically significant difference in PHQ-9 scores from the SBT through the end of the app intervention (F_{78}=43.680, P<.001). Figure 6 shows the change over time for each group, where the WC group saw significant reductions (P<.001) in PHQ-9 scores after starting the app intervention.

Figure 5. Change in Patient Health Questionnaire-9 (PHQ-9) scores over 4 weeks. IS: immediate start; WC: waitlist control.

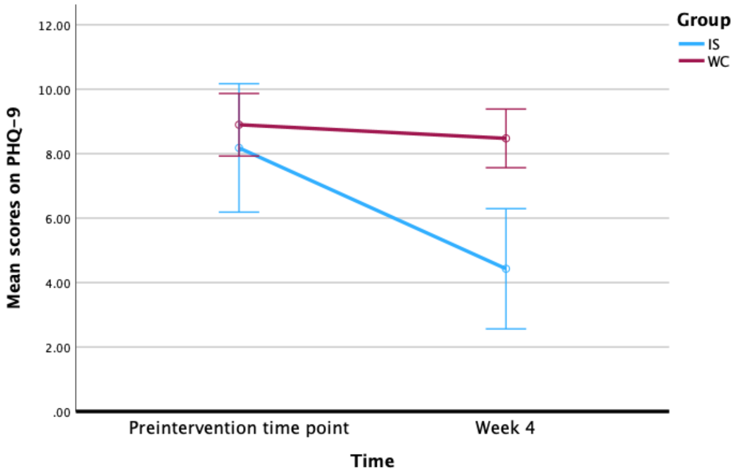
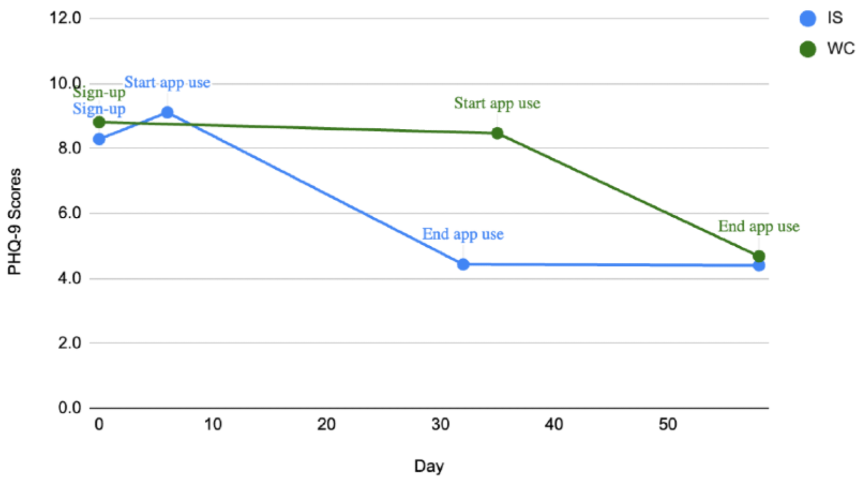


Figure 6. Changes in Patient Health Questionnaire-9 (PHQ-9) from the start of basic training (SBT) to postintervention time point. IS: immediate start; WC: waitlist control.



7DES Scale

In the exploratory analysis, we examined the effect of the Feeling Good App intervention on depression scores as measured by the novel 7DES-d and 7DES-depression measure composite. These slider scales provide information about the severity of current depression symptoms. A repeated measures ANOVA was conducted to measure the changes in depression scores from the SBT to postintervention time point. To account for the participant dropout, analyses for change in time excluded participants who did not complete the measures at both time points. There was a statistically significant change in depression scores throughout the use of the app, where the 4-week Feeling Good App had a statistically significant effect on the 7DES-d (F_{78}=30.508; P<.001) and 7DES-composite (F_{78}=34.36; P<.001).

Bonferroni adjustment pairwise comparisons were used to measure the change in depression scores between each time point. Table 3 describes the score changes on the 7DES-composite and 7DES-d scores across the intervention period. On the basis of these results, we can conclude that the reductions in depression scores on the 7DES scale occurred during IBT and were maintained throughout the app intervention for both groups. We then measured whether this change over time was different between the groups. In the repeated measures ANOVA, there was not a statistically significant interaction between time and group (F_{78}=0.81; P=.44). Therefore, the group did not impact the changes over time in depression scores, where both the IS and the WC groups saw significant (P<.001) decreases in depression scores. Figures 7 and 8 show the change over time by group.

We also conducted a repeated measures ANOVA to measure the changes in scores on the GAD-7 over time. We found that there was a significant within-participants change over time (F_{78}=38.99; P<.001) and a nonsignificant between-participants effect of group (F_1=0.64; P=.53). Using Bonferroni adjustment pairwise comparisons, we found that there was a statistically significant reduction in GAD-7 scores from SBT to postintervention time point for both the IS (mean change 21.14, P<.001) and WC (mean change 17.78, P<.001) groups (Figure 9).

We then measured how the participants' anxiety changed over time as measured by the 7DES-anxiety. Using a repeated measures ANOVA, we found that there was a significant within-participants change based on time (F_{78}=50.58; P<.001) and that there was not a significant between-participants effect of group on this change (F_1=1.03; P=.37). Bonferroni adjustment pairwise comparisons were used to measure the change in 7DES-anxiety scores between each time point. There was a statistically significant difference on the 7DES-anxiety from the SBT to EBT for both IS (mean change 15.92, P<.001) and WC (mean change 20.00, P<.001) groups (Figure 10).

Table 3. Score changes on the 7-Dimension Emotion Sliders (7DES)-composite and 7DES-depression (7DES-d) over time.

Measure	Change across time	P value
7DES-composite		
SBT[a] to EBT[b]		
IS[c]	–13.71	<.001
WC[d]	–17.29	<.001
EBT to postintervention time point		
IS	1.26	>.99
WC	0.27	>.99
7DES-d		
SBT to EBT		
IS	–14.21	<.001
WC	–16.59	<.001
EBT to postintervention time point		
IS	0.66	1.0
WC	1.24	1.0

[a]SBT: start of basic training.

[b]EBT: end of basic training.

[c]IS: immediate start.

[d]WC: waitlist control.

Figure 7. Changes in 7-Dimension Emotion Sliders-depression (7DES-d) scores from preintervention to postintervention time point. IS: immediate start; WC: waitlist control.

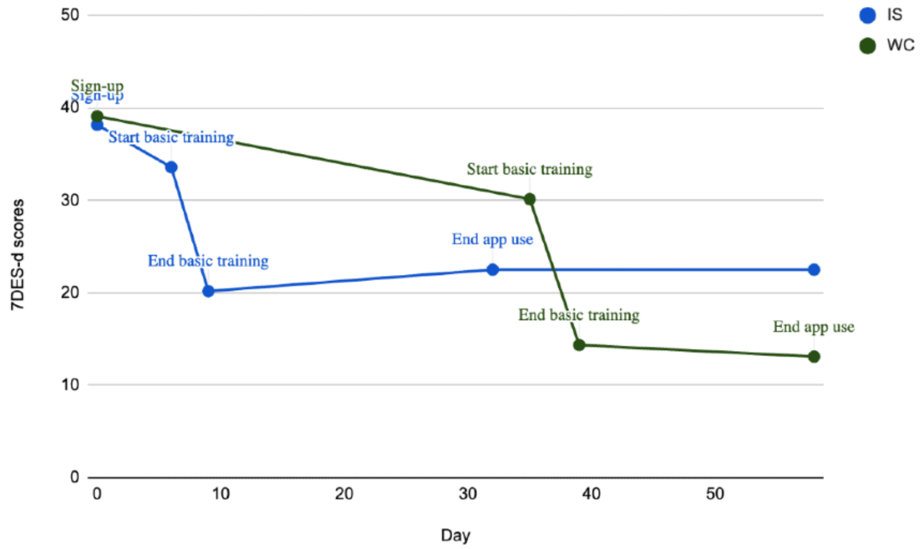


Figure 8. Changes in 7-Dimension Emotion Sliders (7DES)-composite scale scores from preintervention to postintervention time point. IS: immediate start; WC: waitlist control.

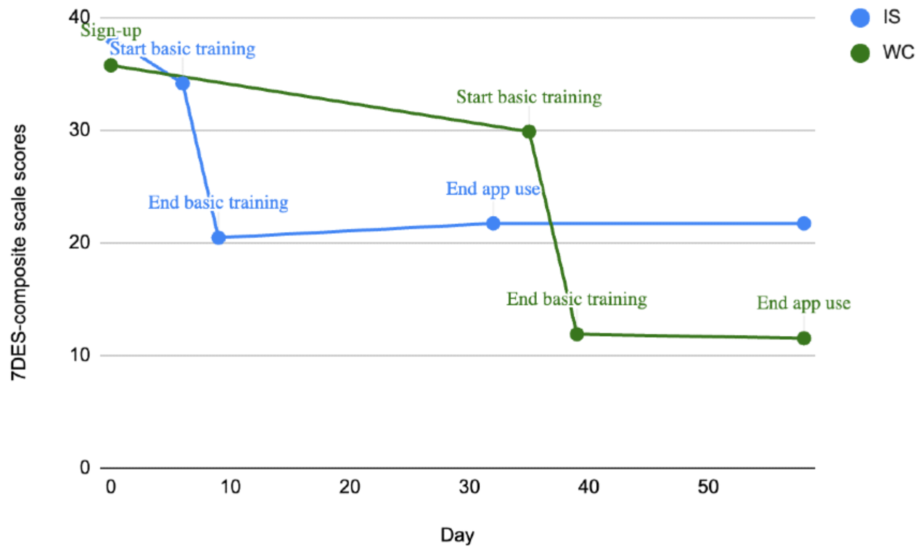


Figure 9. Changes in Generalized Anxiety Disorder-7 (GAD-7) scores from preintervention to postintervention time point. IS: immediate start; SBT: start of basic training; WC: waitlist control.

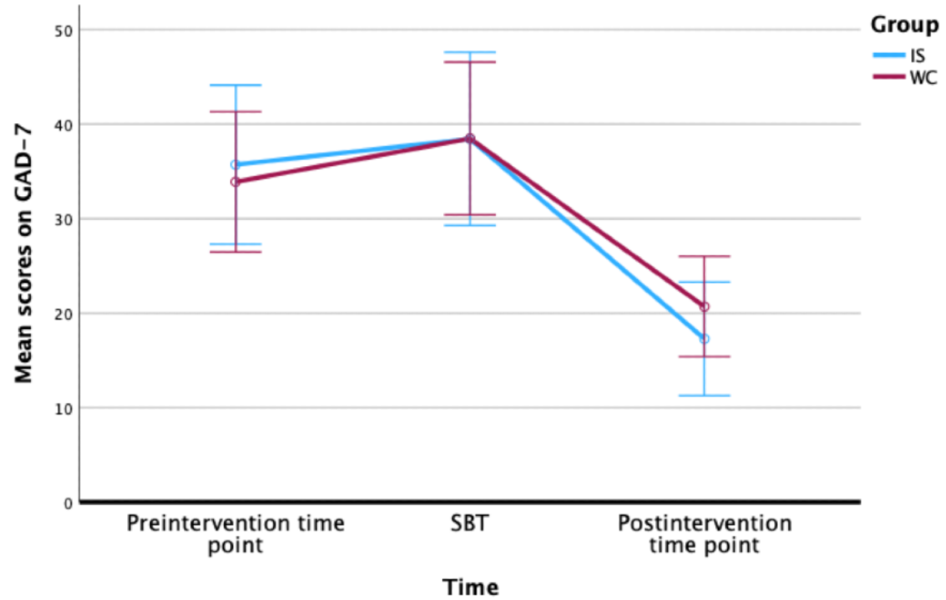
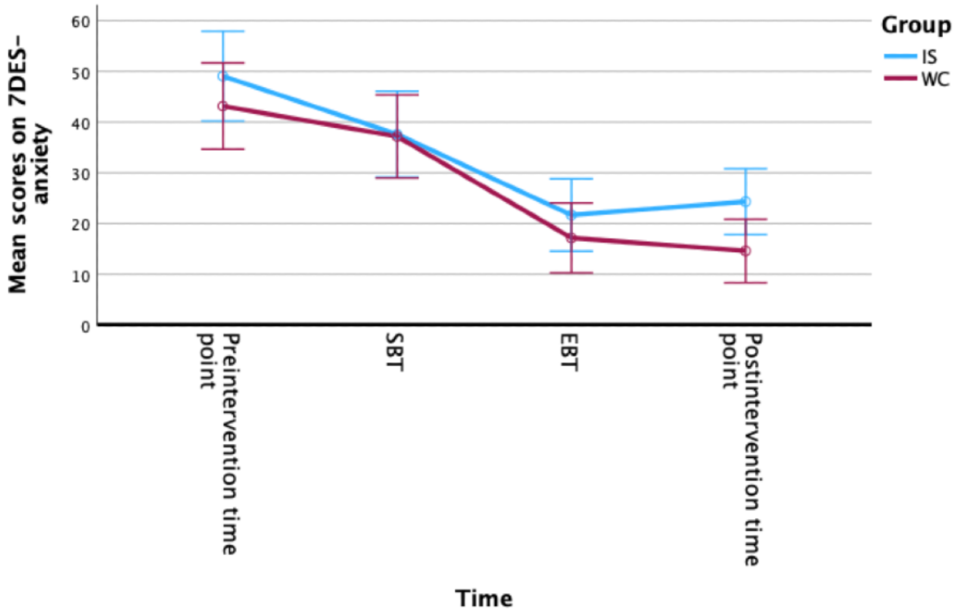


Figure 10. Changes in 7-Dimension Emotion Sliders (7DES)-anxiety item level scores from preintervention to postintervention time point. EBT: end of basic training; IS: immediate start; SBT: start of basic training; WC: waitlist control.



Demographics as Moderators

Finally, we examined the effect of demographic variables on changes in depression scores on the 7DES over time. A univariate ANOVA was used to examine the effect of moderator variables on depression change scores. There were not statistically significant interactions between depression change scores from SBT to EBT and age (F_{56}=1.11; P=.35), gender (F_2=0.02; P=.89), race (F_9=0.69; P=.63), income (F_8=0.80; P=.60), education level (F_5=0.97; P=.43), marital status (F_7=1.80; P=.13), current therapy engagement (F_1=0.22; P=.64), or use of psychopharmacological medications (F_1=0.10; P=.76).

Discussion

Principal Findings

This study is the first academic and reported proof-of-concept evaluation of the Feeling Good App in terms of safety, feasibility, and initial reports of efficacy. It demonstrates a feasible and novel approach to industry and academic collaboration in the process of developing a digital mental health technology translated from existing evidence-informed treatment. This study may be able to serve as a model for future low-risk, low resource-intensive, minimal conflict-of-interest pathways to conduct initial nonsponsored research with various stakeholders. The results support the initial Feeling Good App prototype as most likely safe for users, having acceptable levels of engagement throughout the length of the intervention and evidence of use correlating with significant decreases in symptoms of depression and anxiety for those with moderate to low levels of clinical symptoms of depression.

The qualitative and quantitative measures of safety demonstrated evidence that this app is generally safe to use when excluding users with current safety concerns before use. Safeguards such as providing resources to individuals who indicated any amount of suicidal ideation on the PHQ-9 and encouraging all users to report concerns of safety allowed for users to proceed with the app without significant safety concerns arising throughout the duration of study. Further consideration of safety for users must be considered as the app moves toward commercial use. These may include protocols in place for allowing those with current or past safety concerns to use and benefit from the app since suicidal behaviors and parasuicidal are unfortunately common in those with mood disorders.

Measures of acceptability demonstrate varying points at which users cease use of the app, indicating that improving the app's engagement could benefit its widespread usability. App-based mental health tools are often affected by high dropout rates, particularly those that target depression symptoms. Dropout rates for these depression symptom apps tend to fall approximately 47.8%, and apps with smaller samples and more individualized feedback have lower rates on average than larger, generalized samples [34]. Dropout rates in the Feeling Good study can be considered after basic training, when a large percentage of total improvement has already taken place or at the end of the study and follow-up, which would put it above those reported average dropout rates. The dropout rates through the EBT for both groups combined were 28%, which can be attributed to the IBT's high engagement value and low-time commitment of 1 to 2 days. Over the 4-week period that users engaged with the app, there was a total dropout rate of 66% for the IS group and 64% for the WC group. Further app development may want to focus on improving this dropout, as it is a common problem currently with all forms of treatment for depression and especially for automated digital app– or web-based treatments.

The qualitative data analysis illuminates several salient content-based themes that underscore the multifaceted nature of users' experiences with the mental health app. These themes encompass users' expectations, personalization needs, technical functionality, and engagement dynamics. The findings underscore the importance of user-centered design in creating effective digital mental health interventions. The themes identified suggest the potential for integrating advanced personalization techniques, such as artificial intelligence algorithms, to enhance the app's relevance and resonance with individual users.

Retrospective analysis of this industry-generated existent data set demonstrates a statistically significant (P<.001) reduction in depression scores on the PHQ-9 from the beginning of the app to the end of the initial training and to the end of the ongoing modules. However, it is important to note that the cohort was not recruited from a clinical setting, and the mean scores before and after were indicative of mild to moderate depressive symptoms. This study does not fulfill the requirements of a true efficacy study, rather it includes initial statistical impressions interpreted from the data set. The initial training took place over just 2 days, representing the possibility of rapid improvement. Scores from the 2 groups were not significantly (P=.41) different at the start of the intervention, so the delayed reduction in depression scores for the WC group aligned with their completion of the app, indicating that changes over time were attributed to app use.

Each of the scales from the 7DES (anger, anxiety, guilt, loneliness, inferiority, and hopelessness) saw significant (P<.001) reductions throughout the duration of the initial training. These reductions were maintained through the end of ongoing modules and the users' time using the app. There were no significant reductions between the end of the initial training and the end of the ongoing modules, suggesting the possibility that the app has the potential to maintain decreased depressive and negative emotions feelings over time.

Reductions as seen through the depression slider 7DES specific to the app were consistent with reductions on the PHQ-9, a popular well-validated assessment tool for depression symptoms. This may support efforts and further exploration of this measure to verify whether it is a clinically and ecologically momentarily valid and reliable measure to be used.

These findings of reduced PHQ-9 and 7DES scores represent potential for this app to help individuals reduce symptoms of depression and associated factors of anger, anxiety, guilt, loneliness, inferiority, and hopelessness. These findings should be further investigated in future studies due to the clinical importance of an app that can help individuals increase their mental well-being.

Individuals from the WC group did not see a significant (P=.20) reduction of depressive feelings during the 4-week waiting period. This was surprising, as many times during studies of WCs, there were some improvements seen before beginning an intervention, which may represent expectation, regression to the mean or passage of time, and increased hope for improvement. Research on WC design shows highly variable effects, which depend on the characteristics of sample being studied [40]. One explanation for this lack of waitlist expectation effect in our analysis may be some type of selection bias such as dropout from those individuals who felt their mood improved during the waiting period and chose not to participate. While reasons for dropping out at assignment were not collected, dropout may have been due to remission of symptoms, lack of desire to engage with the program, or technological barriers.

Limitations

Inherent in any clinical research is the risk of selection bias. The app is currently available by sign-up and is likely reaching audiences familiar with the work of Dr Burns. This bias toward individuals with knowledge related to CBT and psychotherapy may have a moderating effect on the reduction of negative feelings. Future research should take place in naturalistic settings with data from more representative populations perhaps in clinical populations or across cultural and diverse demographic contexts worldwide. The app is not publicly accessible through the Apple App Store and is only available to individuals who are invited to participate in the beta version of the app. With the intention to provide resources at a low cost, these findings should consider the accessibility associated with the app's current stage of development. In addition, the study population was not a clinical population and showed only moderate severity levels of depression per PHQ-9.

This study occurred over a relatively short period with few follow-up data. One of the primary themes within the qualitative feedback was users indicating a desire to continue to use the skills and lessons provided by the app in the future. Further studies could examine the impact of these modules at a consistent time period beyond the completion of the initial training.

Completion of the Feeling Good App involves active attention and effort, and individuals' progressions and effort through the completion of the app were not monitored within this analysis. The app can request its users to engage mindfully but cannot account for potential lapses in attention that result in not engaging with the product in its intended context. User retention and engagement are critical to the success of a mobile app–based intervention. While the development team is capable of monitoring engagement, these data were not analyzed in our study. Further research should aim to examine the relationship between time spent in app and modules completed to determine the potential relationship with symptom reduction.

The participants in this study were all given the Feeling Good App and associated assessments, so this study did not include a control group. Future studies that explore the app and its efficacy should compare the users with a control group that is not given access to the app.

Privacy issues were not evaluated, and further assessment of this app using the American Psychiatric Association's evaluation model and guidelines is encouraged and should include accessibility, privacy, security, clinical foundation, engagement style, and therapeutic goal [41].

Finally, the data analyzed were generated from an industry partner, and there was no way to ensure the data's accuracy by a third party. Further design and clinical testing will need to have the data collection occur by a third party to improve the level of evidence needed to establish efficacy and effectiveness.

Conclusions

The following are the summarized conclusions of this study:

- A low-risk, resource-sparing, nonsponsored, and unfunded collaboration between industry and academia is feasible and can occur during the development of an evidence-informed digital mental health app.
- The Feeling Good App in its current form was found acceptable and feasible to deliver to mild to moderately depressed users familiar with TEAM-CBT in a nonclinical setting.
- Safety appears reasonable but may need improvements in assessment, monitoring, and contingency management. Some safeguards are currently in place, while some safety improvements, such as monitoring for suicidal ideation, are needed with commercial use and with clinical populations.
- Acceptability: high dropout scores compared with other apps indicate some improvement in engagement needed.
- Although not designed to test efficacy, this study indicates that those who stay engaged with the Feeling Good App had significant (P<.001) reductions in the severity of depression symptoms.
- Evidence of feasibility and efficacy thus far supports further testing and development of this app by a third party using a prospective randomized controlled trial.

Conflicts of Interest

None declared.

References

1. Depressive disorder (depression). World Health Organization. URL: https://www.who.int/en/news-room/fact-sheets/detail/depression [accessed 2024-04-05]
2. Lépine JP, Briley M. The increasing burden of depression. Neuropsychiatr Dis Treat 2011;7(Suppl 1):3-7 [FREE Full text] [doi: 10.2147/NDT.S19617] [Medline: 21750622]
3. Demyttenaere K, Van Duppen Z. The impact of (the concept of) treatment-resistant depression: an opinion review. Int J Neuropsychopharmacol 2019 Feb 01;22(2):85-92 [FREE Full text] [doi: 10.1093/ijnp/pyy052] [Medline: 29961822]
4. The state of mental health in America. Mental Health America. URL: https://mhanational.org/issues/state-mental-health-america?eType=EmailBlastContent&eId=071fe8fa-9a08-419f-989b-f062af601066&sl_tc=2022_State_Of_MH_In_America&sl_gr=grp_2022_state_of_mh_in_america_report [accessed 2024-04-05]
5. Rice DP, Miller LS. Health economics and cost implications of anxiety and other mental disorders in the United States. Br J Psychiatry 2018 Aug 06;173(S34):4-9. [doi: 10.1192/s0007125000293458]
6. Insel TR. Assessing the economic costs of serious mental illness. Am J Psychiatry 2008 Jun;165(6):663-665. [doi: 10.1176/appi.ajp.2008.08030366] [Medline: 18519528]
7. Ebert DD, Cuijpers P, Muñoz RF, Baumeister H. Prevention of mental health disorders using internet- and mobile-based interventions: a narrative review and recommendations for future research. Front Psychiatry 2017;8:116 [FREE Full text] [doi: 10.3389/fpsyt.2017.00116] [Medline: 28848454]

8. Baumel A, Torous J, Edan S, Kane JM. There is a non-evidence-based app for that: a systematic review and mixed methods analysis of depression- and anxiety-related apps that incorporate unrecognized techniques. J Affect Disord 2020 Aug 01;273:410-421 [FREE Full text] [doi: 10.1016/j.jad.2020.05.011] [Medline: 32560936]
9. Moberg C, Niles A, Beermann D. Guided self-help works: randomized waitlist controlled trial of Pacifica, a mobile app integrating cognitive behavioral therapy and mindfulness for stress, anxiety, and depression. J Med Internet Res 2019 Jun 08;21(6):e12556 [FREE Full text] [doi: 10.2196/12556] [Medline: 31199319]
10. Serrano-Ripoll MJ, Zamanillo-Campos R, Fiol-DeRoque MA, Castro A, Ricci-Cabello I. Impact of smartphone app-based psychological interventions for reducing depressive symptoms in people with depression: systematic literature review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2022 Jan 27;10(1):e29621 [FREE Full text] [doi: 10.2196/29621] [Medline: 35084346]
11. Venkatesan A, Rahimi L, Kaur M, Mosunic C. Digital cognitive behavior therapy intervention for depression and anxiety: retrospective study. JMIR Ment Health 2020 Aug 26;7(8):e21304 [FREE Full text] [doi: 10.2196/21304] [Medline: 32845246]
12. Anton MT, Greenberger HM, Andreopoulos E, Pande RL. Evaluation of a commercial mobile health app for depression and anxiety (AbleTo Digital+): retrospective cohort study. JMIR Form Res 2021 Sep 21;5(9):e27570 [FREE Full text] [doi: 10.2196/27570] [Medline: 34546170]
13. Raevuori A, Vahlberg T, Korhonen T, Hilgert O, Aittakumpu-Hyden R, Forman-Hoffman V. A therapist-guided smartphone app for major depression in young adults: a randomized clinical trial. J Affect Disord 2021 May 01;286:228-238 [FREE Full text] [doi: 10.1016/j.jad.2021.02.007] [Medline: 33743385]
14. Sawyer A, Kaim A, Le HN, McDonald D, Mittinty M, Lynch J, et al. The effectiveness of an app-based nurse-moderated program for new mothers with depression and parenting problems (eMums plus): pragmatic randomized controlled trial. J Med Internet Res 2019 Jun 04;21(6):e13689 [FREE Full text] [doi: 10.2196/13689] [Medline: 31165715]
15. Graham AK, Greene CJ, Kwasny MJ, Kaiser SM, Lieponis P, Powell T, et al. Coached mobile app platform for the treatment of depression and anxiety among primary care patients: a randomized clinical trial. JAMA Psychiatry 2020 Sep 01;77(9):906-914 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.1011] [Medline: 32432695]
16. Birney AJ, Gunn R, Russell JK, Ary DV. MoodHacker mobile web app with email for adults to self-manage mild-to-moderate depression: randomized controlled trial. JMIR Mhealth Uhealth 2016 Jan 26;4(1):e8 [FREE Full text] [doi: 10.2196/mhealth.4231] [Medline: 26813737]
17. Lahtinen O, Aaltonen J, Kaakinen J, Franklin L, Hyönä J. The effects of app-based mindfulness practice on the well-being of university students and staff. Curr Psychol 2023 May 01;42(6):4412-4421 [FREE Full text] [doi: 10.1007/s12144-021-01762-z] [Medline: 33967565]
18. McCloud T, Jones R, Lewis G, Bell V, Tsakanikos E. Effectiveness of a mobile app intervention for anxiety and depression symptoms in university students: randomized controlled trial. JMIR Mhealth Uhealth 2020 Jul 31;8(7):e15418 [FREE Full text] [doi: 10.2196/15418] [Medline: 32735221]
19. Mantani A, Kato T, Furukawa TA, Horikoshi M, Imai H, Hiroe T, et al. Smartphone cognitive behavioral therapy as an adjunct to pharmacotherapy for refractory depression: randomized controlled trial. J Med Internet Res 2017 Nov 03;19(11):e373 [FREE Full text] [doi: 10.2196/jmir.8602] [Medline: 29101095]
20. Everitt N, Broadbent J, Richardson B, Smyth JM, Heron K, Teague S, et al. Exploring the features of an app-based just-in-time intervention for depression. J Affect Disord 2021 Aug 01;291:279-287. [doi: 10.1016/j.jad.2021.05.021] [Medline: 34058610]
21. Huberty J, Puzia ME, Green J, Vlisides-Henry RD, Larkey L, Irwin MR, et al. A mindfulness meditation mobile app improves depression and anxiety in adults with sleep disturbance: analysis from a randomized controlled trial. Gen Hosp Psychiatry 2021 Nov;73:30-37 [FREE Full text] [doi: 10.1016/j.genhosppsych.2021.09.004] [Medline: 34537477]
22. Chan KL, Leung WC, Tiwari A, Or KL, Ip P. Using smartphone-based psychoeducation to reduce postnatal depression among first-time mothers: randomized controlled trial. JMIR Mhealth Uhealth 2019 May 14;7(5):e12794 [FREE Full text] [doi: 10.2196/12794] [Medline: 31094354]
23. Greer JA, Jacobs JM, Pensak N, Nisotel LE, Fishbein JN, MacDonald JJ, et al. Randomized trial of a smartphone mobile app to improve symptoms and adherence to oral therapy for cancer. J Natl Compr Canc Netw 2020 Feb;18(2):133-141. [doi: 10.6004/jnccn.2019.7354] [Medline: 32023526]
24. Larsen ME, Huckvale K, Nicholas J, Torous J, Birrell L, Li E, et al. Using science to sell apps: evaluation of mental health app store quality claims. NPJ Digit Med 2019 Mar 22;2(1):18 [FREE Full text] [doi: 10.1038/s41746-019-0093-1] [Medline: 31304366]
25. Johansen S. Digital Mental Health. New York, NY: American Psychiatric Association; 2023.
26. Burns DD, Nolen-Hoeksema S. Therapeutic empathy and recovery from depression in cognitive-behavioral therapy: a structural equation model. J Consult Clin Psychol 1992;60(3):441-449. [doi: 10.1037/0022-006x.60.3.441]
27. Burns DD, Nolen-Hoeksema S. Coping styles, homework compliance, and the effectiveness of cognitive-behavioral therapy. J Consult Clin Psychol 1991 Apr;59(2):305-311. [doi: 10.1037/0022-006x.59.2.305]
28. Burns D, Westra H, Trockel M, Fisher A. Motivation and changes in depression. Cogn Ther Res 2012 Apr 22;37(2):368-379. [doi: 10.1007/s10608-012-9458-3]
29. Burns DD. Feeling Great: The Revolutionary New Treatment for Depression and Anxiety. Madison, WI: PESI Publishing & Media; 2020.

30. Percentage rollouts. LaunchDarkly Docs. URL: https://docs.launchdarkly.com/home/targeting-flags/rollouts [accessed 2023-12-08]
31. Martin A, Rief W, Klaiberg A, Braehler E. Validity of the brief Patient Health Questionnaire mood scale (PHQ-9) in the general population. Gen Hosp Psychiatry 2006;28(1):71-77. [doi: 10.1016/j.genhosppsych.2005.07.003] [Medline: 16377369]
32. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
33. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
34. Johnson SU, Ulvenes PG, Øktedalen T, Hoffart A. Psychometric properties of the General Anxiety Disorder 7-item (GAD-7) scale in a heterogeneous psychiatric sample. Front Psychol 2019 Aug 6;10:1713 [FREE Full text] [doi: 10.3389/fpsyg.2019.01713] [Medline: 31447721]
35. Jordan P, Shedden-Mora MC, Löwe B. Psychometric analysis of the Generalized Anxiety Disorder scale (GAD-7) in primary care using modern item response theory. PLoS One 2017 Aug 3;12(8):e0182162 [FREE Full text] [doi: 10.1371/journal.pone.0182162] [Medline: 28771530]
36. Rutter LA, Brown TA. Psychometric properties of the Generalized Anxiety Disorder Scale-7 (GAD-7) in outpatients with anxiety and mood disorders. J Psychopathol Behav Assess 2017 Mar;39(1):140-146 [FREE Full text] [doi: 10.1007/s10862-016-9571-9] [Medline: 28260835]
37. Stinson L, Liu Y, Dallery J. Ecological momentary assessment: a systematic review of validity research. Perspect Behav Sci 2022 Jun 06;45(2):469-493 [FREE Full text] [doi: 10.1007/s40614-022-00339-w] [Medline: 35719870]
38. Wolfe KL, Nakonezny PA, Owen VJ, Rial KV, Moorehead AP, Kennard BD, et al. Hopelessness as a predictor of suicide ideation in depressed male and female adolescent youth. Suicide Life Threat Behav 2019 Feb 21;49(1):253-263 [FREE Full text] [doi: 10.1111/sltb.12428] [Medline: 29267993]
39. Labelle R, Breton JJ, Pouliot L, Dufresne MJ, Berthiaume C. Cognitive correlates of serious suicidal ideation in a community sample of adolescents. J Affect Disord 2013 Mar 05;145(3):370-377. [doi: 10.1016/j.jad.2012.08.027] [Medline: 23102862]
40. Cunningham JA, Kypri K, McCambridge J. Exploratory randomized controlled trial evaluating the impact of a waiting list control design. BMC Med Res Methodol 2013 Dec 06;13(1):150 [FREE Full text] [doi: 10.1186/1471-2288-13-150] [Medline: 24314204]
41. Lagan S, Emerson MR, King D, Matwin S, Chan SR, Proctor S, et al. Mental health app evaluation: updating the American Psychiatric Association's framework through a stakeholder-engaged workshop. Psychiatr Serv 2021 Sep 01;72(9):1095-1098. [doi: 10.1176/appi.ps.202000663] [Medline: 33882716]

Abbreviations

7DES: 7-Dimension Emotion Sliders
7DES-d: 7-Dimension Emotion Sliders–depression
CBT: cognitive behavioral therapy
CONSORT: Consolidated Standards of Reporting Trials
EBT: end of basic training
GAD-7: Generalized Anxiety Disorder-7
IBT: intensive basic training
IS: immediate start
OTM: ongoing training module
PHQ-9: Patient Health Questionnaire-9
SBT: start of basic training
WC: waitlist control
TEAM-CBT: Testing Empathy Assessment Methods–Cognitive Behavioral Therapy

Edited by J Torous; submitted 01.09.23; peer-reviewed by M Zhou, D Liu; comments to author 23.11.23; revised version received 08.12.23; accepted 12.04.24; published 10.05.24.

Please cite as:
Bisconti N, Odier M, Becker M, Bullock K
Feasibility and Acceptability of a Mobile App–Based TEAM-CBT (Testing Empathy Assessment Methods–Cognitive Behavioral Therapy) Intervention (Feeling Good) for Depression: Secondary Data Analysis
JMIR Ment Health 2024;11:e52369
URL: https://mental.jmir.org/2024/1/e52369
doi:10.2196/52369
PMID:38728080

Original Paper

Cost-Effectiveness of Digital Mental Health Versus Usual Care During Humanitarian Crises in Lebanon: Pragmatic Randomized Trial

Racha Abi Hana[1,2], MSc; Jinane Abi Ramia[1,2], MPH; Sebastian Burchert[3], MSc; Kenneth Carswell[4], DclinPsy; Pim Cuijpers[1,5], PhD; Eva Heim[6,7], PhD; Christine Knaevelsrud[3], PhD; Philip Noun[2], MSc; Marit Sijbrandij[1], PhD; Mark van Ommeren[4], PhD; Edith van't Hof[4], PhD; Ben Wijnen[8], PhD; Edwina Zoghbi[9], MPH; Rabih El Chammay[2,10*], MD; Filip Smit[1,11,12*], PhD

[1]Clinical, Neuro- and Developmental Psychology Department, Vrije Universiteit Amsterdam, Amsterdam, Netherlands
[2]National Mental Health Programme, Ministry of Public Health, Beirut, Lebanon
[3]Division of Clinical Psychological Intervention, Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany
[4]Department of Mental Health and Substance Use, World Health Organization, Geneva, Switzerland
[5]International Institute for Psychotherapy, Babeş-Bolyai University, Cluj-Napoca, Romania
[6]Institute of Psychology, University of Lausanne, Lausanne, Switzerland
[7]Department of Psychology, University of Zurich, Zurich, Switzerland
[8]Centre of Economic Evaluation and Machine Learning, Trimbos Institute (Netherlands Institute of Mental Health and Addiction), Utrecht, Netherlands
[9]Country Office for Lebanon, World Health Organization, Beirut, Lebanon
[10]Faculty of Medicine, Psychiatry Department, Saint Joseph University, Beirut, Lebanon
[11]Department of Epidemiology and Biostatistics, Amsterdam University Medical Centers, Amsterdam, Netherlands
[12]Department of Mental Health and Prevention, Trimbos Institute (Netherlands Institute of Mental Health and Addiction), Utrecht, Netherlands
[*]these authors contributed equally

Corresponding Author:
Racha Abi Hana, MSc
Clinical, Neuro- and Developmental Psychology Department
Vrije Universiteit Amsterdam
De Boelelaan 1105
Amsterdam, 1105 1081 HV
Netherlands
Phone: 31 3978968
Fax: 31 (0)20 59 8
Email: rasha_abihana@hotmail.com

Abstract

Background: There is evidence from meta-analyses and systematic reviews that digital mental health interventions for depression, anxiety, and stress-related disorders tend to be cost-effective. However, no such evidence exists for guided digital mental health care in low and middle-income countries (LMICs) facing humanitarian crises, where the needs are highest. Step-by-Step (SbS), a digital mental health intervention for depression, anxiety, and stress-related disorders, proved to be effective for Lebanese citizens and war-affected Syrians residing in Lebanon. Assessing the cost-effectiveness of SbS is crucial because Lebanon's overstretched health care system must prioritize cost-effective treatment options in the face of continuing humanitarian and economic crises.

Objective: This study aims to assess the cost-effectiveness of SbS in a randomized comparison with enhanced usual care (EUC).

Methods: The cost-effectiveness analysis was conducted alongside a pragmatic randomized controlled trial in 2 parallel groups comparing SbS (n=614) with EUC (n=635). The primary outcome was cost (in US $ for the reference year 2019) per treatment response of depressive symptoms, defined as >50% reduction of depressive symptoms measured using the Patient Health Questionnaire (PHQ). The secondary outcome was cost per remission of depressive symptoms, defined as a PHQ score <5 at last follow-up (5 months post baseline). The evaluation was conducted first from the health care perspective then from the societal perspective.

Results: Taking the health care perspective, SbS had an 80% probability to be regarded as cost-effective compared with EUC when there is a willingness to pay US $220 per additional treatment response or US $840 per additional remission. Taking the wider societal perspective, SbS had a >75% probability to be cost-saving while gaining response or remission.

Conclusions: To our knowledge, this study is the first cost-effectiveness analysis based on a large randomized controlled trial (n=1249) of a guided digital mental health intervention in an LMIC. From the principal findings, 2 implications flowed, from the (1) health care perspective and (2) wider societal perspective. First, our findings suggest that SbS is associated with greater health benefits, albeit for higher costs than EUC. It is up to decision makers in health care to decide if they find the balance between additional health gains and additional health care costs acceptable. Second, as seen from the wider societal perspective, there is a substantial likelihood that SbS is not costing more than EUC but is associated with cost-savings as SBS participants become more productive, thus offsetting their health care costs. This finding may suggest to policy makers that it is in the interest of both population health and the wider Lebanese economy to implement SbS on a wide scale. In brief, SbS may offer a scalable, potentially cost-saving response to humanitarian emergencies in an LMIC.

Trial Registration: ClinicalTrials.gov NCT03720769; https://clinicaltrials.gov/ct2/show/NCT03720769

International Registered Report Identifier (IRRID): RR2-10.2196/21585

(JMIR Ment Health 2024;11:e55544) doi:10.2196/55544

KEYWORDS

depression; internet-based intervention; economic evaluation; Lebanese; Syrian; digital mental health; digital health; mental health; usual care; Lebanon; anxiety; stress-related disorders; treatment; symptoms; large randomized controlled trial; effectiveness

Introduction

In the last decade, Lebanon registered an influx of around 1 million Syrian displaced people [1], in addition to approximately 0.5 million unregistered refugees. They are at considerable risk of developing mental disorders such as depression, anxiety, and posttraumatic stress disorder and often in need of mental health services [2]. Meanwhile, Lebanon's own population suffered from a series of overlapping humanitarian and economic crises (ongoing political turmoil, hyperinflation, the COVID-19 pandemic, the Beirut port blast), which undermined population mental health [3] and exacerbated the pressure on Lebanon's already overstretched health care system. As a result of these emergencies, Lebanese health care workers have been put under immense pressure, and many sought to leave the country to work overseas [4].

Considering the crises-related needs, transformations were required in line with Lebanon's national mental health strategy, in which one of its objectives was to scale-up digital self-help programs for priority conditions such as depression, anxiety, and posttraumatic stress in a cost-effective way [5]. To this end, the World Health Organization (WHO) helped to develop, test, and implement a guided e-mental health intervention for depression among Lebanese citizens, displaced Syrians, and other people residing in Lebanon [6-10]. The intervention is called Step-by-Step (SbS) and can be downloaded on digital devices. SbS consists of 5 sessions based on evidence-based psychological treatments, primarily behavioral activation. It is delivered with the aid of nonspecialist helpers trained and supervised to provide participants with guidance via messaging or by phone [6].

Elsewhere, we demonstrated that the SbS intervention is effective in reducing depression, anxiety, and posttraumatic stress and in improving personal and social functioning and subjective well-being in Lebanese citizens, Syrians, and other populations residing in Lebanon [11,12]. The objective of this paper was to report on the cost-effectiveness of offering the effective SbS intervention in Lebanon. This study is among the first to evaluate the cost-effectiveness of scaling up a response to crises using a digital intervention in low- and middle-income countries.

Methods

Study Design

The study was designed as a cost-effectiveness analysis alongside a pragmatic randomized controlled trial with 2 parallel groups comparing SbS with enhanced usual care (EUC) with assessments at baseline (t0); 8 weeks post baseline (t1), which was after completion of the intervention; and 20 weeks post baseline (t2), hence 3 months after conclusion of the intervention. A study protocol [10], pilot study [8], and feasibility trial [9] have been published, and effectiveness studies showed positive clinical effects on depression, disability, and symptoms of anxiety and posttraumatic stress [11,12].

Ethics Approval

The WHO (ERC.0002797) and Saint Joseph's University in Beirut (CEHDF862) Ethical Review Committees provided medical ethics approval.

Participants

Participants were Lebanese citizens, displaced Syrians in Lebanon, and other people residing in Lebanon. Participants were recruited via social media, online advertisements, and outreach activities. Participants could download the SbS app for iOS and Android or visit its web version, which provided information about the intervention and the study and included a screener for eligibility. To be included in the study, participants had to (1) be aged 18 years or older, (2) reside in Lebanon, (3) be able to speak and understand Arabic or English, (4) have access to a device connected to the internet, (5) score higher than 10 on the Patient Health Questionnaire (PHQ-9) [13] for depressive symptom severity, and (6) score higher than

16 on the World Health Organization Disability Assessment Scale Schedule-12 (WHODAS) for impaired functioning [14]. The exclusion criterion was imminent risk of suicide, in which case the participant received psychoeducation and was directed to the national suicide prevention lifeline. Consenting participants who fulfilled the inclusion criteria were invited to complete the baseline questionnaire. As an incentive, those who completed the questionnaires received a US $20 phone credit.

Randomization and Masking

Eligible and consenting participants were randomized to either the SbS intervention or EUC using automated permuted block randomization with a 1:1 allocation ratio within blocks of random length between 2 and 8. Randomization was stratified for nationality: Syrians in Lebanon versus Lebanese citizens and other people residing in Lebanon. The randomization algorithm was built into the SbS app and was not accessible by the research team.

Interventions

SbS was a guided digital health intervention to alleviate symptoms of depression, anxiety, and posttraumatic stress [6]. The intervention was based on evidence-based therapeutic techniques. The main therapeutic technique was behavioral activation with additional psychoeducation, stress management, positive self-talk, gratitude practice, reinforcing social support, and relapse prevention [6]. Weekly support to users was provided by phone or messaging by trained nonspecialists called “e-helpers” [9]. The e-helpers were trained over 5 days and continued to receive weekly group supervision by a clinical supervisor as well as individual supervision when needed [11,12]. e-Helpers could only begin their job after passing a competency examination subsequent to the completion of their training. Using a treatment fidelity checklist [11], fidelity checks found minor deviations from the treatment plan, such as when e-helpers skipped practice activities or did not go through the story with users fully [11].

EUC consisted of a psychoeducational message on the SbS app (similar to the first SbS session) and a list of primary health care centers available in different areas in Lebanon. Staff in these health centers were trained in screening, detecting, and managing mental health conditions [15].

Outcome Measures

The central outcome was the PHQ-9 [13]. The PHQ-9 is a 9-item instrument measuring severity of depression, with a cutoff score >10 indicating moderate to severe depression, which has also been validated in Lebanon [16]. For clinical and economic interpretation, the PHQ-9 was converted to treatment response and remission. Response was defined as an improvement of at least 50% between t0 and t2 in depressive symptom severity as measured by the PHQ-9. Remission was defined as a participant’s PHQ-9 score below 5 at t2 [17].

Resource Use and Costs

Questionnaires

Costs stemming from health care uptake and productivity losses were collected using a Lebanese Resource Use questionnaire that was based on both the Trimbos and iMTA (Institute of Medical Technology Assessment) Cost questionnaire for Psychiatric illness (TiC-P) [18], which has good reliability and validity [19], and the cross culturally validated Client Service Receipt Inventory (CSRI) [20] adapted and piloted for use in Lebanon [9]. The Lebanese Resource Use questionnaire was programmed into the SbS app as a self-report questionnaire.

All costs are expressed in US $ for the year 2019 when the study was carried out and when, according to the World Bank, the average exchange rate for US $1 was LBP 1507.50.

Cost of SbS

In the year 2019, the personnel cost of offering SbS amounted to US $59,520 (consisting of the gross annual salaries of a 0.2 full-time equivalent [FTE] clinical supervisor, 1 FTE coordinator, 1 FTE senior e-helper, and 2 FTE e-helper). The annual nonpersonnel cost of offering SbS was US $62,800 (US $36,000 for hosting, maintaining, and periodically upgrading the digital intervention; US $14,800 for renting the office, equipment, and overhead; and US $2000 for advertising). The total personnel and nonpersonnel costs of operating SbS was therefore US $59,520 + $62,800 = $122,320. SbS can serve 4700 users in a year. Therefore, the per-user costs of SbS was US $122,320/4700 = US $26.

Cost of EUC

People randomized to EUC received an online psychoeducational message derived from the first session of the SbS intervention. They also received a list of primary health care facilities with nonspecialized staff trained in the Mental Health Gap Action Programme [15]. The per-user cost of this online message was next to nothing (US $0.01) and was ignored in the subsequent analyses.

Cost of Health Care Utilization and Productivity Costs

Table 1 reports the cost prices per unit health care, such as a visit to a general practitioner, session with a psychologist, or day in a mental ward.

Costs can be regarded from the health care perspective and from the societal perspective. Taking the health care perspective, only the direct medical costs were considered, stemming from the contacts of participants with health services. Costs were evaluated first from the health care perspective and second from the broader societal perspective, thus adding the costs stemming from productivity losses to health care costs. Productivity losses occur when a person stays absent from work (absenteeism) as well as when a person does not feel well, tries to work anyway but is less productive (presenteeism), resulting in work cutback.

Table 1. Unit cost prices in US $ (2019 price level).

Item	Unit	Unit cost price (US $), mean	Range	Source
Intervention				
Step by Step	Usage	26.00	—[a]	Estimate 1[b]
EUC[c] psychoeducation	Online message	0.01	—	Estimate 1
Primary care				
General practitioner (GP)	Contact	23.66	7.32[d]-40.00[e]	GPs
Nurse	Contact	1.83	—	Ministry PH[f]
Social worker	Contact	2.59	—	Ministry PH[f]
Outpatient care				
Psychiatrist	Consult	56.06	24.62[d]-87.50[e]	Ministry PH[f]
Neurologist	Consult	37.33	9.66[d]-65.00[e]	Ministry PH[f]
Psychologist	Session	25.37	13.23[d]-37.50[e]	Ministry PH[f]
Inpatient care				
Psychiatric ward	Day	325.00	150.00[d]-500.00[e]	Hospital
Mental hospital	Day	150.00	150.00[e]	Hospital
Emergency				
Ambulance trip	Transport	99.00	—	Red Cross
Emergency room	Visit	48.75	25.00[d]-72.50[e]	Hospital
Medication				
Antidepressants	DDD[g]	0.63	0.41-0.90	Estimate 2[h]
Anxiolytics	DDD	0.36	0.23-0.55	Estimate 3[i]
Hypnotics	DDD	0.39	0.23-0.55	Estimate 4[j]
Productivity				
Paid work	Workday	32.09	16.66[k]-47.51[l]	Estimate 5[m]
Unpaid work	Workday	7.82	—	Estimate 6[n]

[a]Not applicable.
[b]See the “Resource Use and Costs” section.
[c]EUC: enhanced usual care.
[d]Public.
[e]Private.
[f]Tariffs provided by the Ministry of Public Health (PH).
[g]DDD: daily defined dose.
[h]Average cost price per DDD of frequently prescribed antidepressants (escitalopram, sertraline, amitriptyline, clomipramine, venlafaxine).
[i]Average cost price per DDD of frequently prescribed anxiolytics (bromazepam, alprazolam, lorazepam, hydroxyzine, alprazolam).
[j]Average cost price per daily defined dose of hypnotics (zolpidem and melatonin).
[k]Syrian.
[l]Lebanese.
[m]Based on the Labour Market Assessment in Beirut and Mount Lebanon by the Agency for Technical Cooperation and Development [21] and indexed for the year 2019.
[n]Opportunity costs valued as the average per diem salary of domestic help in Lebanon in 2019.

Base Case Analysis

The health economic evaluation was conducted in agreement with the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) 2022 guideline for trial-based economic evaluations [22]. The costs are reported in US $ for the reference year 2019. The study's time horizon was 20 weeks (ie, 5 months).

With a sample size of 1136, the study would be powered to detect a standardized mean difference of medium size (d≥0.50) between the conditions, as statistically significant at $\alpha \leq .05$ (2-tailed) and a power of $(1 - \beta) \geq 0.90$ while accounting for an expected dropout rate of 70%, which is typical for self-help interventions [23]. The study was powered to evaluate a clinical depression–related outcome but not for a health economic evaluation for which costs are typically associated with large standard errors. In other words, in this study, we could statistically test effect differences but not cost differences. Instead, health-economic inferences were not based on statistical hypothesis testing but on probabilistic medical decision-making techniques.

At each assessment (t0, t1, and t2), cost data were collected retrospectively from the last 4 weeks. Cumulative costs over the full trial duration of 20 weeks were computed using linear interpolation among the t0, t1, and t2 assessments. Incremental costs and incremental effects were computed as the difference of the cumulative costs and effects between the conditions. The incremental cost-effectiveness ratio (ICER) was computed as the cost difference over the effect difference: ICER = $(C_1\text{-}C_0)/(E_1\text{-}E_0)$, where C and E are costs and effects and the subscripts 1 and 0 refer to the SbS and EUC conditions, respectively. The ICER is interpreted as the additional costs for gaining a treatment response and remission.

For intention-to-treat analysis, missing PHQ-9 scores were imputed using regression imputation. The regression imputation model included 2 types of predictors: predictors of outcome (societal costs, PHQ-9, gender, age, education, and WHO-5 Wellbeing as measured at t0) and predictors of missingness (randomization status, partner status, employment status, and WHODAS-12 at t0). The first set of predictors was included for predictive accuracy of the outcome, and the second type was included to better satisfy the missing-at-random assumption [24]. After imputation, the PHQ-9 score was converted into treatment response and remission.

To simultaneously evaluate both incremental costs and effects, seemingly unrelated regression equations (SURE) models were used. In the SURE model, baseline costs were included to adjust for a baseline imbalance because EUC had higher baseline costs than SbS. Since cost data were nonnormally distributed, nonparametric bootstraps (2500 times) were used for the SURE models. The bootstrapped SURE models helped to create 2 figures: (1) the ICER plane on which the simulated ICERs are plotted and (2) the acceptability curve. We return to these figures and their interpretation in the Results section.

Sensitivity Analysis

The base case analysis had treatment response and depressive remission as outcomes. It was first conducted from the health care perspective, then it was repeated from the societal perspective. Missing observations due to dropout were imputed using regression imputation. In a preplanned sensitivity analysis, the base case analysis was repeated using multiple imputation with chained equations (MICE) using predictive mean matching to impute missing observations [25]. This was done to see how robust the results were under varying imputation strategies. All analyses were carried out in Stata 17.0 [26].

Results

Participants

Recruitment of the participants started December 9, 2019, and ended on July 9, 2020. In this 7-month period, 3042 persons were assessed for eligibility, 1676 met the inclusion criteria, and 1249 were randomized: 614 to SbS and 635 to EUC. Figure 1 shows the flow of participants through the trial.

At t1, dropout was 64.2% (394/614) in the SbS group and 51.5% (327/635) in the EUC group. At t2, these rates increased by an additional 24 and 43 participants, respectively, such that total dropout became 68.1% (418/614) in the SbS group and 58.3% (370/635) in the EUC group. Total dropout in the whole sample was 788/1249, or 63.1%. As indicated, a 70% dropout has to be expected in digital self-help interventions [23].

The demographic, clinical, and economic characteristics of the participants are listed in Table 2. It appears that randomization led to an even distribution of these variables over the conditions; however, societal costs appeared somewhat higher in the EUC group than the SbS group (Table 2).

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flowchart. t0: baseline; t1: 8 weeks post baseline, which was after completion of the intervention; t2: 20 weeks post baseline.

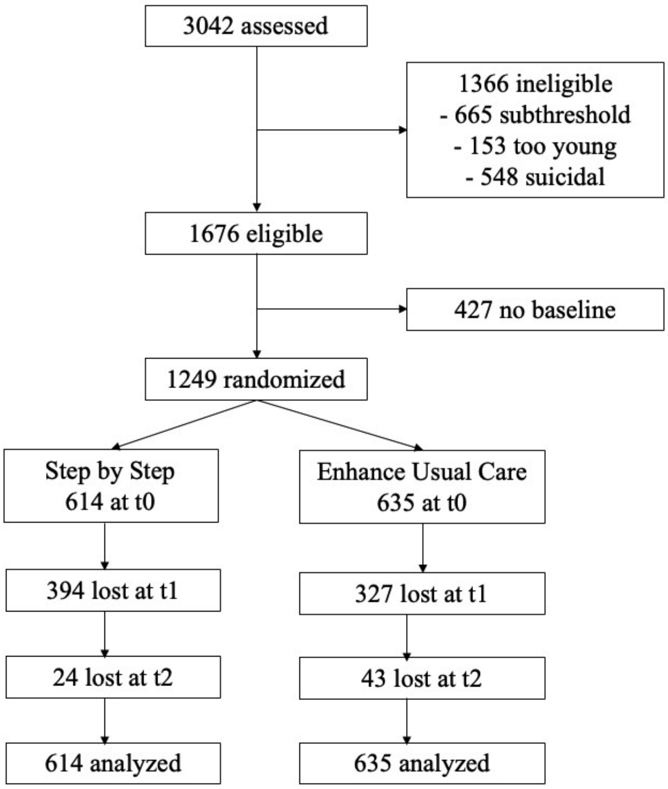


Table 2. Characteristics of the sample at baseline (N=1249).

Characteristic	SbS[a] (n=614)	EUC[b] (n=635)	Total (N=1249)
Age (years), mean	29.0	29.1	29.1
Gender, n (%)			
Female	415 (67.6)	393 (61.9)	808 (64.7)
Male	199 (32.4)	242 (38.1)	441 (35.3)
Marital status, n (%)			
Never married	275 (44.8)	289 (45.6)	564 (45.2)
Married	285 (46.4)	300 (47.2)	585 (46.8)
Other	54 (8.8)	46 (7.2)	100 (8)
Nationality, n (%)			
Lebanese	307 (50)	315 (49.6)	622 (49.8)
Syrian	275 (44.8)	283 (44.6)	558 (44.7)
Other	32 (5.2)	37 (5.8)	69 (5.5)
Education, n (%)			
Primary	138 (22.5)	128 (20.1)	266 (21.3)
Secondary	136 (22.2)	154 (24.3)	290 (23.2)
Vocational	238 (38.8)	261 (41.1)	499 (40)
Academic	102 (16.6)	92 (14.5)	194 (15.5)
Employment status, n (%)			
Employed	143 (23.3)	169 (26.6)	312 (25)
Homemaker	97 (15.8)	98 (15.4)	195 (15.6)
Student	106 (17.3)	107 (16.9)	213 (17.1)
Retired	1 (0.2)	4 (0.6)	5 (0.4)
Unemployed	267 (43.5)	257 (40.5)	524 (42)
Clinical characteristics, mean			
PHQ[c] depression	16.4	16.4	16.4
WHODAS[d] disability	33.0	33.1	33.0
Costs (last 4 weeks; US $)			
Health care costs	42	45	44
Societal costs	107	124	115

[a]SbS: Step-by-Step.

[b]EUC: enhanced usual care.

[c]PHQ: Patient Health Questionnaire.

[d]WHODAS: World Health Organization Disability Assessment Scale.

Cumulative Costs

Table 3 describes the costs in the last 4 weeks at the t0, t1, and t2 assessments and how these costs accumulated over the full 20-week period. Missing cost data at t1 and t2 due to dropouts were imputed; the per-user cost of SbS was not included.

Table 3 shows that the mean health care costs fluctuated per participant slightly over time. The cumulative health care costs were virtually the same in both groups (US $241 [SbS] vs US $243 [EUC]). However, the cumulative productivity costs ended up being lower with SbS (US $151) than with EUC (US $202), indicating that SbS helped to reduce productivity losses. This is also mirrored in the mean societal costs that were US $443 with EUC and US $393 with SbS. A detailed breakdown of health care and productivity costs is provided in Table S1 in Multimedia Appendix 1.

Table 3. Per-participant costs in US $ over time by condition, not including US $26 Step-by-Step (SbS) costs (N=1249).

Costs by group	t0[a] (US $)	t1[b] (US $)	t2[c] (US $)	Cum (t0-t2) cost (US $; 95% CI)
Health care				
EUC[d]	45	48	51	241 (203-279)
SbS	42	55	43	243 (209-276)
Productivity				
EUC	79	17	55	202 (176-229)
SbS	65	15	32	151 (129-172)
Societal				
EUC	124	65	105	443 (396-490)
SbS	107	70	75	393 (354-433)

[a]Baseline.

[b]8 weeks post baseline, which was after completion of the intervention.

[c]20 weeks post baseline.

[d]EUC: enhanced usual care.

Base Case Analysis

Incremental Costs

Taking the health care perspective, including the SbS costs of US $26 per recipient, the bootstrapped incremental health care costs averaged US $28 (95% CI –$22 to $78), suggesting that SbS costs somewhat more than EUC, but this was not statistically significant (bootstrap SE=25.44, z=1.10; P=.27). For context, US $28 would buy a single session with a psychologist in Lebanon in 2019.

Taking the societal perspective, the incremental costs averaged US –$24 (95% CI –$85 to $37), suggesting a small cost reduction favoring SbS over EUC, which was not statistically significant (bootstrap SE=31.15, z=–0.76; P=.45).

It is worth noting that there is a discrepancy between the incremental cost as seen from the health care perspective (US $28) and the incremental costs as seen from the societal perspective (US –$24). The cost reduction that comes into view when taking the societal perspective must be related to the greater productivity (less absenteeism and less presenteeism) among the recipients of SbS.

Incremental Effects

The incremental response rate was 0.23 (95% CI 0.18 to 0.27) favoring SbS in a statistically significant way (bootstrap SE=0.022, z=10.24; P<.001).

The incremental remission rate was 0.06 (95% CI 0.031 to 0.085) favoring SbS over EUC and was statistically significant (bootstrap SE=0.014; z=4.20; P<.001).

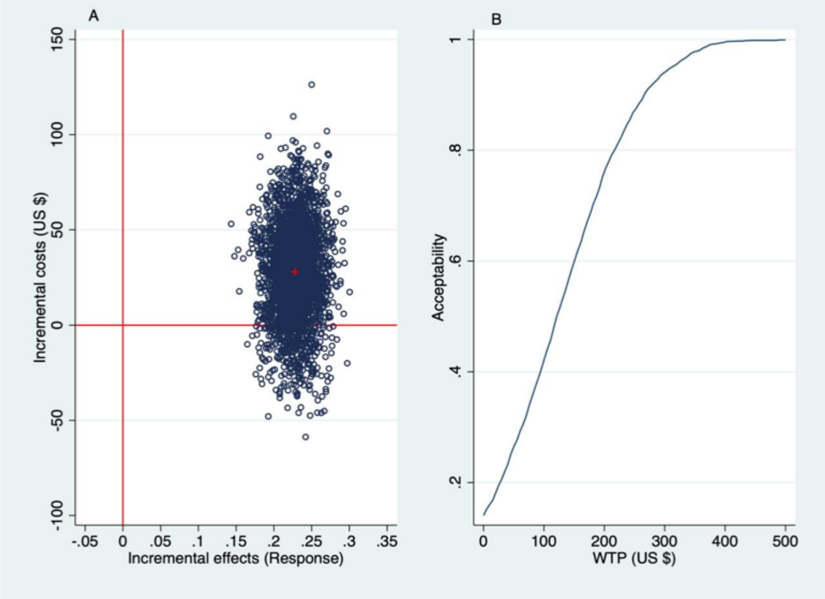
ICER Response: Health Care Perspective

Dividing the incremental health care costs of US $28 by the incremental response rate of 0.23 gives an ICER of US $28/0.23=$122. Thus, SbS costs US $122 more than EUC per treatment responder. It should be noted that the ICER of US $122 is an average and is surrounded by uncertainty, as depicted in Figure 2.

Figure 2A shows the uncertainty around the mean ICER as a scatter of simulated ICERs over the ICER plane. The ICER plane is divided into 4 quadrants, denoted North East (NE), North West (NW), South West (SW), and South East (SE). Of the bootstrapped ICERs, 86% appear in the NE quadrant of the ICER plane, indicating an 86% probability that better effects are obtained by SbS albeit for higher costs than EUC. The remainder of the simulated ICERs appear in the SE quadrant, indicating that health gains are achieved by SbS while cost reductions occurred compared with EUC. Thus, SbS generates a greater health gain albeit for more costs than EUC.

This begs the question about how much one is willing to pay to gain an additional treatment responder. The acceptability curve (Figure 2B) is now key to decision-making. The acceptability curve plots the probability that the intervention is deemed cost-effective (ie, acceptable) on the Y axis (range 0.00-1.00) against various willingness-to-pay (WTP) levels to gain a response on the X axis (range US $0-$500). If there is no WTP to gain an additional response (WTP=US $0), then there is a 14% probability that SbS is acceptable from a cost-effectiveness point of view (because 14% of the simulated ICERs appeared in the SE quadrant). Similarly, the SbS intervention has a 50% probability of being regarded as acceptable if society would be willing to pay US $122 per PHQ-9 response (because ICER=US $122 to gain an additional responder). Assuming a decision-maker would like to have a greater than 50% certainty and instead seeks an 80% probability for acceptability, it follows that the WTP for gaining a treatment responder must be about US $220 (ie, the WTP at which the acceptability curve meets the 80% probability level of acceptability).

Figure 2. Health care perspective: incremental costs per additional Patient Health Questionnaire (9-item version) responder shown in a (A) cost-effectiveness plane and (B) acceptability curve (2500 bootstraps). WTP: willingness to pay.



ICER Response: Societal Perspective

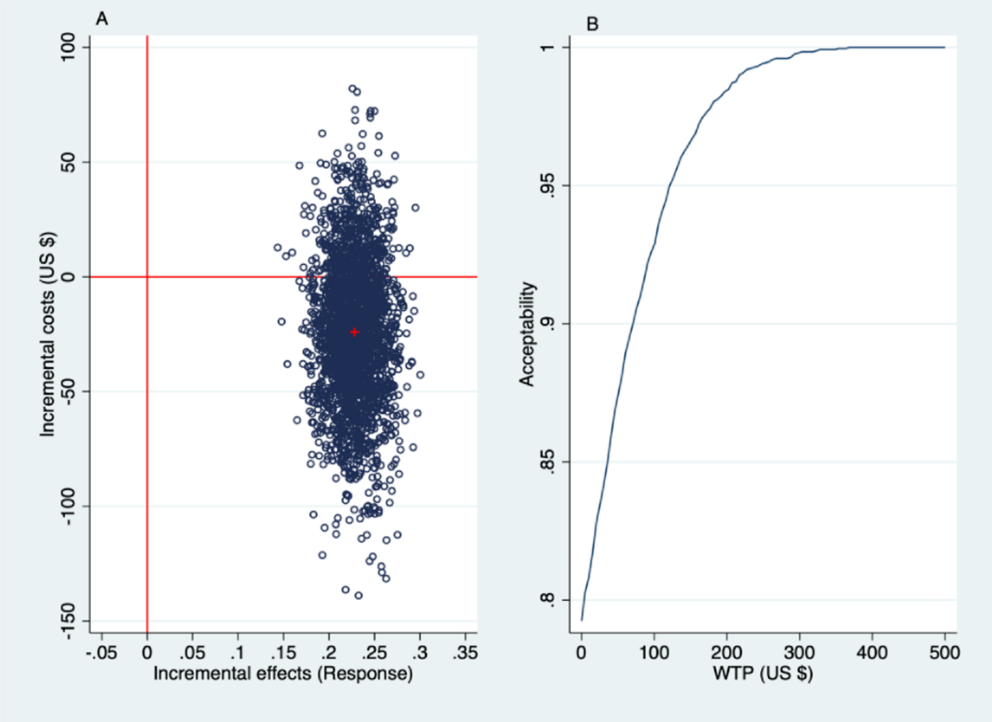
The base case analysis with treatment response as the primary outcome was repeated taking the societal perspective. The incremental effect is as before: a 0.23 greater probability of treatment response with SbS than with EUC but with an incremental societal cost of –$24 (negative cost, hence cost reductions). The fact that societal costs are lower than the health care costs can only be explained by greater productivity of the participants after having received SbS. Absenteeism (and associated costs) may have been reduced because more SbS recipients returned to work than those who received EUC. In addition, a greater proportion of people who received SbS became more productive (ie, less presenteeism).

The incremental cost per additional treatment responder was also negative (US –$24/0.23 = –$105); therefore, the ICER is said to be “dominant,” because better effects are obtained for less cost with SbS than with EUC. Figure 3 presents the outcomes.

Figure 3A depicts the scatter of simulated ICERs. Of these, 21% appear in the NE quadrant, indicating 21% likelihood that the SbS intervention had better effects albeit for higher costs than EUC. However, a more substantial 79% of the simulated ICERs are in the SE quadrant, which is indicative of a 79% probability that gaining a treatment response with SbS is associated with cost reductions compared with EUC; hence, SbS is deemed “dominant” (ie, represents the more favorable treatment option compared with EUC from a cost-effectiveness point of view). Now, there is no need for a decision maker to review the acceptability curve because, in the context of cost savings, any WTP threshold has become irrelevant. Nonetheless, a look at the acceptability curve (in Figure 3B) shows that, at WTP=US $0, the likelihood of acceptability is 0.78 (78%), increasing to 0.95 (95%) at a WTP of US $110.

In sum, taking the health care perspective, SbS has an 80% likelihood to be regarded as cost-effective when there is a WTP of US $220 for an additional PHQ-9 response. However, taking the societal perspective, SbS is associated with a cost reduction and hence “dominant.” To reiterate, the cost reduction as seen from the societal perspective indicates that people become more productive (lesser absenteeism and presenteeism) after receiving SbS.

Figure 3. Societal perspective: incremental costs per additional Patient Health Questionnaire (9-item version) responder shown in a (A) cost-effectiveness plane and (B) acceptability curve (2500 bootstraps). WTP: willingness to pay.

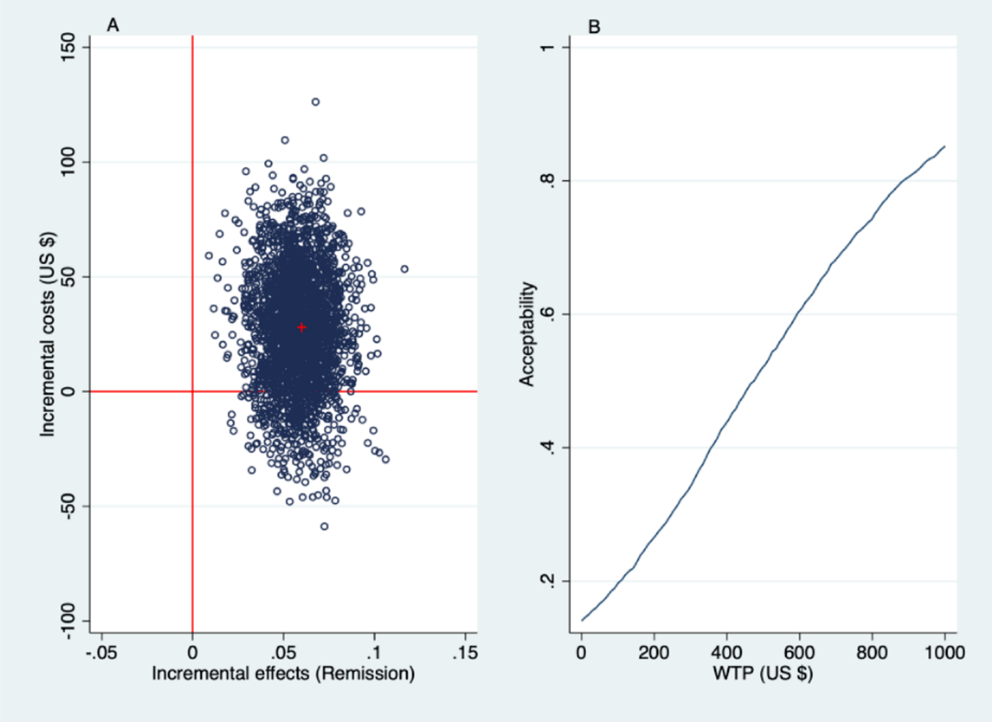


ICER Remission: Health Care Perspective

The base case analysis was repeated for the secondary outcome, remission. The incremental health care costs remained the same as before, at US $28. The remission rate achieved was higher by 0.058 (95% CI 0.031-0.085) with SbS than with EUC, which was statistically significant (bootstrap SE=0.014, z=4.20; P<.001). Division of the incremental health care cost by the incremental remission rate gives US $28/0.058=$474. Thus, to achieve a remission with SbS costs more than with EUC, by $474. Figure 4 presents the results.

Since achieving a remission with SbS costs more than with EUC, the acceptability curve needs to be reviewed for decision-making. At WTP=US $0, there is a 14% probability that SbS is deemed to be acceptable from a cost-effectiveness point of view (because 14% of the simulated ICERs appeared in the SE quadrant). At WTP=US $474, the acceptability reached the 50% probability level (because the ICER=US $474 for health care costs per remission). Should a greater than 50% certainty be required for decision-making in health, say, acceptability at the probability level of 80%, then it follows from Figure 4 that the required WTP threshold is about US $840 per remission.

Figure 4. Health care perspective: incremental costs per additional Patient Health Questionnaire (9-item version) responder shown in a (A) cost-effectiveness plane and (B) acceptability curve (2500 bootstraps). WTP: willingness to pay.



ICER Remission: Societal Perspective

Looking at remission and taking the societal perspective, the incremental costs are negative (US –$24), indicating a cost reduction (due to fewer productivity losses after receiving SbS). This implies the SbS intervention dominates EUC as the more cost-effective treatment option as seen from the wider societal perspective.

When taking the societal perspective, SbS was dominant relative to EUC because costs will be saved in both the base case analysis and sensitivity analyses.

Sensitivity Analysis

Table 4 summarizes the outcomes from the base case analysis (incremental costs per response and per remission based on regression imputation) and compares these with those from the sensitivity analysis (MICE using predictive mean matching).

Taking the health care perspective, Table 4 shows that the ICER is US $121 per responder in the base case analysis (US $101 in the sensitivity analysis) and US $474 per remission in the base case analysis (US $141 in the sensitivity analysis).

Overall, the sensitivity analyses produced results that were fairly similar to the base case analysis but produced a higher remission rate (0.163) relative to the base case analysis (0.058). The estimate of 0.058 can only be interpreted as a small but positive incremental effect of SbS on remission. The estimate of 0.163 is still indicative of a small but positive incremental effect of SbS on remission, and our previous conclusion would not alter in any material way. All in all, the sensitivity analysis produced similar or roughly similar outcomes as the base case analysis. This is important because the study was subject to an expected, but large, dropout; therefore, it is important to assess if the results did not crucially depend on one or another imputation technique but that different imputation techniques indeed produced similar results.

Table 4. Base case and sensitivity analyses.

Analysis, perspective, and outcome	Incremental costs (US $)	Incremental effects	ICER[a] cost (US $) or effect	ICER distribution, %			
				NE	NW	SW	SE
Base case							
Health care							
Response	28	0.228	121	86	0	0	14
Remission	28	0.058	474	86	0	0	14
Societal							
Response	–24	0.228	Dominant	21	0	0	79
Remission	–24	0.058	Dominant	21	0	0	79
Sensitivity							
Health care							
Response	23	0.228	101	76	0	0	24
Remission	23	0.163	141	76	0	0	24
Societal							
Response	–29	0.228	Dominant	24	0	0	76
Remission	–28	0.163	Dominant	24	0	0	76

[a]ICER: incremental cost-effectiveness ratio.

Discussion

Principal Findings

The health-economic evaluation of SbS versus EUC was conducted from 2 perspectives: the health care perspective and the societal perspective. The evaluation showed that the distinction between both perspectives is important. After all, SbS turned out to be associated with additional costs than EUC when seen from the health care perspective but was associated with cost savings as seen from the more encompassing societal perspective, apparently because the health care costs were more than compensated by the greater productivity of the people who received SbS instead of EUC. To be precise, taking the health care perspective, SbS had an 80% probability to be regarded as cost-effective compared with EUC when there was a WTP of US $220 per additional treatment response or US $840 per additional remission. Taking the more encompassing societal perspective, SbS had a more than 75% probability to be cost-saving while gaining treatment response or remission. Access to evidence-based mental health care in Lebanon is limited due to insufficient resources and infrastructure. Digital mental health interventions offer promising alternatives to address this issue, particularly for vulnerable populations like displaced people and those affected by conflict [11]. In addition, stigma surrounding mental health services discourages individuals from seeking help. To address this, interventions like SbS offer remote support, allowing mental health patients to seek private assistance without fear of negative perceptions [27].

Limitations

Our study was not without limitations. First, the findings may have been affected by the high dropout rate. Here, it should be noted that high dropout rates were expected, as these are usually associated with digital self-help interventions [11,12,23,28,29]. In the power calculation, we therefore accounted for the high dropout. In addition, different imputation techniques for missing data produced similar results, attesting to the robustness of our findings despite the high dropout rate. Nonetheless, dropout may have influenced the study's outcomes.

Second, no clinical diagnostic interviews were used to assess depression status. Nevertheless, the PHQ-9 is a reliable instrument and was transformed in clinically relevant metrics such as treatment response and remission using well-established cutoffs.

Third, unit cost prices were mostly based on tariffs and reflected the price levels of the year 2019, which are likely to differ from current price levels in Lebanon where inflation is high.

Fourth, the effects were assessed at 3 months postintervention. Although the effects were maintained over that period, new studies are required to assess longer-term effectiveness and cost-effectiveness of digital self-help interventions and to assess the need to perhaps invite participants to return to SbS, for example after a depressive relapse.

Finally, decision-making based on a health-economic evaluation must consider that there are no universally agreed-on WTP thresholds for gaining a treatment response and remission. Ultimately, it is up to national policymakers what they consider good value for money, which may depend on factors such as the need for health care in the population, possibilities for sustained funding, and likely budget impacts in addition to medical-ethical and equity considerations.

Our study also has some notable strengths. In low- and middle-income countries, resources, expertise, and infrastructure for mental health care research, including health-economic evaluations, are limited [30]. This is one of the larger randomized trials in mental health in a low- to middle-income country and one of the very few studies to assess the cost-effectiveness of a guided, digital self-help intervention for treatment response and remission in such a context. Furthermore, the fact that the study was able to recruit 1249 participants during the COVID-19 outbreak demonstrates the high demand for a low-threshold intervention like SbS.

Comparison With Prior Work

In terms of clinical outcomes, SbS significantly reduced symptoms of depression, anxiety, and posttraumatic stress and improved functioning [11,12]. These findings were consistent with meta-analytic evidence that digital health interventions for depression offered with some guidance, like SbS, are effective in generating beneficial clinical outcomes [31]. A health-economic evaluation of Syrian refugees in Turkey compared a group-based guided self-help course for stress management with EUC and demonstrated a 97.5% probability of cost-effectiveness at a WTP of US $2802 for gaining a quality-adjusted life year [32].

Regarding the cost-effectiveness of guided digital interventions, Mitchell et al [33] performed a systematic review of 27 economic evaluations of internet-based psychological interventions for anxiety disorders and depression and found that 81% of the internet-based treatments were cost-effective. More recently, Rohrbach et al [34] conducted a systematic review and meta-analysis of 37 trial-based economic evaluations of digital health interventions for people with mental disorders compared with care as usual or “no intervention.” They observed that online therapies for mental disorders were more effective at generating health gains for similar costs as usual care across a range of mental disorders. All in all, our results appear consistent with available evidence in indicating that guided digital self-help has a high probability to be regarded as cost-effective when compared with care as usual. However, it should be noted that the reviews and meta-analyses were based on studies conducted in high-income countries, whereas our study presents evidence from a middle-low income country facing a series of overlapping humanitarian and economic crises.

Conclusions

The SbS intervention, in addition to having a statistically significant and clinically meaningful effect on depression, anxiety, and stress-related disorders [11,12], had a more than 75% probability of being cost-saving as seen from the societal perspective.

Taking the health care perspective, SbS is associated with additional costs that are acceptable when there is a WTP of US $220 to achieve a treatment responder or US $840 for gaining a remission. In short, this study shows that digital health can be seen as cost-effective or even cost-saving compared with usual care. Apart from its cost-effectiveness, digital care can bring additional benefits because it can be accessed 24/7 from any location without the need to travel to health services and with less fear of stigma. This can be particularly relevant when access to routine care is hampered.

Throughout crises, health care systems are frequently overburdened, the health care workforce health is affected, and access to specialists for support is limited. In such a context, a digital self-help intervention with limited guidance appears to offer a promising and cost-effective approach to respond to humanitarian crises in low-resource settings such as Lebanon.

Acknowledgments

Step-by-Step (SbS) was developed by the World Health Organization in collaboration with the National Mental Health Programme of the Ministry of Public Health in Lebanon, Freie Universität Berlin, and University of Zurich. The authors acknowledge, with many thanks, the contributions from the Primary Health Care department at the Ministry of Public Health in Lebanon, Dr Randa Hamadeh and Imad Haddad; the directors of the primary care centers who provided unit cost price data; and the National Mental Health Programme and SbS team of Perrine Posbic, Nour Kik, Ahmad Akar, Gaelle Fahed, Layal Hamze, Maria Hayek, Sally Khoury, Yara Hasbany, Jihane bou Sleiman, Ghida Anani, and Charbel Zgheib. The authors alone are responsible for the views expressed in this article, and they do not necessarily represent the views, decisions, or policies of the institutions with which they are affiliated.

Fondation d'Harcourt and Research for Health in Humanitarian Crises (R2HC, managed by Elrha) funded this research. The research was conducted using an open-source software platform established by Freie Universität Berlin. The development of the SbS software platform was funded by the EU H2020 STRENGTHS project, European Union’s Horizon 2020 Research and Innovation Programme Societal Challenges under grant agreement No 733337. The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The opinions expressed in this paper are those of the authors and do not necessarily represent the decisions, policies, nor views of the organizations they serve or of the funder. The European Community is not liable for any use that may be made of the information contained therein.

Data Availability

Deidentified data are available at the open access DANS data repository [35] at [36] and can be acquired through the data management committee of the Faculty of Behavioural and Movement Sciences of the Vrije Universiteit Amsterdam, which can be reached at data.committee.fgb@vu.nl.

Authors' Contributions

All authors contributed to the study. MvO and CK acquired the funding. EH, JAR, KC, EvH, EZ, MvO, and REC were project administrators. REC, KC, CK, SB, EvH, EH, PC, and MvO contributed to the study design. FS, EH, JAB, RAH, and SB adapted the cost questionnaire to the context of this study. RAH, EH, EvH, PN, JAR, and EZ contributed to the data collection and data curation. FS and BW performed the statistical analysis. RAH and FS drafted the manuscript. REC and FS were joint last authors. All authors revised the text critically for important intellectual content and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Mean costs in US $ per participant over time by condition, not including US $26 Step-by-Step costs (n=1249).

[DOCX File , 17 KB - mental_v11i1e55544_app1.docx]

References

1. Syria Regional Refugee Response. United Nations High Commissioner for Refugees (UNHCR) Operational Data Portal (ODP). URL: https://data.unhcr.org/en/situations/syria/location/71 [accessed 2022-09-30]
2. Charlson F, van Ommeren M, Flaxman A, Cornett J, Whiteford H, Saxena S. New WHO prevalence estimates of mental disorders in conflict settings: a systematic review and meta-analysis. The Lancet 2019 Jul;394(10194):240-248. [doi: 10.1016/s0140-6736(19)30934-1]
3. Fouad FM, Barkil-Oteo A, Diab JL. Mental health in Lebanon's triple-fold crisis: the case of refugees and vulnerable groups in times of COVID-19. Front Public Health 2020 Jan 20;8:589264 [FREE Full text] [doi: 10.3389/fpubh.2020.589264] [Medline: 33553090]
4. Kawa N, Abisaab J, Abiad F, Badr K, El-Kak F, Alameddine M, et al. The toll of cascading crises on Lebanon's health workforce. The Lancet Global Health 2022 Feb;10(2):e177-e178. [doi: 10.1016/s2214-109x(21)00493-9]
5. The National Mental Health Programme. Republic of Lebanon Ministry of Public Health. URL: http://www.moph.gov.lb/en/Pages/6/553/nmhp [accessed 2022-12-08]
6. Carswell K, Harper-Shehadeh M, Watts S, Van't Hof E, Abi Ramia J, Heim E, et al. Step-by-Step: a new WHO digital mental health intervention for depression. Mhealth 2018 Aug;4:34-34 [FREE Full text] [doi: 10.21037/mhealth.2018.08.01] [Medline: 30225240]
7. Abi Ramia J, Harper Shehadeh M, Kheir W, Zoghbi E, Watts S, Heim E, et al. Community cognitive interviewing to inform local adaptations of an e-mental health intervention in Lebanon. Glob Ment Health (Camb) 2018 Nov 27;5:e39 [FREE Full text] [doi: 10.1017/gmh.2018.29] [Medline: 30637112]
8. Harper Shehadeh MJ, Abi Ramia J, Cuijpers P, El Chammay R, Heim E, Kheir W, et al. Step-by-Step, an e-mental health intervention for depression: a mixed methods pilot study from Lebanon. Front Psychiatry 2019 Feb 12;10:986 [FREE Full text] [doi: 10.3389/fpsyt.2019.00986] [Medline: 32116815]
9. Heim E, Ramia JA, Hana RA, Burchert S, Carswell K, Cornelisz I, et al. Step-by-step: Feasibility randomised controlled trial of a mobile-based intervention for depression among populations affected by adversity in Lebanon. Internet Interv 2021 Apr;24:100380 [FREE Full text] [doi: 10.1016/j.invent.2021.100380] [Medline: 33747798]
10. van 't Hof E, Heim E, Abi Ramia J, Burchert S, Cornelisz I, Cuijpers P, et al. Evaluating the effectiveness of an e-mental health intervention for people living in Lebanon: protocol for two randomized controlled trials. JMIR Res Protoc 2021 Jan 28;10(1):e21585 [FREE Full text] [doi: 10.2196/21585] [Medline: 33507158]
11. Cuijpers P, Heim E, Abi Ramia J, Burchert S, Carswell K, Cornelisz I, et al. Effects of a WHO-guided digital health intervention for depression in Syrian refugees in Lebanon: A randomized controlled trial. PLoS Med 2022 Jun 23;19(6):e1004025 [FREE Full text] [doi: 10.1371/journal.pmed.1004025] [Medline: 35737665]
12. Cuijpers P, Heim E, Ramia JA, Burchert S, Carswell K, Cornelisz I, et al. Guided digital health intervention for depression in Lebanon: randomised trial. Evid Based Ment Health 2022 Dec 16;25(e1):e34-e40 [FREE Full text] [doi: 10.1136/ebmental-2021-300416] [Medline: 35577536]
13. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
14. Ustun TB, Kostanjesek N, Chatterji S, Rehm J, World Health Organization. Measuring health and disability: manual for WHO Disability Assessment Schedule (WHODAS 2.0). World Health Organization. 2010. URL: https://tinyurl.com/4e4s8ht7 [accessed 2024-05-02]
15. Scaling up care for mental, neurological, and substance use disorders: mhGAP. World Health Organization. URL: https://www.who.int/activities/scaling-up-mental-health-care [accessed 2024-05-02]
16. Sawaya H, Atoui M, Hamadeh A, Zeinoun P, Nahas Z. Adaptation and initial validation of the Patient Health Questionnaire - 9 (PHQ-9) and the Generalized Anxiety Disorder - 7 Questionnaire (GAD-7) in an Arabic speaking Lebanese psychiatric outpatient sample. Psychiatry Res 2016 May 30;239:245-252. [doi: 10.1016/j.psychres.2016.03.030] [Medline: 27031595]

17. Urtasun M, Daray FM, Teti GL, Coppolillo F, Herlax G, Saba G, et al. Validation and calibration of the patient health questionnaire (PHQ-9) in Argentina. BMC Psychiatry 2019 Sep 18;19(1):291 [FREE Full text] [doi: 10.1186/s12888-019-2262-9] [Medline: 31533674]
18. van Roijen L, van Straten A, Tiemens B, Donker MCH. Manual Trimbos/iMTA Questionnaire for Costs Associated with Psychiatric Illness (TIC-P) (in Dutch). ResearchGate. 2002 Jan. URL: https://tinyurl.com/yrd5uue6 [accessed 2024-05-14]
19. Bouwmans C, De Jong K, Timman R, Zijlstra-Vlasveld M, Van der Feltz-Cornelis C, Tan Swan S, et al. Feasibility, reliability and validity of a questionnaire on healthcare consumption and productivity loss in patients with a psychiatric disorder (TiC-P). BMC Health Serv Res 2013 Jun 15;13(1):217 [FREE Full text] [doi: 10.1186/1472-6963-13-217] [Medline: 23768141]
20. Chisholm D, Knapp MRJ, Knudsen HC, Amaddeo F, Gaite L, van Wijngaarden B. Client Socio-Demographic and Service Receipt Inventory--European Version: development of an instrument for international research. EPSILON Study 5. European Psychiatric Services: Inputs Linked to Outcome Domains and Needs. Br J Psychiatry Suppl 2000 Jan 02;177(39):s28-s33. [doi: 10.1192/bjp.177.39.s28] [Medline: 10945075]
21. Agency for Technical Cooperation and Development Labour Market Assessment in Beirut and Mount Lebanon. United Nations High Commissioner for Refugees (UNHCR) Operational Data Portal (ODP). 2014. URL: https://data2.unhcr.org/en/documents/details/46027 [accessed 2024-05-02]
22. Husereau D, Drummond M, Augustovski F, de Bekker-Grob E, Briggs AH, Carswell C, CHEERS 2022 ISPOR Good Research Practices Task Force. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations. BMC Med 2022 Jan 12;20(1):23 [FREE Full text] [doi: 10.1186/s12916-021-02204-0] [Medline: 35022047]
23. Cuijpers P, Karyotaki E, Reijnders M, Purgato M, Barbui C. Psychotherapies for depression in low- and middle-income countries: a meta-analysis. World Psychiatry 2018 Feb 19;17(1):90-101 [FREE Full text] [doi: 10.1002/wps.20493] [Medline: 29352530]
24. Demirtas H. Simulation driven inferences for multiply imputed longitudinal datasets*. Statistica Neerlandica 2004 Dec 06;58(4):466-482. [doi: 10.1111/j.1467-9574.2004.00271.x]
25. Brand J, van Buuren S, le Cessie S, van den Hout W. Combining multiple imputation and bootstrap in the analysis of cost-effectiveness trial data. Stat Med 2019 Jan 30;38(2):210-220 [FREE Full text] [doi: 10.1002/sim.7956] [Medline: 30207407]
26. Stata. URL: https://www.stata.com/ [accessed 2024-05-03]
27. Abi Hana R, Arnous M, Heim E, Aeschlimann A, Koschorke M, Hamadeh RS, et al. Mental health stigma at primary health care centres in Lebanon: qualitative study. Int J Ment Health Syst 2022 May 07;16(1):23 [FREE Full text] [doi: 10.1186/s13033-022-00533-y] [Medline: 35525972]
28. van Ballegooijen W, Cuijpers P, van Straten A, Karyotaki E, Andersson G, Smit JH, et al. Adherence to internet-based and face-to-face cognitive behavioural therapy for depression: a meta-analysis. PLoS One 2014 Jul 16;9(7):e100674 [FREE Full text] [doi: 10.1371/journal.pone.0100674] [Medline: 25029507]
29. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: A systematic and meta-analytic review. J Consult Clin Psychol 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
30. Levin C, Chisholm D. Cost-Effectiveness and Affordability of Interventions, Policies, and Platforms for the Prevention and Treatment of Mental, Neurological, and Substance Use Disorders. In: Patel V, Chisholm D, Dua T, Laxminarayan R, Medina-Mora ME, editors. Disease Control Priorities, Third Edition (Volume 4): Mental, Neurological, and Substance Use Disorders. Washington DC: The International Bank for Reconstruction and Development / The World Bank; 2016:219-236.
31. Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry 2019 Oct 09;18(3):325-336 [FREE Full text] [doi: 10.1002/wps.20673] [Medline: 31496095]
32. Park A, Waldmann T, Kösters M, Tedeschi F, Nosè M, Ostuzzi G, et al. Cost-effectiveness of the self-help plus intervention for adult Syrian refugees hosted in Turkey. JAMA Netw Open 2022 May 02;5(5):e2211489 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.11489] [Medline: 35536574]
33. Mitchell LM, Joshi U, Patel V, Lu C, Naslund JA. Economic evaluations of internet-based psychological interventions for anxiety disorders and depression: a systematic review. J Affect Disord 2021 Apr 01;284:157-182 [FREE Full text] [doi: 10.1016/j.jad.2021.01.092] [Medline: 33601245]
34. Rohrbach PJ, Dingemans A, Evers C, Van Furth EF, Spinhoven P, Aardoom JJ, et al. Cost-effectiveness of internet interventions compared with treatment as usual for people with mental disorders: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2023 Jan 05;25:e38204 [FREE Full text] [doi: 10.2196/38204] [Medline: 36602854]
35. Data Archiving and Networked Services (DANS). URL: https://dans.knaw.nl/en/ [accessed 2024-05-02]
36. Step by step trial data, Syrian displaced people in Lebanon. Data Archiving and Networked Services (DANS). URL: https://doi.org/10.17026/dans-zey-hjpk [accessed 2024-05-02]

Abbreviations

CHEERS: Consolidated Health Economic Evaluation Reporting Standards
CSRI: Client Service Receipt Inventory
EUC: enhanced usual care
FTC: full-time equivalent
ICER: incremental cost-effectiveness ratio
iMTA: Institute of Medical Technology Assessment
MICE: multiple imputation with chained equations
PHQ-9: Patient Health Questionnaire (9-item version)
SbS: Step-by-Step
SURE: seemingly unrelated regression equations
TiC-P: Trimbos and iMTA Cost questionnaire for Psychiatric illness
WHO: World Health Organization
WTP: willingness to pay

Edited by J Torous; submitted 16.12.23; peer-reviewed by R Taher, NH Zainal; comments to author 26.02.24; revised version received 22.03.24; accepted 08.04.24; published 29.05.24.

Please cite as:
Abi Hana R, Abi Ramia J, Burchert S, Carswell K, Cuijpers P, Heim E, Knaevelsrud C, Noun P, Sijbrandij M, van Ommeren M, van't Hof E, Wijnen B, Zoghbi E, El Chammay R, Smit F
Cost-Effectiveness of Digital Mental Health Versus Usual Care During Humanitarian Crises in Lebanon: Pragmatic Randomized Trial
JMIR Ment Health 2024;11:e55544
URL: https://mental.jmir.org/2024/1/e55544
doi:10.2196/55544
PMID:38810255

Original Paper

Characterizing Longitudinal Patterns in Cognition, Mood, And Activity in Depression With 6-Week High-Frequency Wearable Assessment: Observational Study

Francesca Cormack[1,2,3], PhD; Maggie McCue[4], MSc; Caroline Skirrow[1,5], PhD; Nathan Cashdollar[1], PhD; Nick Taptiklis[1], BA; Tempest van Schaik[6], PhD; Ben Fehnert[3,6,7], BSc; James King[3,6,7], MA; Lambros Chrones[4], MD; Sara Sarkey[4], PhD; Jasmin Kroll[1], PhD; Jennifer H Barnett[1,2], PhD

[1]Cambridge Cognition, Cambridge, United Kingdom
[2]Department of Psychiatry, University of Cambridge, Cambridge, United Kingdom
[3]Cognition Kit, Cambridge, United Kingdom
[4]Takeda Pharmaceuticals USA Inc, Lexington, MA, United States
[5]Department of Psychological Science, University of Bristol, Bristol, United Kingdom
[6]Ctrl Group, London, United Kingdom
[7]Fora Health, London, United Kingdom

Corresponding Author:
Jasmin Kroll, PhD
Cambridge Cognition
Tunbridge Court
Bottisham
Cambridge, CB25 9TU
United Kingdom
Phone: 44 7961910560
Email: jasmin.kroll@camcog.com

Abstract

Background: Cognitive symptoms are an underrecognized aspect of depression that are often untreated. High-frequency cognitive assessment holds promise for improving disease and treatment monitoring. Although we have previously found it feasible to remotely assess cognition and mood in this capacity, further work is needed to ascertain the optimal methodology to implement and synthesize these techniques.

Objective: The objective of this study was to examine (1) longitudinal changes in mood, cognition, activity levels, and heart rate over 6 weeks; (2) diurnal and weekday-related changes; and (3) co-occurrence of fluctuations between mood, cognitive function, and activity.

Methods: A total of 30 adults with current mild-moderate depression stabilized on antidepressant monotherapy responded to testing delivered through an Apple Watch (Apple Inc) for 6 weeks. Outcome measures included cognitive function, assessed with 3 brief n-back tasks daily; self-reported depressed mood, assessed once daily; daily total step count; and average heart rate. Change over a 6-week duration, diurnal and day-of-week variations, and covariation between outcome measures were examined using nonlinear and multilevel models.

Results: Participants showed initial improvement in the Cognition Kit N-Back performance, followed by a learning plateau. Performance reached 90% of individual learning levels on average 10 days after study onset. N-back performance was typically better earlier and later in the day, and step counts were lower at the beginning and end of each week. Higher step counts overall were associated with faster n-back learning, and an increased daily step count was associated with better mood on the same (P<.001) and following day (P=.02). Daily n-back performance covaried with self-reported mood after participants reached their learning plateau (P=.01).

Conclusions: The current results support the feasibility and sensitivity of high-frequency cognitive assessments for disease and treatment monitoring in patients with depression. Methods to model the individual plateau in task learning can be used as a sensitive approach to better characterize changes in behavior and improve the clinical relevance of cognitive data. Wearable technology allows assessment of activity levels, which may influence both cognition and mood.

(JMIR Ment Health 2024;11:e46895) doi:10.2196/46895

KEYWORDS

cognition; depression; digital biomarkers; ecological momentary assessment; mobile health; remote testing

Introduction

Background

Major depressive disorder (MDD) is a debilitating condition and the leading cause of disease burden worldwide [1,2]. MDD is characterized primarily by low mood, or reduced interest and pleasure in daily activities [3]. Cognitive deficits are a substantial problem in patients with MDD, with reported impairments in a range of domains, including processing speed, attention, executive function, learning, and memory [4-6]. Despite impaired cognition being widely reported, further studies using objective neuropsychological measures are required [7].

Self-Assessment of MDD

MDD is typically assessed using retrospective self-report, where patients reflect on their experiences over a period of days or weeks. However, this method of reporting is subject to a variety of memory distortions [8,9], and depression itself is linked to impaired recollection, memory bias, and overgeneralization [10]. A direct comparison between retrospective recall and repeated real-time assessments has shown negative emotional biases in patients with depression [11,12] which persist beyond the depressive episode in a subset of patients [13]. Various explanatory factors have been proposed, such as the number of previous episodes and demographic factors, but overall results have been inconsistent [14].

If patients' recollections correspond poorly with their actual experiences, this is likely to distort our understanding of disease course and treatment response [15]. Discrepancies between objectively measured cognitive function and patients' self-report have been demonstrated, with the latter being influenced by depressed mood [16-18]. This may be true even in cases of significant cognitive impairment, where self-reported cognitive function is associated with subjective complaints of depressive symptoms but not objective cognitive outcomes [19]. Hence, it is plausible that self-reported and objective assessments are measuring different cognitive capabilities, with objective measures needed to reveal underlying cognitive function.

Results from cognitive tests may also vary in relation to within-individual differences, including dietary effects and sleep-wake cycles [20,21]. These variances can make it difficult to differentiate clinically meaningful change from measurement error [22]. Higher frequency sampling is thought to generate more stable and reliable estimates of constructs of interest [22], by reducing state effects on punctual relationships, which can obscure the signal in study interventions, and by narrowing the margin of error [23].

Real-Time Measurement of Cognition, Mood, and Activity

Advances in portable technology have enabled the precise, unobtrusive recording of real-time psychological, behavioral, and physiological measures. Higher frequency assessments in the context of everyday life can identify real life changes that are associated with clinical improvement [24], improve sensitivity for detecting change [25] and help to identify shifts in depressive symptoms [26]. Additionally, this approach allows for the characterization of the temporal relationship of symptoms over time in relation to changes in an individual's behavior and environmental influences [27,28] creating a profile of the dynamic relationships between cognitive function, psychological processes, and biological processes of an individual [29].

However, this approach also has important implications for sampling strategy and data analysis. Diurnal changes in mood and affect have been reported in individuals with depression, where negative symptoms are more prevalent in the morning [30] and increased positive affect is seen later in the day [31,32]. Further, research indicates cognitive function follows the same pattern as daily changes in mood, with worse performance in the morning and better performance in the evening on a range of cognitive measures, including memory, attention, and psychomotor speed [33]. In the general population, more positive and less negative moods are typically seen on weekends, with improvements starting on Fridays [34], while more activity, measured by higher step counts, is noted during weekdays in comparison with weekends [35,36]. To account for the complexity of these fluctuations and how patients themselves may experience depression, novel data collection methods, such as ecological momentary assessments, are a promising opportunity to address this gap. Current scientific literature is mostly laboratory-based. Although ecological momentary assessment studies have shed light on various aspects of the disease, such as rumination and emotion reactivity [37], however, less is known about the relationship between cognition function and mood in MDD.

Goal of This Study

Here we examine data from a 6-week feasibility study of daily cognitive, mood, and activity assessment in adults with mild-to-moderate depression stabilized on antidepressant treatment. Previously, in this same sample, we demonstrated excellent compliance with daily cognitive and mood assessments on a smart watch and good agreement with validated full-length cognitive and self-report measures [38]. High adherence rates were found (95%) with no deterioration over the course of the study. Adherence was not associated with depressive symptoms or cognitive functioning.

In this study, we aim to optimize methods for analyzing high-frequency longitudinal data and characterize relationships between cognition, mood, and activity data in order to facilitate future intervention studies. This study aims to examine (1) longitudinal changes in mood, cognition, activity levels, and heart rate over 6 weeks; (2) diurnal and weekday-related changes; and (3) the cooccurrence of fluctuations between mood, cognitive function, and activity.

Methods

This was an observational study aimed at characterizing longitudinal patterns in cognition, mood, and activity in patients with MDD over 6 weeks with high-frequency assessment enabled by a wearable device.

Participants

Full recruitment details have previously been reported by Cormack et al [38], who demonstrated the feasibility of high-frequency testing for a prolonged period of 6 weeks in participants with depression. In brief, participants were recruited for a primary psychiatric diagnosis of MDD with mild to moderate depression (as defined by Patient Health Questionnaire [PHQ]-9 scores between 5 and 15, inclusive). They were recruited through a patient recruitment company with links to primary care providers and patients with depression groups. Participants aged between 18 and 65 years, inclusive, who were able to read and understand English, were eligible for participation. Individuals were excluded if they had a personal history of another psychiatric disorder (except nonprimary anxiety); mental, neurological, or neurodegenerative disorders; or substance abuse or dependence. In total, 30 participants with MDD receiving antidepressant monotherapy treatment were enrolled.

Measures

Cognitive function and mood were assessed with the Cognition Kit app (joint venture between Ctrl Group and Cambridge Cognition), loaded onto the Apple Watch (high-resolution touch-screen watch), and paired with an iPhone. Participants were asked to wear study equipment between 8 AM and 10 PM for 6 weeks. This duration was chosen as it is the time it typically takes for the improvement in mood to be seen following the start of administration of the antidepressant medication.

Prompts for cognitive assessment were given 3 times daily (morning, afternoon, and evening). Cognitive function was assessed with the Cognition Kit N-Back test, which has shown sensitivity to impairments in MDD [39]. Research suggests that n-back task performance is a marker of cognitive function, including aspects of working memory, task switching, and attention [40]. During test administration, 9 symbols randomly selected from a pool of 227 were presented briefly, one at a time, over 30 trials. Participants were asked to respond when any symbol was the same as the one presented 2 trials previously. The primary outcome measure was d-prime (the ratio of hits [correct detection of an n-back match] to false alarms [response during no match]). Under the current implementation of the paradigm, values of d-prime range from –3.33 to 3.33. Test-retest reliability of high frequency testing using n-back d-prime is 0.8 in a mixed sample of participants with neurodegenerative disease, immune-mediated inflammatory disorders, and healthy controls [41]. Each assessment took 30 seconds to complete, after which participants were shown their test score.

Prompts for mood assessment were given up to twice daily (afternoon and evening), with no prompt delivered in the evening if participants had completed the assessment in the afternoon. Mood was assessed with the following three questions adapted from the PHQ-2 and Perceived Deficits Questionnaire: How much have the following problems bothered you over the past day? (1) Lack of interest or pleasure in doing things; (2) Feeling down, depressed, or hopeless; and (3) Trouble concentrating on things (eg, newspapers or television). Responses were coded on a 4-point scale of severity of symptoms (1=no problem to 4=greatly). Responses were summated to provide a summary mood score. The PHQ-2, a short, validated form of the PHQ-9, has high accuracy and sensitivity to depression [42].

Total daily step count and average heart rate, measures associated with depressive symptom severity, [43,44], were acquired passively through the Apple Watch. The Apple Watch has been found to be a reliable and valid tool for assessing heart rate variability and is extremely accurate as a daily step count [45,46]. Similarity the n-back test has been previously used in wearable technology studies that incorporated high-frequency testing, demonstrating its feasibility [38,47].

Procedure

Participants attended the study site, where researchers introduced the devices and tasks to participants, who were given the opportunity to practice using the software and hardware and ask questions. Testing was completed in the following 6 weeks (42 days), with data uploaded to secure study servers when transfer through Wi-Fi or roaming was possible. The study was completed with a home visit at 6 weeks, during which study hardware was returned.

Daily assessments were completed as part of a larger test battery, including full-length rating scales, and cognitive tests, and a semistructured qualitative interview, as described previously [38].

Statistical Analysis

Data Cleaning

Data were harmonized by defining the first study day as the first day in which participants responded to assessment prompts for cognitive and mood assessments. Earlier days, only passive data were excluded from analyses. For activity and heart rate measures, non-wearing days (defined as days with <100 recorded steps [48,49], or where no heart rate was recorded) were excluded from the analyses (25 observations out of 1160).

Response frequency within each 1-hour period collapsed across days (eg, 6 AM-6:59 AM, 7 AM-7:59 AM, and 8 AM-8:59 AM) was examined separately for mood and d-prime to identify periods with sparse data. One-hour periods with n<25 assessments overall were excluded. This left n-back data from 6 AM-11:59 PM (n=29 observations dropped) and mood data from 12 noon until 5:59 PM (n=98 observations dropped). For all included 1-hour periods, there were between 41-287 data points available (mean 182 mood, 180 n-back assessments per 1-hour period).

Change Over Time

Since outcome measures are sampled repeatedly, hierarchical models are required to account for observations nested within individuals [50]. Random effects of participants with random intercept and random slope were applied to all models to control for between-participant variability, allowing the intercept and regression coefficient to vary between participants [51]. Model parameters were estimated using maximum likelihood, and model fits were compared with model likelihood ratios. This allowed the direct comparison of models with different fixed-effect structures using chi-square analysis and the selection of the best-fitting model. Diagnostic tests included examination of the normality of residuals and their spread for each predictor variable.

Raw data were transformed using log and square root transformations into normally distributed data as appropriate. A series of longitudinal mixed-effects models examined change in each outcome measure over the 6-week duration of the study to identify time-related trends in outcome measures. Each outcome measure was designated as the response variable, and a fixed effect of time-on-task was specified. Intercept-only models and linear, quadratic, and cubic trends of time-on-task were examined. For unequally spaced timepoints (cognitive assessment or mood), a continuous autoregressive correlation structure was applied. For summated daily indices (step count or heart rate), a first-order autoregressive covariance structure was used.

Diurnal and Weekday Effects

Fixed effects of diurnal variation or weekday were examined by appending these as additional predictors (fixed effects) to the best-fitting models. The time of day was treated as a continuous variable. Weekday was dummy coded from 1 to 7 (1=Monday to 7=Sunday). As in previous work, the quadratic and linear fixed effects of time of day and day of week were examined [32]. For measures with variation both in time-of-day and day-of-week (d-prime and mood), models first examined the effects of time-of-day and then incorporated weekday effects.

Characterizing Individual N-Back Learning Curves

N-back learning curves were characterized following previously described methods [52]. This includes (1) the “starting point,” the level at which performance begins; (2) the “asymptote,” the theoretically best score achievable toward which task performance tends with unlimited assessments; (3) the “slope,” the rate at which learning occurs; and (4) the “learning rate,” how quickly a prespecified level of performance is reached.

Intercepts extracted for each participant from the best-fitting mixed model for change over time using the analysis steps described above were implemented as individual starting points. These were subtracted from d-prime at each assessment occasion to provide a baseline-adjusted d-prime. As in previous work, the data were smoothed [53] by applying a 3-assessment moving average to the data series, corresponding to the typical number of assessments per day.

For each individual participant, nonlinear regression was used to fit an inverse curve ($Y=a–(b/X)$), yielding personalized estimates for a (asymptote, the theoretically best score achievable: as $X \rightarrow \infty$, $Y \rightarrow a$) and b (slope) for baseline adjusted d-prime (Y) over each consecutive n-back assessment (X). A 90% learning rate was defined as the number of trials required by each individual participant to reach 90% of their potential beyond their starting point ($Y=0.9a$ when $X=10*b/a$) [52]. The stable maximum d-prime was calculated for participants by summating their asymptotes and intercepts.

Exploratory correlations of individual asymptotes and slopes were completed with summary measures from mood and activity assessments. The data were first examined for normality, and parametric and nonparametric correlations were completed as appropriate.

Covariation of Fluctuations With Mood

Fluctuations over time were quantified using random effects. Residuals derived from the best-fitting hierarchical models described above, regressing out significant effects of time-on-task (to control for practice effects), diurnal effects, or weekday effects as relevant to each outcome measure. Residuals reflect the deviation of each individual from their own slope at each moment in time (better or worse, or higher or lower), quantifying the difference between each observation and what would be statistically expected. Since all measures, with the exception of d-prime, reflected overall daily measures, mean daily residuals for d-prime were computed by averaging d-prime residuals for each day, providing a consistent time scale for covariation analyses. Hierarchical models were used to examine covariation between measures. All models specified a first-order autoregressive covariance structure and random intercepts. Diagnostic tests included examination of the normality of residuals and their spread for each predictor variable.

A step-by-step approach was taken to examine the covariation of mood with d-prime and other assessment domains. First, a mixed model was used to predict fluctuations in mood based on fluctuations within all other domains (fixed effects: d-prime, activity, and heart rate). Next, the direction of causality was examined using a time-lagged approach. Analyses examined whether the fluctuations in mood were predicted by fluctuations in the predictor variable on the previous or following day. Only outcome measures that were significant predictors in concurrent covariation analyses were taken forward and included in the lagged models.

Analyses were repeated after excluding the period during which 90% of n-back learning took place (computed from individual participant learning curves), with the exception of participants who showed no significant slope where all data were included. This helped to identify correlation between measures after the influences of task learning were minimized. Where significant associations with mean daily d-prime fluctuations were seen, analyses were repeated when constrained only to days in which 3 n-back assessments were available. This aimed to test whether covariation between d-prime and mood was affected by regression to the mean (where days with fewer n-back assessments are likely to show greater variability).

Ethical Considerations

The study was reviewed and approved by the Proportionate Review Sub-Committee of the Wales Research Ethics Committee at Swansea University (17/WA/0042) and performed in accordance with the current version of the Declaration of Helsinki. All participants provided written informed consent before being enrolled.

Results

Sample Characteristics

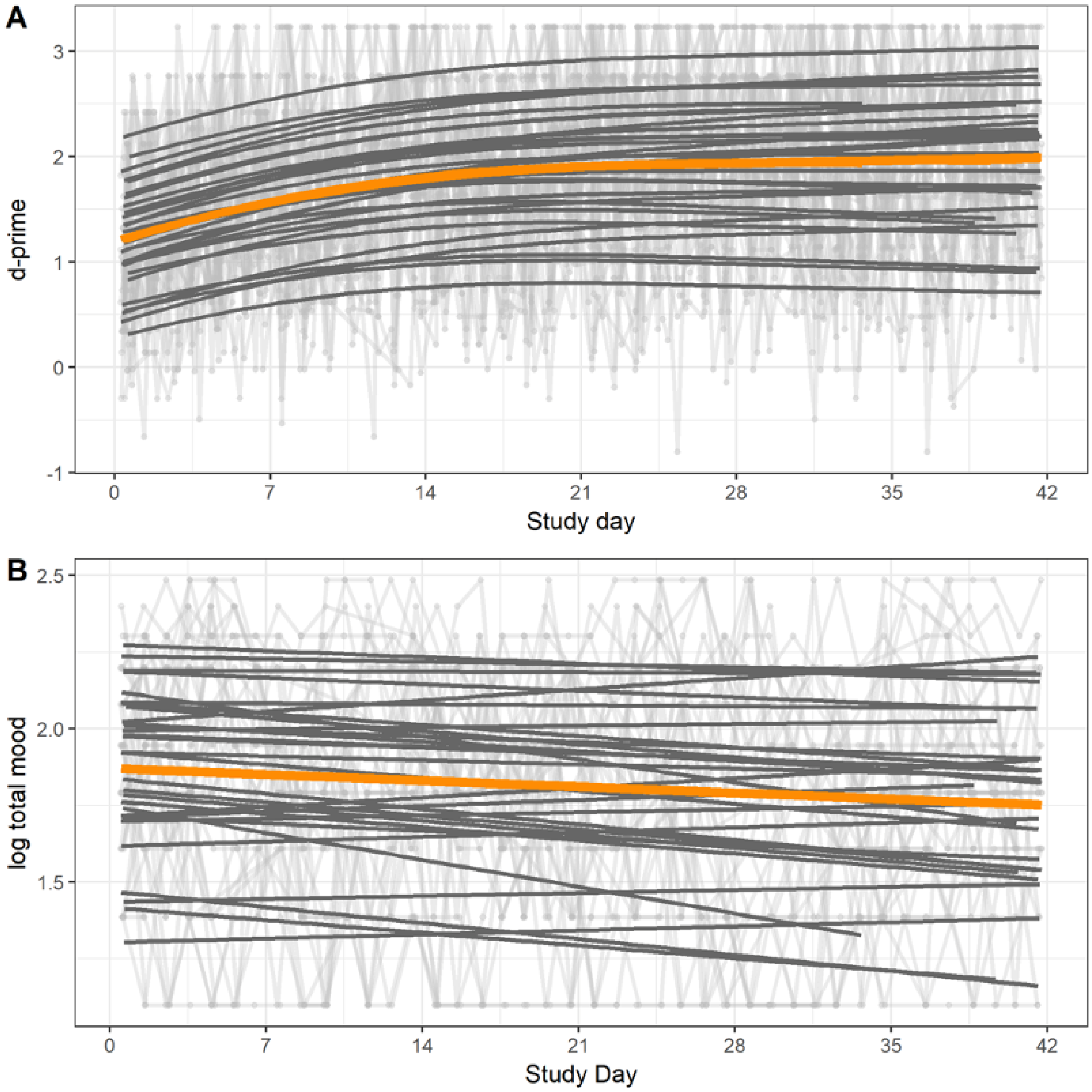
The final sample (N=30) included 19 women and 11 men, aged between 19 and 63 years (mean 37.2, SD 10.4 years). All participants received antidepressant monotherapy and had been on their current medication for an average of 9.9 (SD 9.5 months; range 0.4-94.3). Current medications included selective serotonin reuptake inhibitors (n=20), serotonin and norepinephrine reuptake inhibitors (n=5), tricyclic antidepressants (n=4), and serotonin antagonists and reuptake inhibitors (n=1). The mean depression symptom severity as measured by the PHQ-9 was 9.1 (SD 3.1; range 5-15).

Modelling Changes in Mood, Cognition, Activity Levels, and Heart Rate Over 6 Weeks

Model selection statistics and resultant model parameters are presented in Table S1 in Multimedia Appendix 1. A cubic trend provided the best fit for change in cognitive performance over 6 weeks. Similarly, a linear trend provided the best fit for within-subjects change in mood over 6 weeks, in which depressive symptoms showed a subtle reduction over the course of the study. Heart rate and activity levels showed no overall change over time.

Data for d-prime and mood are presented in Figure 1. Here, individual daily scores (light gray lines) are shown alongside fitted random effects (bold gray) and the fitted fixed effect (orange). This illustrates the variability across participants in the direction and magnitude of change over the 6 weeks and the remaining variability from these model fits.

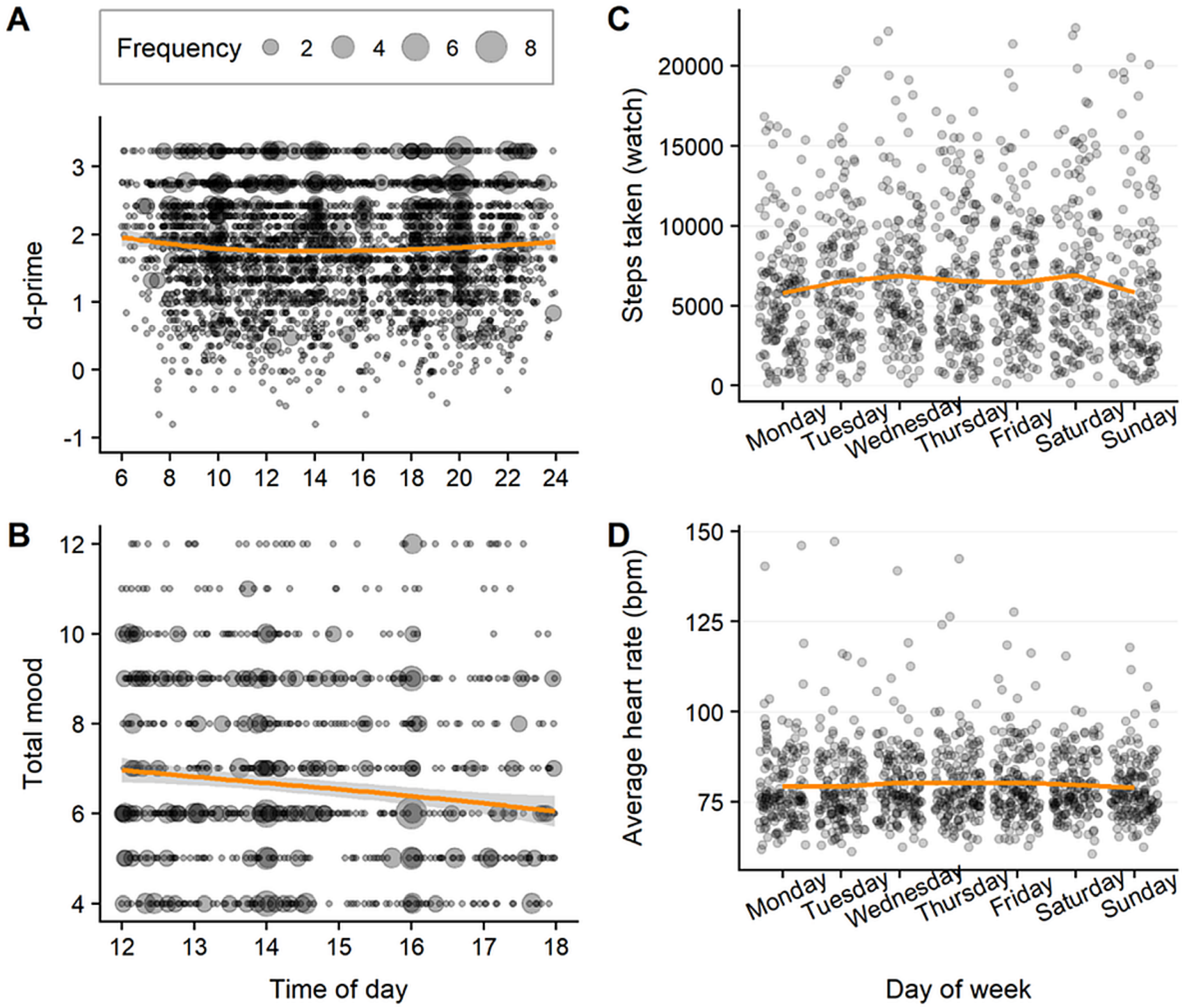
Figure 1. Change over study duration (A) in d-prime (up to 3 daily) is shown on the y-axis, with higher scores equating to better performance and (B) total mood as reflected by the y-axis (higher scores equate to a more depressed mood): random effects (bold gray lines), fixed effects (orange lines), model fits, and individual scores (pale gray lines) for individual participants.



Diurnal and Weekday-Related Changes

D-prime showed a significant quadratic effect of time-of-day, with better performance seen first thing in the morning and later at night (Figure 2A; time-of-day: estimate=–0.02, SE 0.01; *t*=–2.33; *P*=.02; time-of-day: estimate=0.001, SE 0.0006; *t*=2.48; *P*=.01). The model including time-of-day improved model fit (fit statistics: Akaike information criterion [AIC]=5952.15, Bayesian information criterion [BIC]=6019.10, likelihood ratio χ^2=6.16; *P*=.01). No significant linear or quadratic effect of the weekday was observed (*P*≥.15).

Figure 2. Diurnal and weekday effects in data: (A) Bubble chart of d-prime over time-of-day from early morning until midnight with loess regression line, with higher scores showing better performance; (B) bubble chart of total mood over time-of-day from noon until evening with loess regression line, with higher total mood denoting more severe depressive symptoms; (C) total steps taken over weekdays with mean line; and (D) average heart rate (beats per minute) over weekdays with mean line.



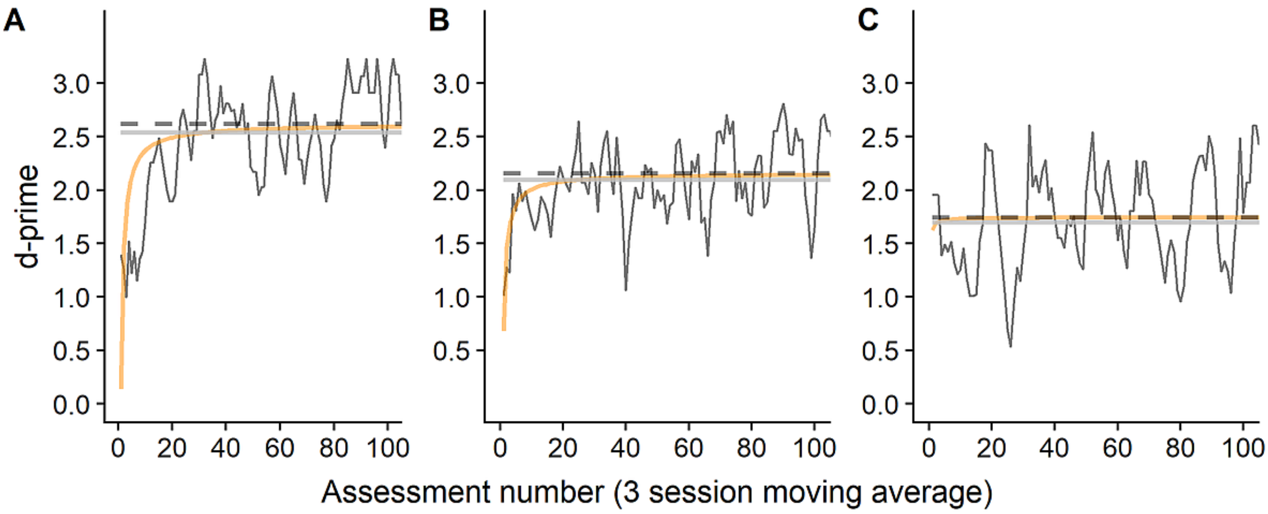
Mood showed a subtle improvement over the course of the day, which did not reach significance thresholds and did not improve model fit (Figure 2B; time-of-day: estimate=–0.01, SE 0.006]; *P*=.06; model fit statistics: AIC=436.22, BIC=476.20, likelihood ratio=3.55; *P*=.06). The effects of day-of-week were also not significant (*P*≥.14).

Significant changes in activity and heart rate were also found over weekdays (Figure 2C-2D). Step counts were lower at the week onset and end (weekday: estimate=3.52, SE=1.50; *t*=3.52; *P*=.004; weekday: estimate=–0.64, SE 0.17; *t*=–3.66; *P*=.003). As expected, including these quadratic weekday effects improved model fit (fit statistics: AIC=10131.19, BIC=10166.41, likelihood ratio=–5058.60; *P*=.002). Similarly, heart rate showed a subtle quadratic effect of weekday with lower heart rates recorded on week onset and end (weekday: estimate=0.01, SE=0.006; *t*=1.96; *P*=.05, weekday: estimate=–0.002, SE=0.0008; *t*=–1.98; *P*=.05), with improved model fit (AIC=–2312.81, BIC=–2272.56, likelihood ratio=3.91; *P*=.05).

Characterizing the N-Back Learning Curve

An adequate fit of the inverse learning curve model was seen for 27 out of 30 participants of whom estimates for both slope and asymptote were significant (minimum *t*=2.18; *P*<.04). In participants with adequate fit, 90% learning rates were reached after an average of 22.4 assessments (range 13-31), occurring after a mean of 10 days (range 6-24). Participants with data showing poor fit displayed a flatter trajectory and lower slope over the period of assessment than those with a better fit (Figure 3). Stable maximal d-prime was seen at a mean of 1.94 (range 0.74-2.93), with 4 individuals having maximums within the top quintile of positive n-back scores (at 2.59 and above) and no participants with 95% CIs incorporating the highest possible score.

Figure 3. Examples of individual learning curves (dark gray lines), and inverse curves for n-back performance (orange line; y=asymptote–(slope/x)) over period of assessment. Asymptotes are shown in black dashed lines, while 90% learning rate is shown in pale gray. Subjects A (slope=2.48) and B (slope =1.47) show moderate and adequate fit for the inverse curve, while a poor fit is seen for subject C (slope=0.12).



A total of 3 participants did not show a learning effect over the period of assessment, characterized by a nonsignificant slope. These participants performed overall above chance in n-back assessments (mean scores between 0.9 and 1.7), indicating adequate understanding of test objectives but no clear learning pattern.

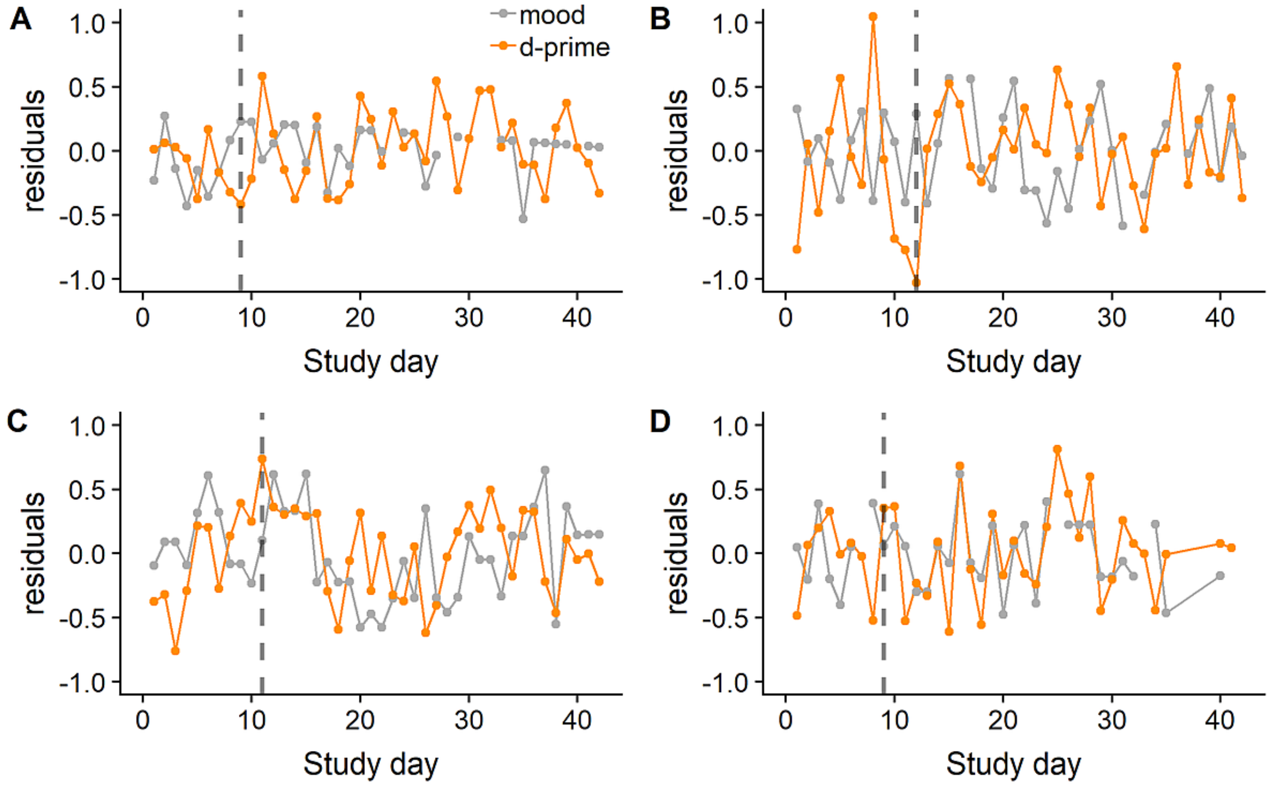
Learning slope and asymptote correlated moderately with one another (ρ=0.70, 95% CI 0.39-0.84; *P*<.001), but neither slope nor asymptote correlated with the intercept (minimum *P*=.40). Correlational analysis examining the relationship between learning parameters and mean mood and activity during the monitoring period revealed no significant association between learning slope or asymptote with mean mood (ρ=–0.07 to 0.05; minimum *P*=.74), but a significant association emerged with mean activity (slope: ρ=0.44 [0.05-0.75]; *P*=.02; asymptote: ρ=0.45 [0.08-0.68]; *P*=.02).

Covariation of Fluctuations Between Mood, Cognitive Function, and Activity

When examining the full assessment period, mood fluctuations did not significantly covary with fluctuations in daily heart rate (*t*=1.2; *P*=.23), or d-prime (*t*=–1.8; *P*=.07). However, daily mood fluctuations were seen with concurrent fluctuations in step count (estimate=–0.001, SE 0.0004; *t*=–3.93; *P*<.001), with a relatively higher levels of activity associated with relatively better mood. Examining the direction of these effects using lagged models for the entire assessment period revealed that increased step count the day before was associated with better mood on the following day (previous day activity residual: estimate=–0.001, SE 0.0004; *t*=–2.40; *P*=.02). However, a relatively more positive mood did not predict increased activity on the following day (*t*=–0.49; *P*=.62).

Fluctuations between outcome variables and mood were reexamined after excluding the period during which 90% n-back learning rates were reached. Daily fluctuations in d-prime (estimate=–0.06, SE 0.02; *t*=–2.51; *P*=.01) and activity (estimate=–0.002, SE 0.0005; *t*=–3.50; *P*<.001) were associated with fluctuations in mood (Figure 4 shows individual examples of covariation of mood with d-prime). The covariation of mood with heart rate was nonsignificant (*t*=1.74; *P*=.08). Estimates and significance levels remained similar after restricting analyses to days where all 3 n-back assessments were available (d-prime: estimate=–0.07, SE 0.03; *t*=–2.57; *P*=.01; activity: estimate=–0.002, SE 0.0005; *t*=–4.42; *P*<.001). Taking a lagged approach, but after excluding the learning period, increased step count was again associated with better mood on the next day (estimate=–0.001, SE 0.0005; *t*=–2.00; *P*=.05). However, a relatively more positive mood did not predict an increased step count on the following day (*t*=–0.34; *P*=.73). This approach also indicated that relatively higher d-prime scores were not associated with better mood on the previous or following day (t range=0.61 to –0.77; *P* range=0.44 to 0.72).

Figure 4. Covariation of daily mood residuals and mean daily d-prime residuals before and after 90% learning level reached (denoted in dashed line) for 4 participants. Note that mood residuals are inverted, so that higher residuals denote better outcomes for both mood and n-back performance.



Discussion

Principal Results and Comparison With Previous Work

This study characterizes high-frequency assessments of mood, cognition, and activity levels over a 6-week period in patients with MDD. By modelling individual learning curves from cognitive testing, we show that, after excluding an initial learning period, Cognition Kit N-Back test performance shows sensitivity to daily fluctuations in mood. The n-back test showed rapid early performance improvement followed by more incremental learning as participants neared a stable performance level, a trend that is replicated from previous findings for mean daily data in this sample [38]. This pattern is similar to that seen in laboratory-based cognitive assessments [53,54] and in practice effects identified during higher-frequency mobile cognitive assessments completed over 1 week of testing [22,55]. Participants completed 90% of task learning after a mean of 22 assessments (approximately 10 days after study onset). All participants achieved this level of improvement after 31 assessments (23 days). As such, a 10-day run-in before the introduction of any intervention may help to reduce the majority of learning effects in this n-back task, and subsequent performance can be referenced relative to an individual's learning plateau as a proxy for baseline. Careful consideration of learning effects may be particularly important in intervention studies examining temporal associations between symptoms and accurate digital phenotyping [56]. As such, after the removal of learning effects, the association between mood and n-back performance can be objectively measured with a brief 3-times daily assessment. After excluding the initial learning period from the analysis, we found that better mood was associated with higher n-back scores on measurements taken on the same day. This approach demonstrates a potential method to "baseline" individual performance in high-frequency assessments after their learning plateau is reached in order to disentangle clinically meaningful relationships between mood and cognition. The absence of a significant relationship between mood and n-back performance during the learning period could be attributable to the development of specific learning strategies applied by participants during the learning phase [57,58] that obscure the relationship between mood and cognition.

Indeed, experimental studies have shown better cognitive performance when a positive mood is induced [59-61], although findings are not always consistent (eg, Nusbaum et al [62]). In the case of the latter study [62], this may be due to an examination of cognitive flexibility as opposed to working memory, as these may have distinct underlying mechanisms contributing toward overall cognitive performance [63]. Working memory impairments in patients with depression have been widely observed and are associated with the number of hospitalizations and the overall prognosis [64,65]. Performance in this domain is associated with negative symptom severity, including lack of motivation and apathy [66], in addition to positive valence [60]. In this study, changes may represent longer-term improvements in response to existing treatment regimens or may reflect the natural history of remission and fluctuating symptom levels commonly reported in depression [67]. It has been suggested that enhanced patient understanding of mood and mood changes over time may help to improve depressive symptoms and their management [68,69]. Whether these methods themselves have an impact on mood over time requires further clarification [70].

When examining diurnal and weekday effects, participants showed modestly better cognitive test performance early in the morning and late at night, with a slight decrease in function throughout the middle parts of the day. These findings are in agreement with the reported "afternoon slump" in cognition in healthy adults [71]. Neither diurnal nor day-of-week effects on mood were seen, in contrast to previous reports [31,34]. However, in broad agreement with previous research [35,36], we identified day-of-week effects for activity levels as measured by step count and similar patterns in mean heart rate. Step counts showed a quadratic effect of day of week, with lower counts registered at the beginning and end of each week. These findings highlight the importance of carefully considering the timing of assessments to ensure consistency in data sampling and control trends in data over time. Longer-term trends in mood data independent of any treatment effects highlight the importance of adequately controlling for subtle overall trends in time in future interventional studies, ideally using a randomized design and a placebo-controlled comparison group. Using a multilevel modeling approach, we were able to examine whether mood fluctuated synchronously with other measures of interest over time. We identified covariations in mood and activity levels, in keeping with a body of research showing improvements in depression symptoms with exercise [72,73], or simple walking interventions [74]. Our results show that a relatively increased mood was associated with a relatively increased step count on the previous day (but not the following day), which indicated that the beneficial effects of exercise on well-being may well be protracted [75].

Limitations

While associations between activity levels (step count), mood, and cognition were identified, step counts are likely to have been influenced by wearing patterns [38], thereby reducing the reliability of the data. The summarized daily step counts also fail to elucidate whether activity was acute or prolonged, vigorous, or light, and the exact timing of activity changes in relation to mood and cognitive assessments. Accelerometers can help to continuously monitor activity, and concurrent GPS and travel diary information has been found to help to classify and identify the duration of walking [76]. Previous studies have examined associations between physical activity on mood in assessments triggered through GPS distance tracking [77,78], a method that could be used to refine the timing and accuracy of activity assessments in this context. Additionally, since this study focused on patients with mild-to-moderate MDD who volunteered for participation, it is unclear whether the results would generalize to patients with different clinical severity.

Conclusions

This study indicates the importance of incorporating objective measures of cognitive testing and provides insight into fluctuations in mood and cognition in patients with MDD. The feasibility of remote high-frequency testing in MDD is promising for future research in this field and has important implications for clinical interventions. While these methods can be used to monitor nuanced fluctuations between mood and cognition in a real-life setting, they may also be useful as a treatment tool. Understanding objective cognitive function in depression may also be used to target patients' acceptance of their objective difficulties in cognition and may be particularly relevant for interventions such as CBT and mindfulness [79]. Overall, while assessments of activity need to be further refined and improved, the effects of step count on mood and cognition support the concurrent capture of data on activity, which may be an important contributor to variations in mood and changes in cognition in everyday life in patients with depression.

Conflicts of Interest

SS is an employee of Takeda Pharmaceuticals, the sponsor of the research. LC was an employee of Takeda Pharmaceuticals USA, Inc, at the time of study. JK and FC are employees at Cambridge Cognition. The other authors have no conflicts to declare.

Multimedia Appendix 1

Final model selection and model parameters for daily assessment outcomes over 6-week period.

[DOCX File , 17 KB - mental_v11i1e46895_app1.docx]

References

1. Ferrari AJ, Charlson FJ, Norman RE, Patten SB, Freedman G, Murray CJL, et al. Burden of depressive disorders by country, sex, age, and year: findings from the Global Burden of Disease Study 2010. PLoS Med 2013;10(11):e1001547 [FREE Full text] [doi: 10.1371/journal.pmed.1001547] [Medline: 24223526]
2. Moussavi S, Chatterji S, Verdes E, Tandon A, Patel V, Ustun B. Depression, chronic diseases, and decrements in health: results from the World Health Surveys. Lancet 2007;370(9590):851-858. [doi: 10.1016/S0140-6736(07)61415-9] [Medline: 17826170]
3. Diagnostic and Statistical Manual of Mental Disorders, 5th Edition. Arlington, V: American Psychiatric Publishing; 2013.
4. Hammar A, Ardal G. Cognitive functioning in major depression—a summary. Front Hum Neurosci 2009;3:26 [FREE Full text] [doi: 10.3389/neuro.09.026.2009] [Medline: 19826496]

5. McClintock SM, Husain MM, Greer TL, Cullum CM. Association between depression severity and neurocognitive function in major depressive disorder: a review and synthesis. Neuropsychology 2010;24(1):9-34. [doi: 10.1037/a0017336] [Medline: 20063944]
6. Ahern E, Semkovska M. Cognitive functioning in the first-episode of major depressive disorder: a systematic review and meta-analysis. Neuropsychology 2017;31(1):52-72. [doi: 10.1037/neu0000319] [Medline: 27732039]
7. Kaser M, Zaman R, Sahakian BJ. Cognition as a treatment target in depression. Psychol Med 2017;47(6):987-989 [FREE Full text] [doi: 10.1017/S0033291716003123] [Medline: 27938430]
8. Robinson MD, Clore GL. Belief and feeling: evidence for an accessibility model of emotional self-report. Psychol Bull 2002;128(6):934-960. [doi: 10.1037/0033-2909.128.6.934] [Medline: 12405138]
9. Ebner-Priemer UW, Trull TJ. Ambulatory assessment: an innovative and promising approach for clinical psychology. European Psychologist 2009;14(2):109-119. [doi: 10.1027/1016-9040.14.2.109]
10. Dillon DG, Pizzagalli DA. Mechanisms of memory disruption in depression. Trends Neurosci 2018;41(3):137-149 [FREE Full text] [doi: 10.1016/j.tins.2017.12.006] [Medline: 29331265]
11. Ben-Zeev D, Young MA. Accuracy of hospitalized depressed patients' and healthy controls' retrospective symptom reports: an experience sampling study. J Nerv Ment Dis 2010;198(4):280-285. [doi: 10.1097/NMD.0b013e3181d6141f] [Medline: 20386257]
12. Ben-Zeev D, Young MA, Madsen JW. Retrospective recall of affect in clinically depressed individuals and controls. Cogn Emot 2009;23(5):1021-1040. [doi: 10.1080/02699930802607937]
13. Everaert J, Vrijsen JN, Martin-Willett R, van de Kraats L, Joormann J. A meta-analytic review of the relationship between explicit memory bias and depression: depression features an explicit memory bias that persists beyond a depressive episode. Psychological Bulletin 2022;148(5-6):435-463. [doi: 10.1037/bul0000367]
14. Semkovska M, Quinlivan L, O'Grady T, Johnson R, Collins A, O'Connor J, et al. Cognitive function following a major depressive episode: a systematic review and meta-analysis. Lancet Psychiatry 2019;6(10):851-861. [doi: 10.1016/S2215-0366(19)30291-3] [Medline: 31422920]
15. Bos FM, Schoevers RA, aan het Rot M. Experience sampling and ecological momentary assessment studies in psychopharmacology: a systematic review. Eur Neuropsychopharmacol 2015;25(11):1853-1864. [doi: 10.1016/j.euroneuro.2015.08.008] [Medline: 26336868]
16. Popkin SJ, Gallagher D, Thompson LW, Moore M. Memory complaint and performance in normal and depressed older adults. Exp Aging Res 1982;8(3-4):141-145. [doi: 10.1080/03610738208260271] [Medline: 7169071]
17. Antikainen R, Hänninen T, Honkalampi K, Hintikka J, Koivumaa-Honkanen H, Tanskanen A, et al. Mood improvement reduces memory complaints in depressed patients. Eur Arch Psychiatry Clin Neurosci 2001;251(1):6-11. [doi: 10.1007/s004060170060] [Medline: 11315519]
18. Naismith SL, Longley WA, Scott EM, Hickie IB. Disability in major depression related to self-rated and objectively-measured cognitive deficits: a preliminary study. BMC Psychiatry 2007;7:32 [FREE Full text] [doi: 10.1186/1471-244X-7-32] [Medline: 17634111]
19. Zlatar ZZ, Muniz M, Galasko D, Salmon DP. Subjective cognitive decline correlates with depression symptoms and not with concurrent objective cognition in a clinic-based sample of older adults. J Gerontol B Psychol Sci Soc Sci 2018;73(7):1198-1202 [FREE Full text] [doi: 10.1093/geronb/gbw207] [Medline: 28329816]
20. Smith AP, Clark R, Gallagher J. Breakfast cereal and caffeinated coffee: effects on working memory, attention, mood, and cardiovascular function. Physiol Behav 1999;67(1):9-17. [doi: 10.1016/s0031-9384(99)00025-6] [Medline: 10463623]
21. Wright KP, Lowry CA, Lebourgeois MK. Circadian and wakefulness-sleep modulation of cognition in humans. Front Mol Neurosci 2012;5:50 [FREE Full text] [doi: 10.3389/fnmol.2012.00050] [Medline: 22529774]
22. Schweitzer P, Husky M, Allard M, Amieva H, Pérès K, Foubert-Samier A, et al. Feasibility and validity of mobile cognitive testing in the investigation of age-related cognitive decline. Int J Methods Psychiatr Res 2017;26(3):e1521 [FREE Full text] [doi: 10.1002/mpr.1521] [Medline: 27539327]
23. Moore RC, Depp CA, Wetherell JL, Lenze EJ. Ecological momentary assessment versus standard assessment instruments for measuring mindfulness, depressed mood, and anxiety among older adults. J Psychiatr Res 2016;75:116-123 [FREE Full text] [doi: 10.1016/j.jpsychires.2016.01.011] [Medline: 26851494]
24. Barge-Schaapveld DQ, Nicolson NA, van der Hoop RG, De Vries MW. Changes in daily life experience associated with clinical improvement in depression. J Affect Disord 1995;34(2):139-154. [doi: 10.1016/0165-0327(95)00012-c] [Medline: 7665806]
25. Lenderking WR, Hu M, Tennen H, Cappelleri JC, Petrie CD, Rush AJ. Daily process methodology for measuring earlier antidepressant response. Contemp Clin Trials 2008;29(6):867-877. [doi: 10.1016/j.cct.2008.05.012] [Medline: 18606249]
26. van de Leemput IA, Wichers M, Cramer AOJ, Borsboom D, Tuerlinckx F, Kuppens P, et al. Critical slowing down as early warning for the onset and termination of depression. Proc Natl Acad Sci U S A 2014;111(1):87-92 [FREE Full text] [doi: 10.1073/pnas.1312114110] [Medline: 24324144]
27. Trull TJ, Ebner-Priemer U. Ambulatory assessment. Annu Rev Clin Psychol 2013;9:151-176 [FREE Full text] [doi: 10.1146/annurev-clinpsy-050212-185510] [Medline: 23157450]

28. Marzano L, Bardill A, Fields B, Herd K, Veale D, Grey N, et al. The application of mHealth to mental health: opportunities and challenges. Lancet Psychiatry 2015;2(10):942-948. [doi: 10.1016/S2215-0366(15)00268-0] [Medline: 26462228]
29. Sliwinski MJ, Mogle JA, Hyun J, Munoz E, Smyth JM, Lipton RB. Reliability and validity of ambulatory cognitive assessments. Assessment 2018;25(1):14-30 [FREE Full text] [doi: 10.1177/1073191116643164] [Medline: 27084835]
30. Leibenluft E, Noonan BM, Wehr TA. Diurnal variation: reliability of measurement and relationship to typical and atypical symptoms of depression. J Affect Disord 1992;26(3):199-204. [doi: 10.1016/0165-0327(92)90016-y] [Medline: 1460170]
31. Peeters F, Berkhof J, Delespaul P, Rottenberg J, Nicolson NA. Diurnal mood variation in major depressive disorder. Emotion 2006;6(3):383-391. [doi: 10.1037/1528-3542.6.3.383] [Medline: 16938080]
32. Crowe E, Daly M, Delaney L, Carroll S, Malone KM. The intra-day dynamics of affect, self-esteem, tiredness, and suicidality in major depression. Psychiatry Res 2019;279:98-108. [doi: 10.1016/j.psychres.2018.02.032] [Medline: 29661498]
33. Porterfield T, Cook M, Deary IJ, Ebmeier KP. Neuropsychological function and diurnal variation in depression. J Clin Exp Neuropsychol 1997;19(6):906-913. [doi: 10.1080/01688639708403771] [Medline: 9524885]
34. Stone AA, Schneider S, Harter JK. Day-of-week mood patterns in the United States: on the existence of 'Blue Monday', 'Thank God it's Friday' and weekend effects. J Posit Psychol 2012;7(4):306-314. [doi: 10.1080/17439760.2012.691980]
35. Hirvensalo M, Telama R, Schmidt MD, Tammelin TH, Yang X, Magnussen CG, et al. Daily steps among Finnish adults: variation by age, sex, and socioeconomic position. Scand J Public Health 2011;39(7):669-677. [doi: 10.1177/1403494811420324] [Medline: 21893605]
36. Ludwig VM, Bayley A, Cook DG, Stahl D, Treasure JL, Asthworth M, et al. Association between depressive symptoms and objectively measured daily step count in individuals at high risk of cardiovascular disease in South London, UK: a cross-sectional study. BMJ Open 2018;8(4):e020942 [FREE Full text] [doi: 10.1136/bmjopen-2017-020942] [Medline: 29654044]
37. Colombo D, Fernández-Álvarez J, Patané A, Semonella M, Kwiatkowska M, García-Palacios A, et al. Current state and future directions of technology-based ecological momentary assessment and intervention for major depressive disorder: a systematic review. J Clin Med 2019;8(4):465 [FREE Full text] [doi: 10.3390/jcm8040465] [Medline: 30959828]
38. Cormack F, McCue M, Taptiklis N, Skirrow C, Glazer E, Panagopoulos E, et al. Wearable technology for high-frequency cognitive and mood assessment in major depressive disorder: longitudinal observational study. JMIR Ment Health 2019;6(11):e12814 [FREE Full text] [doi: 10.2196/12814] [Medline: 31738172]
39. Snyder HR. Major depressive disorder is associated with broad impairments on neuropsychological measures of executive function: a meta-analysis and review. Psychol Bull 2013;139(1):81-132 [FREE Full text] [doi: 10.1037/a0028727] [Medline: 22642228]
40. Espeland MA, Katula JA, Rushing J, Kramer AF, Jennings JM, Sink KM, et al. Performance of a computer-based assessment of cognitive function measures in two cohorts of seniors. Int J Geriatr Psychiatry 2013;28(12):1239-1250 [FREE Full text] [doi: 10.1002/gps.3949] [Medline: 23589390]
41. Cormack F, Ticcinelli V, Taptiklis N, Kudelka J, Emmert K, Maetzler W, et al. App-based cognitive assessment and monitoring—a feasibility study in patients with immune-mediated inflammatory and neurodegenerative disorders. Neurosci Appl. 2022. URL: https://isctm.org/public_access/Autumn2022/Abstracts/Cormack_Abstract.pdf [accessed 2024-02-29]
42. Levis B, Sun Y, He C, Wu Y, Krishnan A, Bhandari PM, et al. Accuracy of the PHQ-2 alone and in combination with the PHQ-9 for screening to detect major depression: systematic review and meta-analysis. JAMA 2020;323(22):2290-2300 [FREE Full text] [doi: 10.1001/jama.2020.6504] [Medline: 32515813]
43. Schiweck C, Piette D, Berckmans D, Claes S, Vrieze E. Heart rate and high frequency heart rate variability during stress as biomarker for clinical depression. A systematic review. Psychol Med 2019;49(2):200-211. [doi: 10.1017/S0033291718001988] [Medline: 30134999]
44. Ramsey CM, Lynch KG, Gehrman PR, Vairavan S, Narayan VA, Li QS, et al. Daily steps and depressive symptoms: a longitudinal evaluation of patients with major depressive disorder in the Precision Medicine in Mental Health Care study. J Affect Disord 2022;300:334-340. [doi: 10.1016/j.jad.2021.12.116] [Medline: 34979178]
45. Veerabhadrappa P, Moran MD, Renninger MD, Rhudy MB, Dreisbach SB, Gift KM. Tracking steps on Apple Watch at different walking speeds. J Gen Intern Med 2018;33(6):795-796 [FREE Full text] [doi: 10.1007/s11606-018-4332-y] [Medline: 29633143]
46. Fuller D, Colwell E, Low J, Orychock K, Tobin MA, Simango B, et al. Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review. JMIR Mhealth Uhealth 2020;8(9):e18694 [FREE Full text] [doi: 10.2196/18694] [Medline: 32897239]
47. Brose A, Schmiedek F, Lövdén M, Lindenberger U. Daily variability in working memory is coupled with negative affect: the role of attention and motivation. Emotion 2012;12(3):605-617. [doi: 10.1037/a0024436] [Medline: 21787075]
48. Richardson CR, Buis LR, Janney AW, Goodrich DE, Sen A, Hess ML, et al. An online community improves adherence in an internet-mediated walking program. Part 1: results of a randomized controlled trial. J Med Internet Res 2010;12(4):e71 [FREE Full text] [doi: 10.2196/jmir.1338] [Medline: 21169160]
49. Moy ML, Weston NA, Wilson EJ, Hess ML, Richardson CR. A pilot study of an internet walking program and pedometer in COPD. Respir Med 2012;106(9):1342-1350 [FREE Full text] [doi: 10.1016/j.rmed.2012.06.013] [Medline: 22795984]

50. Bringmann LF, Vissers N, Wichers M, Geschwind N, Kuppens P, Peeters F, et al. A network approach to psychopathology: new insights into clinical longitudinal data. PLoS One 2013;8(4):e60188 [FREE Full text] [doi: 10.1371/journal.pone.0060188] [Medline: 23593171]
51. Verhagen SJW, Hasmi L, Drukker M, van Os J, Delespaul PAEG. Use of the experience sampling method in the context of clinical trials. Evid Based Ment Health 2016;19(3):86-89 [FREE Full text] [doi: 10.1136/ebmental-2016-102418] [Medline: 27443678]
52. Feldman LS, Cao J, Andalib A, Fraser S, Fried GM. A method to characterize the learning curve for performance of a fundamental laparoscopic simulator task: defining "learning plateau" and "learning rate". Surgery 2009;146(2):381-386. [doi: 10.1016/j.surg.2009.02.021] [Medline: 19628099]
53. Yamashita M, Kawato M, Imamizu H. Predicting learning plateau of working memory from whole-brain intrinsic network connectivity patterns. Sci Rep 2015;5:7622 [FREE Full text] [doi: 10.1038/srep07622] [Medline: 25557398]
54. Collie A, Maruff P, Darby DG, McStephen M. The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test-retest intervals. J Int Neuropsychol Soc 2003;9(3):419-428. [doi: 10.1017/S1355617703930074] [Medline: 12666766]
55. Dupuy M, Misdrahi D, N'Kaoua B, Tessier A, Bouvard A, Schweitzer P, et al. Mobile cognitive testing in patients with schizophrenia: a controlled study of feasibility and validity. J de Ther Comport et Cogn 2018;28(4):204-213. [doi: 10.1016/j.jtcc.2018.02.002]
56. Griffin B, Saunders KEA. Smartphones and wearables as a method for understanding symptom mechanisms. Front Psychiatry 2019;10:949 [FREE Full text] [doi: 10.3389/fpsyt.2019.00949] [Medline: 32009990]
57. Gathercole SE, Dunning DL, Holmes J, Norris D. Working memory training involves learning new skills. J Mem Lang 2016;105:19-42 [FREE Full text] [doi: 10.1016/j.jml.2018.10.003] [Medline: 31235992]
58. Tenison C, Anderson JR. Modeling the distinct phases of skill acquisition. J Exp Psychol Learn Mem Cogn 2016;42(5):749-767. [doi: 10.1037/xlm0000204] [Medline: 26551626]
59. Nadler RT, Rabi R, Minda JP. Better mood and better performance. Learning rule-described categories is enhanced by positive mood. Psychol Sci 2010;21(12):1770-1776. [doi: 10.1177/0956797610387441] [Medline: 20974709]
60. Storbeck J, Maswood R. Happiness increases verbal and spatial working memory capacity where sadness does not: emotion, working memory and executive control. Cogn Emot 2016;30(5):925-938. [doi: 10.1080/02699931.2015.1034091] [Medline: 25947579]
61. Yang H, Yang S, Isen AM. Positive affect improves working memory: implications for controlled cognitive processing. Cogn Emot 2013;27(3):474-482. [doi: 10.1080/02699931.2012.713325] [Medline: 22917664]
62. Nusbaum AT, Wilson CG, Stenson A, Hinson JM, Whitney P. Induced positive mood and cognitive flexibility: evidence from task switching and reversal learning. Collabra Psychol 2018;4(1):25 [FREE Full text] [doi: 10.1525/collabra.150]
63. Periáñez JA, Lubrini G, García-Gutiérrez A, Ríos-Lago M. Construct validity of the Stroop color-word test: influence of speed of visual search, verbal fluency, working memory, cognitive flexibility, and conflict monitoring. Arch Clin Neuropsychol 2021;36(1):99-111 [FREE Full text] [doi: 10.1093/arclin/acaa034] [Medline: 32514527]
64. Harvey PO, Le Bastard G, Pochon JB, Levy R, Allilaire JF, Dubois B, et al. Executive functions and updating of the contents of working memory in unipolar depression. J Psychiatr Res 2004;38(6):567-576. [doi: 10.1016/j.jpsychires.2004.03.003] [Medline: 15458852]
65. Nikolin S, Tan YY, Schwaab A, Moffa A, Loo CK, Martin D. An investigation of working memory deficits in depression using the n-back task: a systematic review and meta-analysis. J Affect Disord 2021;284:1-8. [doi: 10.1016/j.jad.2021.01.084] [Medline: 33581489]
66. Zhu Y, Womer FY, Leng H, Chang M, Yin Z, Wei Y, et al. The relationship between cognitive dysfunction and symptom dimensions across schizophrenia, bipolar disorder, and major depressive disorder. Front Psychiatry 2019;10:253 [FREE Full text] [doi: 10.3389/fpsyt.2019.00253] [Medline: 31105603]
67. Andrews G. Placebo response in depression: bane of research, boon to therapy. Br J Psychiatry 2001;178:192-194 [FREE Full text] [doi: 10.1192/bjp.178.3.192] [Medline: 11230026]
68. Kauer SD, Reid SC, Crooke AHD, Khor A, Hearps SJC, Jorm AF, et al. Self-monitoring using mobile phones in the early stages of adolescent depression: randomized controlled trial. J Med Internet Res 2012;14(3):e67 [FREE Full text] [doi: 10.2196/jmir.1858] [Medline: 22732135]
69. Groot PC. Patients can diagnose too: how continuous self-assessment aids diagnosis of, and recovery from, depression. J Ment Health 2010;19(4):352-362. [doi: 10.3109/09638237.2010.494188] [Medline: 20636115]
70. Hartmann JA, Wichers M, Menne-Lothmann C, Kramer I, Viechtbauer W, Peeters F, et al. Experience sampling-based personalized feedback and positive affect: a randomized controlled trial in depressed patients. PLoS One 2015;10(6):e0128095 [FREE Full text] [doi: 10.1371/journal.pone.0128095] [Medline: 26034983]
71. Wertz AT, Ronda JM, Czeisler CA, Wright KP. Effects of sleep inertia on cognition. JAMA 2006;295(2):163-164. [doi: 10.1001/jama.295.2.163] [Medline: 16403927]
72. Schuch FB, Vancampfort D, Richards J, Rosenbaum S, Ward PB, Stubbs B. Exercise as a treatment for depression: a meta-analysis adjusting for publication bias. J Psychiatr Res 2016;77:42-51 [FREE Full text] [doi: 10.1016/j.jpsychires.2016.02.023] [Medline: 26978184]

73. Schuch FB, Deslandes AC, Stubbs B, Gosmann NP, da Silva CTB, de Almeida Fleck MP. Neurobiological effects of exercise on major depressive disorder: a systematic review. Neurosci Biobehav Rev 2016;61:1-11. [doi: 10.1016/j.neubiorev.2015.11.012] [Medline: 26657969]
74. Robertson R, Robertson A, Jepson R, Maxwell M. Walking for depression or depressive symptoms: a systematic review and meta-analysis. Ment Health Phys Act 2012;5(1):66-75. [doi: 10.1016/j.mhpa.2012.03.002]
75. Daley AJ, Welch A. The effects of 15 min and 30 min of exercise on affective responses both during and after exercise. J Sports Sci 2004;22(7):621-628. [doi: 10.1080/02640410310001655778] [Medline: 15370492]
76. Kang B, Moudon AV, Hurvitz PM, Reichley L, Saelens BE. Walking objectively measured: classifying accelerometer data with GPS and travel diaries. Med Sci Sports Exerc 2013;45(7):1419-1428 [FREE Full text] [doi: 10.1249/MSS.0b013e318285f202] [Medline: 23439414]
77. Reichert M, Tost H, Reinhard I, Schlotz W, Zipf A, Salize HJ, et al. Exercise versus nonexercise activity: e-diaries unravel distinct effects on mood. Med Sci Sports Exerc 2017;49(4):763-773 [FREE Full text] [doi: 10.1249/MSS.0000000000001149] [Medline: 27824691]
78. Reichert M, Tost H, Reinhard I, Zipf A, Salize HJ, Meyer-Lindenberg A, et al. Within-subject associations between mood dimensions and non-exercise activity: an ambulatory assessment approach using repeated real-time and objective data. Front Psychol 2016;7:918 [FREE Full text] [doi: 10.3389/fpsyg.2016.00918] [Medline: 27445891]
79. Serra-Blasco M, Torres IJ, Vicent-Gil M, Goldberg X, Navarra-Ventura G, Aguilar E, et al. Discrepancy between objective and subjective cognition in major depressive disorder. Eur Neuropsychopharmacol 2019;29(1):46-56. [doi: 10.1016/j.euroneuro.2018.11.1104] [Medline: 30503099]

Abbreviations

AIC: Akaike information criterion
BIC: Bayesian information criterion
MDD: major depressive disorder
PHQ: Patient Health Questionnaire

Edited by J Torous; submitted 02.03.23; peer-reviewed by H Kim, M Tai-Seale; comments to author 09.06.23; revised version received 28.11.23; accepted 23.12.23; published 31.05.24.

Please cite as:
Cormack F, McCue M, Skirrow C, Cashdollar N, Taptiklis N, van Schaik T, Fehnert B, King J, Chrones L, Sarkey S, Kroll J, Barnett JH
Characterizing Longitudinal Patterns in Cognition, Mood, And Activity in Depression With 6-Week High-Frequency Wearable Assessment: Observational Study
JMIR Ment Health 2024;11:e46895
URL: https://mental.jmir.org/2024/1/e46895
doi:10.2196/46895
PMID:38819909

Original Paper

News Media Framing of Suicide Circumstances and Gender: Mixed Methods Analysis

Jasmine C Foriest[1,2], MCHES, MS, MPH; Shravika Mittal[1], MS; Eugenia Kim[1], MS; Andrea Carmichael[2], MPH; Natalie Lennon[2,3], MPH; Steven A Sumner[2], MSc, MD; Munmun De Choudhury[1], PhD

[1]School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, United States

[2]National Center for Injury Prevention and Control, Centers for Disease Control and Prevention, Atlanta, GA, United States

[3]Accenture, Arlington, VA, United States

Corresponding Author:
Jasmine C Foriest, MCHES, MS, MPH
School of Interactive Computing
Georgia Institute of Technology
225 North Ave NW
Atlanta, GA, 30332
United States
Phone: 1 7064643658
Email: jay@gatech.edu

Abstract

Background: Suicide is a leading cause of death worldwide. Journalistic reporting guidelines were created to curb the impact of unsafe reporting; however, how suicide is framed in news reports may differ by important characteristics such as the circumstances and the decedent's gender.

Objective: This study aimed to examine the degree to which news media reports of suicides are framed using stigmatized or glorified language and differences in such framing by gender and circumstance of suicide.

Methods: We analyzed 200 news articles regarding suicides and applied the validated Stigma of Suicide Scale to identify stigmatized and glorified language. We assessed linguistic similarity with 2 widely used metrics, cosine similarity and mutual information scores, using a machine learning–based large language model.

Results: News reports of male suicides were framed more similarly to stigmatizing ($P<.001$) and glorifying ($P=.005$) language than reports of female suicides. Considering the circumstances of suicide, mutual information scores indicated that differences in the use of stigmatizing or glorifying language by gender were most pronounced for articles attributing legal (0.155), relationship (0.268), or mental health problems (0.251) as the cause.

Conclusions: Linguistic differences, by gender, in stigmatizing or glorifying language when reporting suicide may exacerbate suicide disparities.

(JMIR Ment Health 2024;11:e49879) doi:10.2196/49879

KEYWORDS

suicide; framing; disparities; reporting guidelines; gender; stigma; glorification; glorify; glorifying; suicidal; self harm; suicides; stigmatizing; stigmatization; disparities; reporting; news; journalist; journalists; journalism; machine learning; NLP; natural language processing; LLM; LLMs; language model; language models; linguistic; linguistics; reporter; reporters; gender; digital mental health; mHealth; media

Introduction

Background

Suicide is a leading cause of death in the United States, claiming over 48,000 lives in 2021 [1]. Rates of suicide and self-harm are significantly different for males and females. In 2021, US age-adjusted suicide rates were higher among males (22.8/100,000) than females (5.7/100,000) [1]. However, reported rates of nonfatal self-harm in 2020, the most recent year for which such data are available, were higher among females (189.4/100,000) than males (123.8/100,000) [2].

Gender and Suicide

The observed gender disparities in suicide death rates between males and females are echoed in gender differences in other suicide-related behaviors such as ideation, attempts, and mechanisms [3]. Women have higher rates of lifetime suicidal ideation [4] and suicide attempts [5]. Wang et al [3] also report that men experience higher rates of suicide likely due to the increased lethality of the mechanism. Circumstances of suicide, or negative life events that precipitate suicide-related behavior [6], also vary among men and women according to the analysis of suicides in the National Violent Death Reporting System [6]. For example, previous analysis correlates negative financial and employment circumstances with increased male suicides in 2009 [7]. Suicide-related behavior is therefore highly gendered, and research that seeks to inform understanding and prevention of suicide must consider gender as a layer of analysis. While we acknowledge the call for expanded US research on gender differences across all suicide-related behaviors [8], this study includes gender as an essential layer of analysis to understand reports of suicide deaths.

Unsafe Media Reporting about Suicide

Unsafe news reporting is among many risk factors that contribute to suicide [9]. For example, studies have shown that sensationalized reporting, including that of celebrity suicides, is positively correlated with a rise in suicides [10,11]. Other research has shown that specific characteristics of news coverage, for example, descriptions of an epidemic or suicide myths, may also shape population-level effects on suicide [12]. However, the need to better understand how specific language in framing suicide news shapes its perception is ongoing. In particular, one critical research need is to better elucidate the differences in unsafe reporting by gender when considering the differences in suicide-related behavior. While discussions of all suicide-related behavior are likely subject to unsafe reporting, this study aligns with the focus of existing work examining unsafe media portrayals of individual suicide deaths in particular.

Stigma and Glorification in News Reports

The suicide reporting guidelines detail best practices for news organizations to safely report suicides and generally focus on content that should not be included, such as the mechanism or specific location of suicide [13]. However, assessing subtle nuances about how suicide may be discussed or framed in other harmful ways is more challenging. In this study, we explore two particular ways of framing suicide in news reports, which are stigmatizing and glorifying descriptions.

Although elucidating the precise language that stigmatizes or glorifies suicide is challenging, researchers have developed the validated Stigma of Suicide Scale (SOSS) [14], which helps to linguistically define these constructs as detailed below. The SOSS furthers research in this area by measuring specific words and how they may stigmatize or glorify perspectives. The scale reports a list of words that can be used to further understand, identify, and explore such framings. For example, “shallow,” “pathetic,” and “immoral” are associated with stigmatizing perspectives (ie, those perspectives that ascribe a negative attitude or perception toward those who die by suicide). Similarly, “understandable,” “brave,” and “motivated” are associated with glorifying perspectives (ie, those perspectives that seek to normalize suicide) [14].

Identifying Problematic Framing

Carefully framing suicide deaths in the news media is essential as the language people use shapes perceived reality [15] and can influence suicide-related behavior [12]. Previous articles on suicide reporting have mostly focused on assessing the degree of explicit adherence to elements of the suicide-safe reporting guidelines [16]. While recent articles have applied machine learning and natural language processing to suicide news, they have largely focused on using these techniques to automate identifying more structured elements present in the suicide reporting guidelines [13]. Thus, in this research, we aim to develop natural language processing methods to measure and explore the more nuanced problematic framings of suicide using stigmatizing or glorifying language. Furthermore, we explore such framings by gender and by the circumstance of suicide. We pose two main research questions: (1) how does the overall framing of suicide differ by gender identity? and (2) how does such framing differ by gender when considering the specific circumstance of suicide?

Methods

Data Collection

Data for this study were comprised of news articles from All the News [17], a leading benchmark data set used in the field of computer science and natural language processing–based news research; the public availability of the data set allows for independent validation of results by the academic community and facilitates transparent head-to-head comparison of different natural language processing approaches. The data set includes 143,000 publicly accessible articles published by 15 print and digital American publications, from 2015 to 2017. Articles included in this data set are in English and come from publishers across political alignments. Examples of publishers include the New York Times, CNN, Washington Post, Breitbart, Fox News, Buzzfeed News, and others [17]. We identified articles from this data set relevant to suicide using suicide-related keywords previously curated by public health experts [18]. Previous research has used the All the News data set for identifying political polarization [19], differences between human- and machine-generated news [20], and news recommendations [21]. We randomly sampled 800 suicide-related articles, in English, and excluded 600 articles mentioning homicide-suicide, suicide bombings, suicide attempts, suicide euphemisms, and fictional portrayals of suicide [22], such as those illustrated in fictional books, television, and movies, as these classifications in media are researched differently in suicidology. In total, 200 articles met the inclusion criteria of mentioning a real-life suicide death not involving the homicide of another person and constitute the data set in this study.

Qualitative Methods

We used 2 coders for a deductive approach to annotate article mentions of gender and circumstances of the suicide to answer our research questions. We derived our definitions of circumstances from Chancellor et al's [23] validated annotation scheme of risk factors for suicide. Chancellor et al [23] developed the scheme using social media data of 200 Reddit community posts from r/SuicideWatch with a focus on construct validity. This scheme offers a formal technique to identify clinical suicide–related information in lay contexts.

Quantitative Methods

Framing—Stigma of Suicide Scale

As noted above, we derived the frames of stigmatization and glorification from the SOSS study [14]. This validated scale consists of linguistic descriptors of suicide. To derive the scale, Batterham et al [14] used principal component analysis of survey results from the public, rating 80 one-word descriptors of someone who dies by suicide to produce a list of words associated with stigmatization and glorification of suicide. The SOSS includes a third frame of isolation or depression (ie, "unhappy," "depressed," and "sad"), which we briefly report in this study but do not focus on the subsequent analyses as we found no discernable differences between genders, perhaps because depressed mood is a common symptom preceding suicide [24], and such language can be applied generically.

Linguistic Representation of the Stigma of Suicide Scale

To get a comprehensive linguistic representation of stigmatizing and glorifying language as represented by the SOSS, we used natural language modeling techniques instead of relying on lexicon-based approaches that restrict themselves to finding exact matches of a fixed set of keywords. We adopted language modeling techniques to better capture linguistic context and nuanced writing style, considering semantic and syntactic relationships between words and phrases in sentences [25,26]. Existing research has shown that such context-based approaches outperform lexicon-based ones on natural language understanding tasks [27,28]. We used the keywords identified by Batterham et al [14] for stigmatization and glorification to generate vector representations for each frame. Vector representations are large numerical representations of text used in natural language processing, and the numerical representations help identify other words that tend to co-occur or lie in the vicinity of particular words of interest. We used Bidirectional Encoder Representations from Transformers (BERT; Google) [29] along with their relevant extracted synonyms using WordNet (Princeton University) [30], a large English database containing synonymous groups (synsets) of nouns, verbs, adjectives, and adverbs. BERT is a leading large language model that is pretrained on billions of documents from the internet and encodes text to generate word-embedding representations for all the words or phrases associated with a SOSS dimension. Adding representations of synonyms strengthens the semantic representation of SOSS dimensions. After this addition, we manually examined the dictionaries to remove irrelevant words captured by WordNet. We calculated Cohen κ to assess interrater reliability (IRR) for each SOSS dimension. Finally, we averaged the relevant individual word embeddings to generate the final vector representation of each SOSS dimension.

Representation of News Articles

We generated a unique vector representation for each news article in our data set to answer research question 1 and to analyze how suicide is framed using the SOSS dimensions. We again used BERT, programmed in Python (Python Software Foundation) using Hugging Face transformers [31], to extract a sentence-level embedding representation for each news article. This allowed us to capture nuanced attributes such as context, framing and writing style, and the relationship between words in a sentence. Note that we did not aggregate the individual embeddings of all the words present in a news article to generate its linguistic representation. Sentence-level BERT embeddings ensured that we captured relevant dependency between words used in the news article. As a result, each news article was characterized using a vector representation to compare it against the SOSS dimension embeddings.

Suicide Framing Based on the Axis of Gender

We analyzed how suicide is framed in the news based on the victim's identified gender by comparing the embedding representation of each news article with the 3 vectors of the 3 SOSS dimensions (stigmatization, glorification or normalization, and isolation or depression). We adopted the cosine similarity scoring framework, as used in previous work [32], to score each article in our data set against the SOSS dimensions. These cosine similarity scores range from 1 to 1, where a higher positive score represents a stronger alignment between a news article and the corresponding SOSS dimension.

Further, we divided all the articles into groups based on the annotated gender identity of the victim. Toward inclusive representation of gender identity in our analysis, we annotated for cisgender, where one's gender identity aligns with their assigned sex, and transgender, where one's gender identity does not align with their assigned sex [33]. This resulted in the ultimate formation of 2 groups corresponding to the most represented gender identities in our data set, which are female and male. Note that some news articles made references to multiple victims. To handle such cases, articles mentioning same-gender identity victims were treated as reporting a single female or male death, and those with mixed-gender mentions were processed by manually extracting relevant text for each victim, respectively, to inform the analysis.

Ethical Considerations

The study did not meet the criteria for institutional research board review as it did not involve interactions with human or other living subjects, private or personally identifiable data, or any pharmaceuticals or medical devices [34]. The data set is comprised only of public, open-source news articles. While this data set is public, widely used, and institutional research board exempt, all possible care was taken to ensure the results of this study were communicated sensitively and safely given the context of suicide.

Results

Overview

A total of 221 real-life suicide deaths were reported from the 200 articles. We identified almost 70% (68.4%, 153/221) of these deaths to be cisgender male, 28.4% (64/221) cisgender female, and 0.4% (1/221) were identified as transgender male. Less than 2% (3/221) of deaths were reported without specifying the decedent's gender. The main analysis focused on articles reporting suicide deaths of cisgender men and women due to the underrepresentation of noncisgender identities in the sample. This paper targets the circumstances of suicide, more than one of which can be attributed to a single death and characterizes them in the results that follow.

Circumstances of Suicide

In total, 153 male decedents in our study had 159 circumstances and 64 female decedents had 75 circumstances (Table 1). We found 21 unique circumstance types among our sample of articles. Analyses described in this paper characterize the top 7 circumstances (Table 1) that occurred most frequently, which are (1) unspecified (n=60), (2) legal problem (n=41), (3) explicit statement of mental health symptoms or diagnosis (n=38), (4) social or relationship problem (n=35), (5) physical health problem (n=33), (6) financial or job problem (n=19), and (7) preceding suicidality (n=13).

Table 1. Summary of top 7 circumstances of suicide by gender annotated in 200 news articles.

Gender	Unspecified, n (%)	Legal problem, n (%)	Mental health symptoms or diagnosis, n (%)	Social or Relationship, n (%)	Physical health, n (%)	Financial or job, n (%)	Preceding suicidality, n (%)	Full sample, n
Cisgender men	35 (22)	31 (19.5)	29 (18.2)	21 (13.2)	24 (15.1)	10 (6.3)	9 (5.7)	159
Cisgender women	24 (32)	9 (12)	8 (10.7)	13 (17.3)	8 (10.7)	9 (12)	4 (5.3)	75
Transgender[a]	0 (0)	0 (0)	0 (0)	1 (100)	0 (0)	0 (0)	0 (0)	1
Unspecified[a]	1 (25)	1 (25)	1 (25)	0 (0)	1 (25)	0 (0)	0 (0)	4

[a]Excluded from additional analyses due to underrepresentation in the data set.

Framing Suicide by Gender

The distributions in Figure 1 illustrate the proportion of articles reporting a male (orange) or female (blue) suicide death with linguistic overlap with each SOSS attitude of stigma, isolation or depression, and glorification. In Figure 2, and subsequent figures, dotted lines represent mean cosine similarity scores across the 2 genders that are male (orange) or female (blue). The isolation or depression frame is excluded from the remaining analyses as no difference occurred in cosine similarity. Cosine similarity measures how similar 2 vectors are with increasing values representing increased similarity. A Mann-Whitney *U* test revealed that articles reporting male suicide deaths contained statistically significant more linguistic overlap with the stigma attitude on average than papers reporting female suicide deaths ($P<.001$). Such articles reporting male deaths also contained statistically significant more linguistic overlap with the glorification attitude on average than articles reporting female suicide deaths ($P=.005$). We found no statistically significant difference in the average linguistic similarity between papers reporting male or female suicide deaths ($P=.07$) regarding the isolation attitude.

Figure 1. Illustration of the relationship between research questions and analysis.

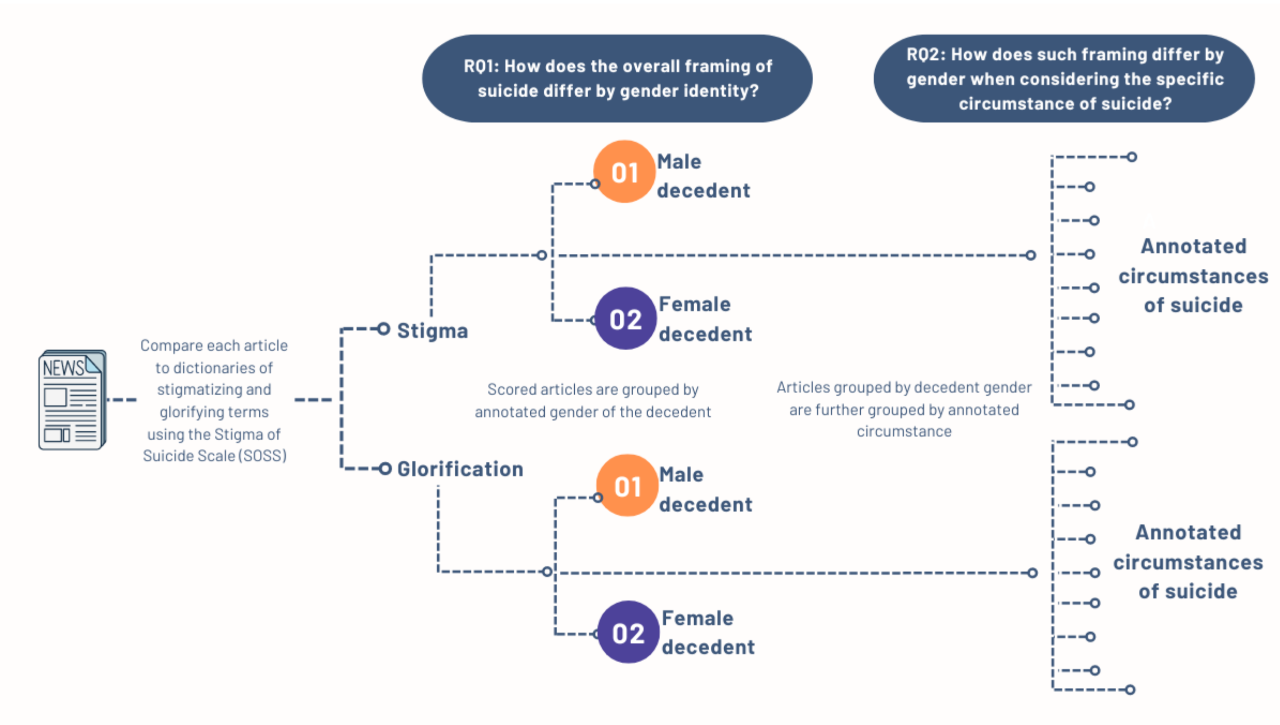
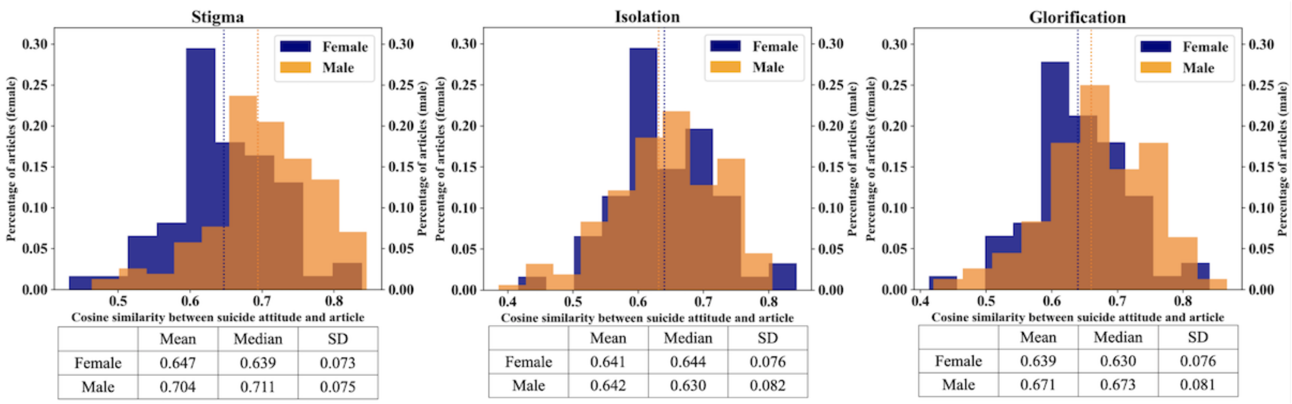


Figure 2. Percentage distribution plots of cosine similarity scores by gender compared to the Stigma of Suicide Scale dimensions.



Assessing IRR using Cohen κ for each SOSS dimension demonstrated acceptable reproducibility, ≥75%, between coders with scores of 0.84, 0.89, and 0.94 for stigma, isolation or depression, and glorification, respectively. Assessing IRR using Cohen κ for decedent gender and circumstance demonstrated acceptable reproducibility (≥75%) between coders with a score of 0.81.

Framing Circumstances of Suicide by Gender

We reported mutual information (MI) scores, which are widely used in the information theory literature, to determine differences by gender in stigmatizing or glorifying language as used under a particular circumstance of death. MI is an indicator of similarity (score farther from 0) or dissimilarity (score closer to 0) between two distributions [35]. In Table 2, the MI scores compare articles reporting male and female suicides to the stigma and glorification of SOSS attitudes for the 7 most frequently reported circumstances in the sample. Regarding stigmatizing language, reports of male and female suicide deaths were least similar when attributed to a legal problem (0.155), a social or relationship problem (0.268), or when no circumstance is described (0.312). Regarding glorification language, reports of male and female suicides were least similar when attributed to a legal problem (0.132), an explicit statement of mental health symptoms or diagnosis (0.251), or a physical health problem (0.320).

Table 2. Mutual information scores between male and female suicide deaths by circumstance. All the News data set—United States, 2015–2017. The closer the score is to 0, the less mutual information is present between the groups.

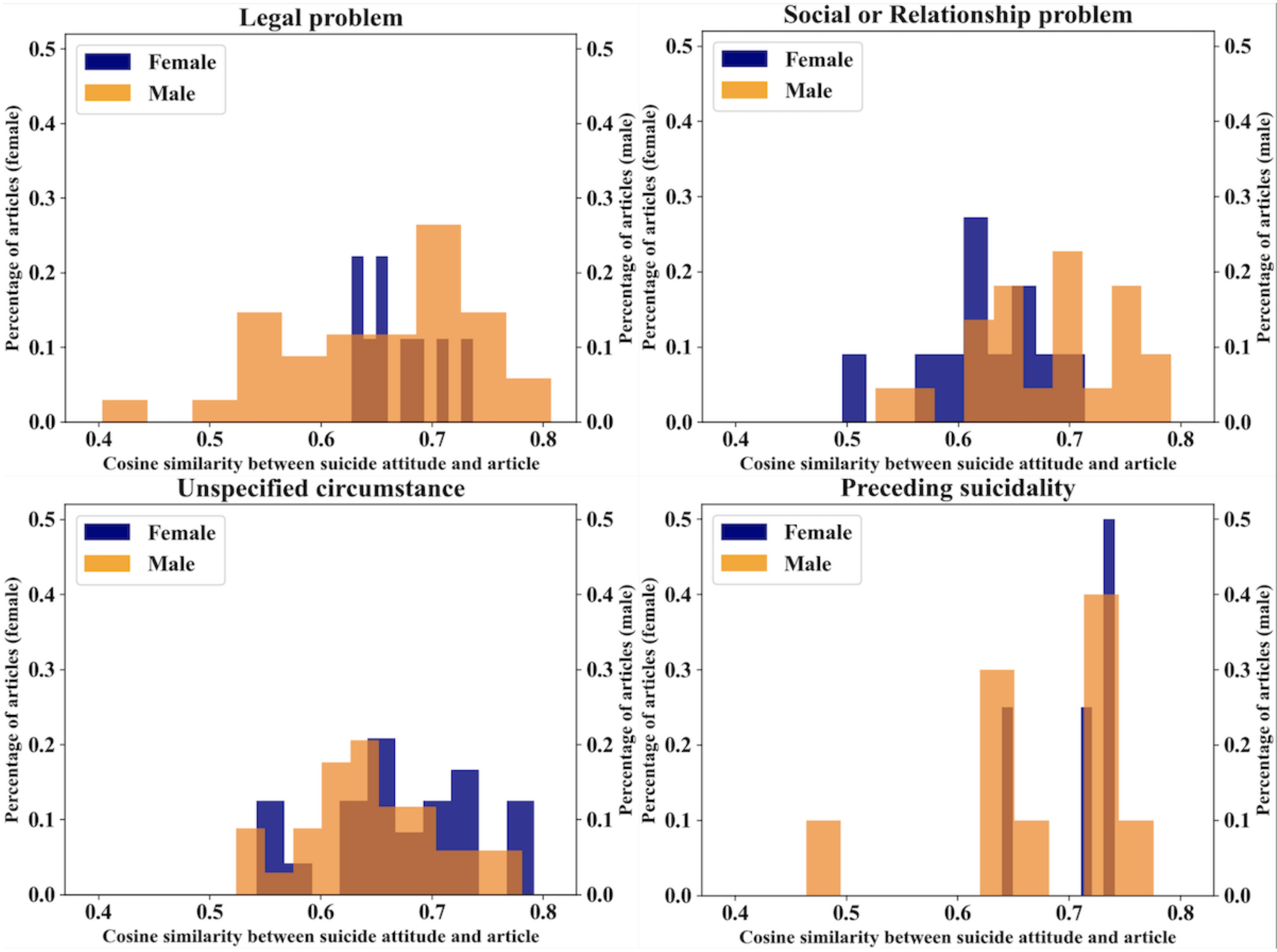
Circumstance	Stigma MI[a]	Glorification MI
Unspecified	0.312[b]	0.578
Legal problem	0.155[b]	0.132[b]
Explicit statement of mental health symptoms or diagnosis	0.421	0.251[b]
Social or relationship problem	0.268[b]	0.522
Physical health problem	0.443	0.320[b]
Financial or job problem	0.349	0.849
Preceding suicidality	0.343	0.441

[a]MI: mutual information.

[b]Circumstance with the least linguistic similarity between males and females for the stigma and glorification frame is further characterized in the text.

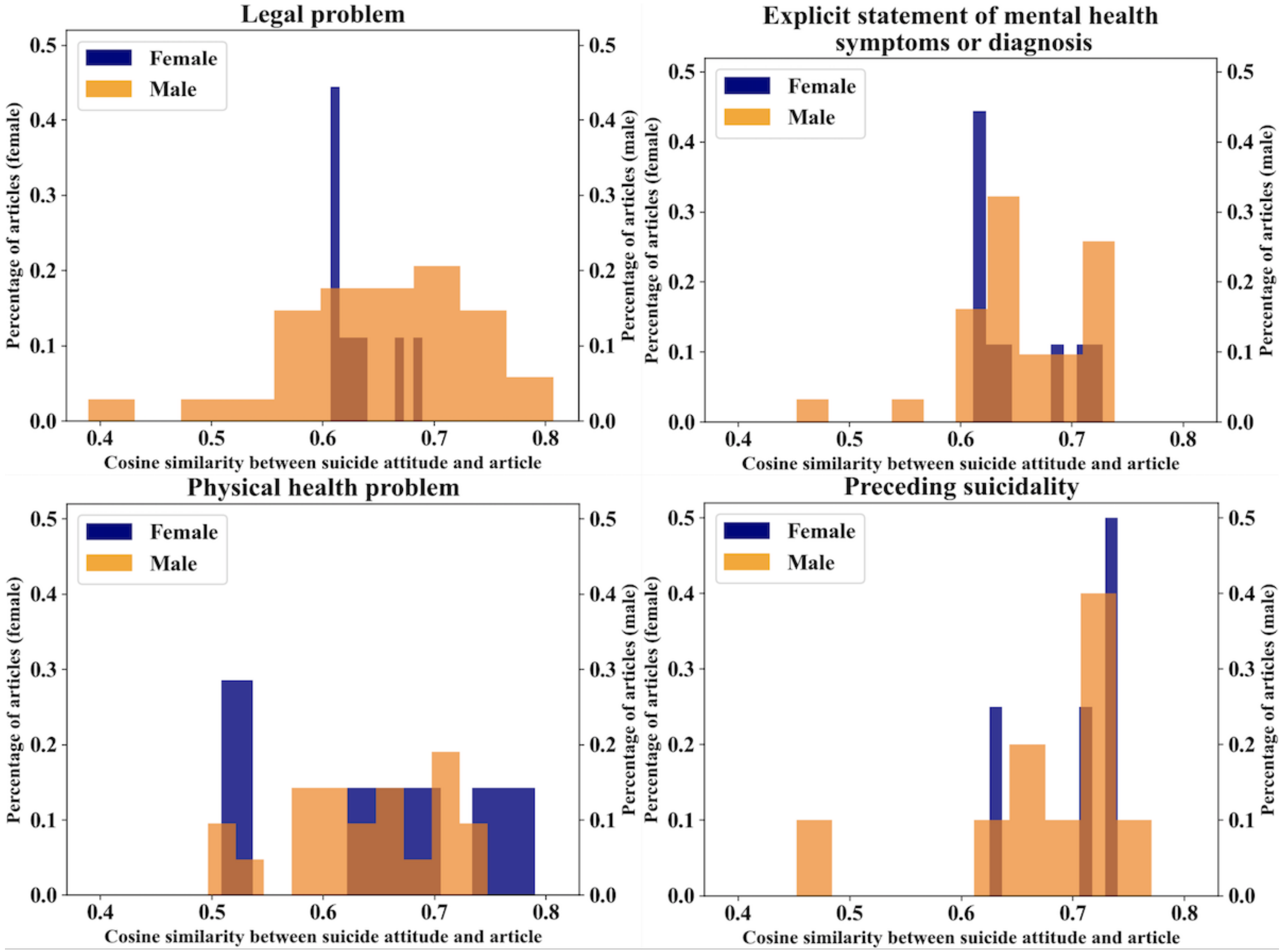
Article similarity to the stigma attitude for the circumstances with the greatest difference in MI is displayed in Figure 3. Mean cosine similarity measurement revealed that reports of female suicides attributed to a legal problem (0.671) or without any circumstance specified (0.673) were more linguistically similar to stigmatizing framing than reports of male suicides attributed to the same circumstances (Figure 3). However, male deaths attributed to a social or relationship issue (0.682) had greater linguistic similarity to stigmatizing framing than reports of female deaths of the same circumstance (0.624).

Figure 3. Percentage distribution of articles by gender and circumstance compared to the attitude of “stigma”.



Similar results hold for the circumstances with the least MI by gender regarding the glorification attitude. On average, reports of male suicides attributed to a legal problem (0.654), mental health symptom or diagnosis (0.657), or a physical health problem (0.639) were more linguistically similar to the glorification attitude than those reporting female suicides (Figure 4).

Figure 4. Percentage distribution of articles by gender and circumstance compared to the attitude of "glorification".



Suicidality had the fourth lowest MI score for both the stigma (0.343) and glorification (0.441) attitudes. The difference between the distributions of articles for this circumstance was less than the other circumstances; the linguistic similarity of reports of these suicides both to the stigma (Figure 3) and glorification (Figure 4) attitudes was higher than for other circumstances, regardless of decedent gender. The average cosine similarity to the glorification attitude for the suicidality circumstance was 0.702 for females and 0.656 for males in Figure 4. The average cosine similarity to the stigma attitude for suicidality as a circumstance was 0.708 for female deaths and 0.671 for male deaths.

Discussion

Key Findings

This study describes the reported gender of decedents and the circumstance of suicide in news articles and provides evidence that linguistic framing differs by both gender and circumstance.

Our comparison of SOSS attitudes in news articles revealed that descriptions of male suicides, on average, had higher linguistic scores for stigmatization and glorification than female suicides. This finding is important considering the observation that males die by suicide at higher rates than females in the United States, largely as a consequence of the greater use of firearms in suicide attempts [3]. Stigmatizing language, which can reduce help-seeking behavior [36], and language that glorifies or normalizes suicide acts [37], which may induce suicide-related contagion [38], are thereby concerning aspects of media reporting on suicide. This merits further attention as part of efforts to reduce high rates of suicide among males and other populations in need.

Circumstances of Suicide and Stigmatizing Language

We used MI scores to identify differences in stigmatizing and glorifying language by gender and circumstance, discerning for which groups these differences were most pronounced. The scores indicated that the greatest differences, by gender, in stigmatizing language were those that discussed a legal problem, social or relationship problem, or no specified circumstance.

Language in reports of female deaths attributed to legal problems had higher linguistic similarity to attitudes of stigma than reports of male deaths of the same circumstance. However, the inverse was true for male deaths attributed to a social or relationship problem. The proportion of articles in our sample attributing legal and social or relationship problems as circumstances to each gender is opposite of the proportions of disclosed circumstances in the National Violent Death Reporting System [6]. Increased stigmatizing language in reports of women experiencing legal problems and men experiencing relationship problems may be attributed to their perceived deviation from social norms [39]. Reports of female deaths not attributed to any circumstance were more linguistically similar to the stigma attitude than reports of male deaths—demonstrating the presence of discernable stigmatizing language even without specifying the circumstances of suicide. This shows that reporting suicide in more general details, per the safe reporting guidelines, may not eliminate stigmatization when framing death.

Circumstances of Suicide and Glorification Language

Similarly, when examining attitudes of glorification, news articles contain different language when reporting suicide deaths of men and women. We observed the greatest differences in language reflecting attitudes of glorification in articles that attributed death to a legal problem, explicit statement of mental health symptom or diagnosis, or a physical health problem. Specifically, reports of male deaths attributed to these problems had higher linguistic similarity to attitudes of glorification than those of female deaths for the same circumstances. Identifying these circumstances as being prone to using glorifying language for male suicide decedents may offer insights into areas where greater disparity exists in safe reporting by gender for more focused education among news organizations and their staff. For example, a correlation study of the SOSS used for this analysis indicates that male gender among respondents is correlated with decreased suicide literacy and increased stigma and glorifying attitudes of those who die by suicide [40]. As a 2021 study finds a greater stigma of suicide among medical professionals compared with the public [41], news publishers should consider education to increase suicide literacy among journalists.

Again, it should be noted that for a sizeable minority of deaths (22% for males and 32% for females), there was no specified circumstance. This may partly reflect the effect of the safe suicide reporting guidelines, which advise against oversimplifying or speculating the reason for the suicide [13]. Still, we find that harmful language persists in the absence of unsafe details.

Shoring Up Safe Reporting of Suicide

This work brings light to the knowledge about the potential hazards of new reports of suicide beyond existing safe reporting guidelines. Current guidelines encourage the omission of specific details from journalistic reporting of suicide deaths such as method, location, or personal details of the decedent [13]. The relationship between adherence to safe reporting guidelines and suicide prevention is well established [16,42,43]. However, our findings suggest that stigmatizing and glorifying framing, facilitated by harmful language, persists in news reports of suicide regardless of adherence to safe reporting guidelines. This is especially exhibited by the presence of such language even when a circumstance is not attributed to the suicide death (Figure 3). Previous work highlights that stigmatizing and glorifying language has implications on public perceptions of suicide [44-47] and prevention strategies of help-seeking [36,37]. Yet, strategies to mitigate such language in the reporting of suicide deaths could be expanded upon in contemporary guidelines [13].

In response to this gap, we call on future work to investigate the relationship between guideline adherence and harmful framing in news reports of suicide deaths and all other suicide-related behaviors. Such investigation informs the extension of existing journalistic guidelines to include unsafe language. The use of computational approaches in this study illustrates the feasibility of this task. Such methods may support the development of tools to detect the harmful language in news reporting of suicide as demonstrated in a similar recommendation of 3 models of artificial intelligence tools to improve adherence to safe reporting guidelines [48] as a key prevention strategy.

Additional Implications

Although gender minorities experience elevated risk and rates of suicide [9], our data do not allow us to discern if news media reports of suicide reflect these disparities or to compare reporting language between them. Gender minorities were underrepresented in this sample of suicide reports in the news even when the information was available. Existing research that explores suicide and gender focuses on risk factors for men and women [49]. This focus may be influenced by limited disclosure of nonconforming gender identity or default reporting of sex assigned at birth, often precluding public health surveillance [49]. Facing the same challenge, this work explored framing language in news reporting across all gender identities. Still, the absence of these identities in news reports prevents a complete understanding of stigmatizing language in these populations.

Our analysis revealed an unexpected finding. While we did not detect a discernable difference, by gender, in the linguistic similarity to stigma or glorification regarding the circumstance of preceding suicidality, reports of suicide attributing to this circumstance displayed greater similarity to both stigmatizing and glorifying attitudes compared with reports attributed to other circumstances. This suggests that a history of suicidal behavior is often reported in an unsafe way.

Limitations

Limitations of this study include its sample size, our capacity to describe or account for the audience consuming news from the venues analyzed from this American outlet–focused data set, the age of the articles, and the lack of availability of some sociodemographic attributes in annotated articles, all of which may preclude generalizability of the findings. Although the study team intended to annotate for age, race, and occupation in addition to gender, recognizing that identity is intersectional, these details were largely unavailable to inform the analysis, possibly due to adherence to the safe reporting guidelines. Although gender was available in the sample, gender minorities were underrepresented. Considering larger and more diverse corpora of news may allow for deeper analysis based on an intersectional lens.

Conclusions

Previous studies have investigated stigma and framing in news reports of suicide [47] or have focused on specific stigmatizing suicide terms [50]; however, this study examined suicide reporting language using a state-of-the-art natural language processing approach and revealed important differences by gender, and particularly stigmatizing and glorifying language for specific circumstances.

Safe reporting and messaging following a suicide is an important approach that can be used to lessen harm and reduce the future risk of suicide [51]. Journalistic adherence to media guidelines can prevent imitative suicide [42]; however, further opportunities exist to understand which elements of language, such as the 2 on which we focused in this study, are the most impactful [52]. While information management strategies such as those prescribed by the safe reporting guidelines seek to prevent harmful suicide reporting [53], actual efficacy and fidelity of adherence may differ based on the sociodemographic characteristics of the decedent, such as gender. The subtle, yet harmful framing of who experiences suicide in news media might influence prevention, intervention, and help-seeking behaviors [36], particularly among already disparaged or marginalized populations. Identifying sociodemographic differences in such framing supports the need to further understand suicide disparities and develop tailored stigma reduction strategies.

Acknowledgments

This project was supported, in part, by an appointment to the Research Participation Program at the Centers for Disease Control, and Prevention (CDC) administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the US Department of Energy and CDC. JCF, SM, and MDC were supported through a National Institutes of Health grant 1R01MH117172. The findings and conclusions in this paper are those of the authors and do not necessarily represent the official position of the CDC.

Authors' Contributions

JCF performed conceptualization, data curation, and methodology, and wrote the original draft. SM contributed to formal analysis, methodology, writing, and visualization. NL wrote, reviewed, and edited this paper. AC provided writing, reviewing, and editing expertise. EK provided writing, reviewing, and editing expertise in addition to data curation. SAS supervised the writing, reviewing, and editing. MDC conceptualized, supervised, acquired funding, and contributed to writing, reviewing, and editing this paper.

Conflicts of Interest

None declared.

References

1. About multiple cause of death, 1999-2020. Centers for Disease Control and Prevention. 2021. URL: http://wonder.cdc.gov/mcd-icd10.html [accessed 2022-10-13]
2. WISQARS (Web-Based Injury Statistics Query and Reporting System). Atlanta, GA: Centers for Disease Control and Prevention, National Center for Injury Prevention and Control; 2011 Apr 01. URL: https://wisqars.cdc.gov/ [accessed 2022-09-01]
3. Wang J, Sumner SA, Simon TR, Crosby AE, Annor FB, Gaylor E, et al. Trends in the incidence and lethality of suicidal acts in the United States, 2006 to 2015. JAMA Psychiatry 2020 Jul 01;77(7):684-693 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.0596] [Medline: 32320023]
4. Schrijvers DL, Bollen J, Sabbe BG. The gender paradox in suicidal behavior and its impact on the suicidal process. J Affect Disord 2012 Apr;138(1-2):19-26. [doi: 10.1016/j.jad.2011.03.050] [Medline: 21529962]
5. Bommersbach TJ, Rosenheck RA, Petrakis IL, Rhee TG. Why are women more likely to attempt suicide than men? Analysis of lifetime suicide attempts among US adults in a nationally representative sample. J Affect Disord 2022 Aug 15;311:157-164. [doi: 10.1016/j.jad.2022.05.096] [Medline: 35598742]
6. Chen T, Roberts K. Negative life events and suicide in the National Violent Death Reporting System. Arch Suicide Res 2021;25(2):238-252. [doi: 10.1080/13811118.2019.1677275] [Medline: 31638469]
7. Chang S, Stuckler D, Yip P, Gunnell D. Impact of 2008 global economic crisis on suicide: time trend study in 54 countries. BMJ 2013;347:f5239 [FREE Full text] [doi: 10.1136/bmj.f5239] [Medline: 24046155]
8. Carretta RF, McKee SA, Rhee TG. Gender differences in risks of suicide and suicidal behaviors in the USA: a narrative review. Curr Psychiatry Rep 2023;25(12):809-824. [doi: 10.1007/s11920-023-01473-1] [Medline: 37930559]
9. Centers for Disease Control and Prevention. Disparities in Suicide. 2022. URL: https://www.cdc.gov/suicide/disparities/ [accessed 2022-08-12]
10. Phillips DP. The influence of suggestion on suicide: substantive and theoretical implications of the Werther effect. Am Sociol Rev 1974;39(3):340-354. [Medline: 11630757]
11. Sinyor M, Tran US, Garcia D, Till B, Voracek M, Niederkrotenthaler T. Suicide mortality in the United States following the suicides of Kate Spade and Anthony Bourdain. Aust N Z J Psychiatry 2021;55(6):613-619. [doi: 10.1177/0004867420976844] [Medline: 33300363]
12. Niederkrotenthaler T, Voracek M, Herberth A, Till B, Strauss M, Etzersdorfer E, et al. Role of media reports in completed and prevented suicide: Werther v. Papageno effects. Br J Psychiatry 2010;197(3):234-243. [doi: 10.1192/bjp.bp.109.074633] [Medline: 20807970]

13. SAVE. Best practices and recommendations for reporting on suicide. Reporting on Suicide. 2020. URL: https://reportingonsuicide.org/recommendations/ [accessed 2022-10-12]
14. Batterham PJ, Calear AL, Christensen H. The Stigma of Suicide Scale. Psychometric properties and correlates of the stigma of suicide. Crisis 2013;34(1):13-21. [doi: 10.1027/0227-5910/a000156] [Medline: 22846447]
15. Whorf BL. In: John BC, editor. Language, Thought, Reality: Selected Writings of Benjamin Lee Whorf. Cambridge, MA: Technology Press of MIT; 1956:278.
16. Sumner SA, Burke M, Kooti F. Adherence to suicide reporting guidelines by news shared on a social networking platform. Proc Natl Acad Sci U S A 2020;117(28):16267-16272 [FREE Full text] [doi: 10.1073/pnas.2001230117] [Medline: 32631982]
17. Thompson A. All the news. 2017. URL: https://components.one/datasets/all-the-news-2-news-articles-dataset [accessed 2022-10-12]
18. Choi D, Sumner SA, Holland KM, Draper J, Murphy S, Bowen DA, et al. Development of a machine learning model using multiple, heterogeneous data sources to estimate weekly US suicide fatalities. JAMA Netw Open 2020;3(12):e2030932 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.30932] [Medline: 33355678]
19. Hagar N, Wachs J, Horvát E. Writer movements between news outlets reflect political polarization in media. New Media & Society 2021;25(8):2034-2056. [doi: 10.1177/14614448211027173]
20. Tewari S, Zabounidis R, Kothari A, Bailey R, Alm CO. Perceptions of human and machine-generated articles. Digital Threats 2021;2(2):1-16. [doi: 10.1145/3428158]
21. Sridhar S, Sanagavarapu S. Content based news recommendation engine using Hybrid BiLSTM-ANN feature modelling. : IEEE; 2021 Presented at: Kitakyushu, Japan; 2021 August 16-20; Kitakyushu, Japan. [doi: 10.1109/icievicivpr52578.2021.9564198]
22. Pirkis J, Blood RW. Suicide and the media. Part II: portrayal in fictional media. Crisis 2001;22(4):155-162. [doi: 10.1027//0227-5910.22.4.155] [Medline: 11848659]
23. Chancellor S, Sumner SA, David-Ferdon C, Ahmad T, De Choudhury M. Suicide risk and protective factors in online support forum posts: Annotation Scheme Development and Validation Study. JMIR Ment Health 2021;8(11):e24471 [FREE Full text] [doi: 10.2196/24471] [Medline: 34747705]
24. Hawton K, van Heeringen K. Suicide. Lancet 2009 Apr 18;373(9672):1372-1381. [doi: 10.1016/S0140-6736(09)60372-X] [Medline: 19376453]
25. Clark K, Khandelwal U, Levy O, Manning C. What does BERT look at an analysis of BERT's attention. Florence, Italy: Association for Computational Linguistics; 2019 Presented at: Computation and Language; 2019 Jun 11; Florence, Italy. [doi: 10.18653/v1/w19-4828]
26. Coenen A, Reif E, Yuan A, Kim B, Pearce A, Viégas F, et al. Visualizing and measuring the geometry of BERT. Proceedings of the 33rd International Conference on Neural Information Processing Systems 2019:8594-8603 [FREE Full text] [doi: 10.5555/3454287.3455058]
27. Rudkowsky E, Haselmayer M, Wastian M, Jenny M, Emrich Š, Sedlmair M. More than Bags of Words: Sentiment Analysis with Word Embeddings. Communication Methods and Measures 2018;12(2-3):140-157 [FREE Full text] [doi: 10.1080/19312458.2018.1455817]
28. Shao Y, Taylor S, Marshall N, Morioka C, Zeng-Treitler Q. Clinical text classification with word embedding features vs. bag-of-words features. USA: IEEE; 2018 Presented at: 2018 IEEE International Conference on Big Data (Big Data); 2018 Dec 10-13; Seattle, WA, USA. [doi: 10.1109/bigdata.2018.8622345]
29. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: Computation and Language (cs.CL); 2019 May 24; Minneapolis, Minnesota p. 4171-4186. [doi: 10.18653/v1/N19-1423]
30. Miller GA. WordNet: a lexical database for English. Commun ACM 1995;38(11):39-41. [doi: 10.1145/219717.219748]
31. Wolf T, Debut L, Sanh V. Transformers: state-of-the-art Natural Language Processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Online: Association for Computational Linguistics; 2020 Presented at: The 2020 Conference on Empirical Methods in Natural Language Processing; 16-20 November; Virtual p. 38-45. [doi: 10.18653/v1/2020.emnlp-demos.6]
32. Kwak H, An J, Jing E, Ahn Y. FrameAxis: characterizing microframe bias and intensity with word embedding. PeerJ Comput Sci 2021;7:e644 [FREE Full text] [doi: 10.7717/peerj-cs.644] [Medline: 34395864]
33. Johns MM, Lowry R, Andrzejewski J, Barrios LC, Demissie Z, McManus T, et al. Transgender identity and experiences of violence victimization, substance use, suicide risk, and sexual risk behaviors among high school students - 19 states and large urban school districts, 2017. MMWR Morb Mortal Wkly Rep 2019 Jan 25;68(3):67-71. [doi: 10.15585/mmwr.mm6803a3] [Medline: 30677012]
34. Exempt IRB review. Georgia Tech Office of Research Integrity Assurance. URL: https://oria.gatech.edu/irb/submitting-protocol/exempt-irb-review [accessed 2024-06-05]
35. Tzannes NS, Noonan JP. The mutual information principle and applications. Inf Control 1973 Feb;22(1):1-12. [doi: 10.1016/s0019-9958(73)90448-8]

36. Barney LJ, Griffiths KM, Jorm AF, Christensen H. Stigma about depression and its impact on help-seeking intentions. Aust N Z J Psychiatry 2006;40(1):51-54. [doi: 10.1080/j.1440-1614.2006.01741.x] [Medline: 16403038]
37. Niederkrotenthaler T, Reidenberg DJ, Till B, Gould MS. Increasing help-seeking and referrals for individuals at risk for suicide by decreasing stigma. Am J Prev Med 2014;47(3 Suppl 2):S235-S243. [doi: 10.1016/j.amepre.2014.06.010]
38. Niederkrotenthaler T, Braun M, Pirkis J, Till B, Stack S, Sinyor M, et al. Association between suicide reporting in the media and suicide: systematic review and meta-analysis. BMJ 2020 Mar 18;368:m575 [FREE Full text] [doi: 10.1136/bmj.m575] [Medline: 32188637]
39. Dempsey RC, Fedorowicz SE, Wood AM. The role of perceived social norms in non-suicidal self-injury and suicidality: a systematic scoping review. PLoS One 2023;18(6):e0286118 [FREE Full text] [doi: 10.1371/journal.pone.0286118] [Medline: 37352219]
40. Batterham PJ, Calear AL, Christensen H. Correlates of suicide stigma and suicide literacy in the community. Suicide Life Threat Behav 2013;43(4):406-417. [doi: 10.1111/sltb.12026] [Medline: 23556504]
41. Eilers JJ, Kasten E, Schnell T. Comparison of stigmatization of suicidal people by medical professionals with stigmatization by the general population. Healthcare (Basel) 2021 Jul 15;9(7):896 [FREE Full text] [doi: 10.3390/healthcare9070896] [Medline: 34356273]
42. Bohanna I, Wang X. Media guidelines for the responsible reporting of suicide: a review of effectiveness. Crisis 2012;33(4):190-198. [doi: 10.1027/0227-5910/a000137] [Medline: 22713977]
43. Roth R, Abraham J, Zinzow H, Wisniewski P, Khasawneh A, Chalil Madathil K. Evaluating news media reports on the 'Blue Whale Challenge' for adherence to suicide prevention safe messaging guidelines. Proc ACM Hum-Comput Interact 2020 May 29;4(CSCW1):1-27. [doi: 10.1145/3392831]
44. Nathan NA, Nathan KI. Suicide, stigma, and utilizing social media platforms to gauge public perceptions. Front Psychiatry 2019;10:947 [FREE Full text] [doi: 10.3389/fpsyt.2019.00947] [Medline: 31998162]
45. Lee H, An S. Social stigma toward suicide: effects of group categorization and attributions in Korean health news. Health Commun 2016;31(4):468-477. [doi: 10.1080/10410236.2014.966894] [Medline: 26485582]
46. Li A, Huang X, Jiao D, O'Dea B, Zhu T, Christensen H. An analysis of stigma and suicide literacy in responses to suicides broadcast on social media. Asia Pac Psychiatry 2018 Mar;10(1):e12314. [doi: 10.1111/appy.12314] [Medline: 29383880]
47. Boudry V. Suicide story frames contribute to stigma. Newsp Res J 2008 Mar 01;29(2):55-69. [doi: 10.1177/073953290802900205]
48. Ransing R, Menon V, Kar S, Arafat SMY. Artificial intelligence-based models for augmenting media reporting of suicide: challenges and opportunities. Glob Psychiatry Arch 2021;4(2):123-129. [doi: 10.52095/gp.2021.3037.1017]
49. Martínez AAB. A critical literature review of the research on suicide from a gender perspective. Soc Med 2019;12(2):118-124.
50. Arendt F. Framing suicide - investigating the news media and public's use of the problematic suicide referents Freitod and Selbstmord in German-speaking countries. Crisis 2018 Jan;39(1):70-73. [doi: 10.1027/0227-5910/a000467] [Medline: 29442550]
51. Suicide prevention resource for action. Centers for Disease Control and Prevention. 2023 Jun 15. URL: https://www.cdc.gov/suicide/resources/prevention.html [accessed 2023-06-19]
52. Stack S. Media guidelines and suicide: a critical review. Soc Sci Med 2020 Oct;262:112690. [doi: 10.1016/j.socscimed.2019.112690] [Medline: 32067758]
53. Zhang R, Wang MS, Toubiana M, Greenwood R. Stigma beyond levels: advancing research on stigmatization. Acad Manag Ann 2021;15(1):188-222. [doi: 10.5465/annals.2019.0031]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
CDC: Centers for Disease Control and Prevention
IRR: interrater reliability
MI: mutual information
SOSS: Stigma of Suicide Scale

Edited by J Torous; submitted 19.09.23; peer-reviewed by A Fang, E Jaafar; comments to author 08.12.23; revised version received 07.03.24; accepted 25.03.24; published 03.07.24.

Please cite as:
Foriest JC, Mittal S, Kim E, Carmichael A, Lennon N, Sumner SA, De Choudhury M
News Media Framing of Suicide Circumstances and Gender: Mixed Methods Analysis
JMIR Ment Health 2024;11:e49879
URL: https://mental.jmir.org/2024/1/e49879
doi:10.2196/49879
PMID:

Original Paper

Bayesian Networks for Prescreening in Depression: Algorithm Development and Validation

Eduardo Maekawa[1,2], MSc; Eoin Martino Grua[1,2], PhD; Carina Akemi Nakamura[3], PhD; Marcia Scazufca[3,4], PhD; Ricardo Araya[5], PhD; Tim Peters[6], PhD; Pepijn van de Ven[1,2], PhD

[1]Department of Electronic and Computer Engineering, University of Limerick, Limerick, Ireland

[2]Health Research Institute, University of Limerick, Limerick, Ireland

[3]Departamento de Psiquiatria, Faculdade de Medicina da Universidade de Sao Paulo, Universidade de Sao Paulo, Sao Paulo, Brazil

[4]Instituto de Psiquiatria, Hospital das Clinicas da Faculdade de Medicina da Universidade de Sao Paulo, Faculdade de Medicina, Universidade de Sao Paulo, Sao Paulo, Brazil

[5]Centre for Global Mental Health, King's College London, London, United Kingdom

[6]Bristol Dental School, University of Bristol, Bristol, United Kingdom

Corresponding Author:
Eduardo Maekawa, MSc
Department of Electronic and Computer Engineering
University of Limerick
Plassey Park Road
Limerick, V94 T9PX
Ireland
Phone: 353 830150601
Email: eduardo.maekawa@ul.ie

Abstract

Background: Identifying individuals with depressive symptomatology (DS) promptly and effectively is of paramount importance for providing timely treatment. Machine learning models have shown promise in this area; however, studies often fall short in demonstrating the practical benefits of using these models and fail to provide tangible real-world applications.

Objective: This study aims to establish a novel methodology for identifying individuals likely to exhibit DS, identify the most influential features in a more explainable way via probabilistic measures, and propose tools that can be used in real-world applications.

Methods: The study used 3 data sets: PROACTIVE, the Brazilian National Health Survey (Pesquisa Nacional de Saúde [PNS]) 2013, and PNS 2019, comprising sociodemographic and health-related features. A Bayesian network was used for feature selection. Selected features were then used to train machine learning models to predict DS, operationalized as a score of ≥10 on the 9-item Patient Health Questionnaire. The study also analyzed the impact of varying sensitivity rates on the reduction of screening interviews compared to a random approach.

Results: The methodology allows the users to make an informed trade-off among sensitivity, specificity, and a reduction in the number of interviews. At the thresholds of 0.444, 0.412, and 0.472, determined by maximizing the Youden index, the models achieved sensitivities of 0.717, 0.741, and 0.718, and specificities of 0.644, 0.737, and 0.766 for PROACTIVE, PNS 2013, and PNS 2019, respectively. The area under the receiver operating characteristic curve was 0.736, 0.801, and 0.809 for these 3 data sets, respectively. For the PROACTIVE data set, the most influential features identified were postural balance, shortness of breath, and how old people feel they are. In the PNS 2013 data set, the features were the ability to do usual activities, chest pain, sleep problems, and chronic back problems. The PNS 2019 data set shared 3 of the most influential features with the PNS 2013 data set. However, the difference was the replacement of chronic back problems with verbal abuse. It is important to note that the features contained in the PNS data sets differ from those found in the PROACTIVE data set. An empirical analysis demonstrated that using the proposed model led to a potential reduction in screening interviews of up to 52% while maintaining a sensitivity of 0.80.

Conclusions: This study developed a novel methodology for identifying individuals with DS, demonstrating the utility of using Bayesian networks to identify the most significant features. Moreover, this approach has the potential to substantially reduce the number of screening interviews while maintaining high sensitivity, thereby facilitating improved early identification and intervention strategies for individuals experiencing DS.

(JMIR Ment Health 2024;11:e52045) doi:10.2196/52045

KEYWORDS

Bayesian network; target depressive symptomatology; probabilistic machine learning; stochastic gradient descent; patient screening; depressive symptom; machine learning model; machine learning; survey; prediction; socioeconomic data sets; utilization; depression; mental health; digital mental health; artificial intelligence; AI; prediction; prediction modeling; patient; mood; anxiety; mood disorders; mood disorder; eHealth; mobile health; mHealth; telehealth

Introduction

Background

Improving the identification and management of depression in primary care remains a global challenge. A meta-analysis has revealed that in primary care, approximately 50% of patients with depression receive diagnoses, while around 15% acquire treatment [1]. While screening for depressive symptomatology (DS) holds significance, it alone falls short of being effective [2].

When evaluating individuals with DS, one approach involves the use of screening tools to determine who may require treatment and further investigation. Among these tools, the 9-item Patient Health Questionnaire (PHQ-9) is a widely used, self-administered questionnaire [3].

In certain scenarios, mental health information (such as the PHQ-9) may not be available, whereas other health or behavioral information that can be linked to an increased risk of depression may be abundant. Examples of such data are demographics, medical history, lifestyle indicators, and socioeconomic status. In such scenarios, it may be useful to leverage the available data to identify individuals likely to have DS such that these individuals can be targeted proactively for interventions. This proactive strategy has the potential to provide necessary support and care before symptoms escalate or result in severe consequences.

Data-driven approaches using machine learning offer an appealing opportunity to design better prescreening methodologies, particularly within primary care settings. Machine learning algorithms can identify patterns in the data that may not be obvious to human experts [4-6].

In many studies involving machine learning models, insights beyond the model development phase are not provided. The results usually provide a simple *Yes* or *No* for the presence of depression, usually operating as “black boxes” with an absence of transparency in the decision-making process [7,8]. There is a lack of understanding as to which features are most important in predicting DS.

There are instances where machine learning models offer some level of explainability [9-12]. However, the insights generated by these models often only consider the predictive power of the predictors. While this approach can help identify the most impactful predictor, it does not provide a probabilistic measure, which is essential for dealing with uncertainty. The use of probability-based measures in machine learning models can provide clinicians with a more nuanced understanding of an individual’s likelihood of experiencing depression.

Moreover, most machine learning models developed for detecting DS fail to provide applications on how to use the models effectively [13-16]. These studies do not develop practical tools to help users benefit from the machine learning models, focusing solely on the model development process. To encourage the use of machine learning models in clinical practice, it is essential to develop machine learning models that are designed with the end user in mind. This includes creating tools that can help practitioners effectively integrate the model into their practice.

The current literature on the use of machine learning models for predicting DS lacks a comprehensive integration of explainability and transparency to identify the important features associated with DS using only general health and socioeconomic data and in the absence of tools such as the PHQ-9 or any other depression-related features. In addition, there is a shortage of practical applications using machine learning to support clinical practice. In our proposed method, we aim to address these gaps.

Objective

The first objective of this study is to establish a replicable prescreening proof-of-concept methodology for the detection of individuals with DS. Using solely general health and socioeconomic data, we aim to demonstrate the potential for such data in identifying individuals who might benefit from further screening. The second objective is to enable the identification of the most influential predictors with probabilistic insights into their importance, all based on the same general health and socioeconomic data. The third objective is to develop a tool that enables specialists to use the benefits of the methodology in their practice.

This paper’s structure comprises a *Methods* section detailing data preparation, the feature selection technique, and probabilistic insight extraction. It introduces a machine learning algorithm and its applications. The *Results* section presents cleaned data sets, chosen features, the most influential predictors, and the performance of the models and showcases the models’ utility in a practical scenario. The *Discussion* section concludes by summarizing findings, acknowledging limitations, and discussing implications.

Methods

Overview

Our proposed methodology for developing a machine learning model to assess people with DS was applied to 3 distinct Brazilian data sets. The first, known as the PROACTIVE data set [17], comprised individuals aged ≥60 years residing in socioeconomically deprived areas of Guarulhos city. The participants were registered in 20 primary care clinics in Guarulhos and were approached, according to a randomly ordered list, for a DS screening interview conducted either in person or via phone using a personalized app [18].

The other 2 data sets, Pesquisa Nacional de Saúde (PNS) 2013 and PNS 2019 [19], resulted from a Brazilian national health survey that assessed individuals aged ≥18 years in different sociodemographic groups and health behaviors. The surveys were conducted in 2013 and 2019 using a household approach where they applied stratified sampling.

All 3 studies used the PHQ-9, which is a 9-item questionnaire that serves as a screening tool for assessing DS. All questions are related to the previous 2 weeks, with responses to each question scored from 0 to 3, where 0 means “Not at all” and 3 means “Nearly every day.” The PHQ-9 cut-off score commonly used for DS is ≥10 [20-22], and we used this to create a binary classification target for the machine learning model. Summary statistics of the sociodemographic data and the prevalence of DS are presented in Tables S1, S2, and S3 in Multimedia Appendix 1 for the PROACTIVE, PNS 2013, and PNS 2019 data sets, respectively.

Ethical Considerations

The PROACTIVE trial received approval from the Comitê de Ética em Pesquisa Faculdade de Medicina da Universidade de São Paulo and authorization from the Guarulhos Health Secretary (number 2.836.569). The Brazilian National Health Ethics Research Committee of the Brazilian National Health Council approved the PNS 2013 (number 328.159) and PNS 2019 (number 3.529.376) surveys. Anonymized versions of the PNS 2013 and PNS 2019 surveys are publicly available for download and analysis. All participants provided informed consent.

Data Preparation

First, we randomly divided each data set into training (70%) and test (30%) sets. The development of the models is performed using the training data set with the test data set strictly used to test the performance of the created models on data not yet seen. Next, we created our response variable by summing the recorded responses of all the PHQ-9 score items for each participant and recording a 1 when the total sum was ≥10 and 0 otherwise. We only included participants who answered all 9 items, and the respondents were aged ≥18 years.

We dropped all features from the data sets related to depression, as our aim was to solely use data readily available in health platforms, such as general health and socioeconomic information. We also dropped features and patients with more than 20% missing values [23] and features with just 1 level in the responses obtained. Furthermore, we categorized all numeric features with more than 30 levels into 4 bins by quartiles (25, 50, 75, and 100).

We then transformed all features into ordinal numbers. To address any remaining missing values, we adopted an approach discussed in the study by Enders [24] by creating a new missing class and assigning it a value of 0. This choice was based on the fact that not all features had a value of 0 and that 0 is a value close to the range of the feature values. This also facilitates model development, as the standardization process that we applied after the missing imputation step is less sensitive when the range values are close to each other. Standardization is applied to ensure that all features have a similar scale, which helps in comparing the importance of different features and improving the model’s overall performance [25].

Feature Selection

The development of a machine learning model generally involves selecting relevant features to be used as inputs. A total of 1 approach to this task is constructing a Bayesian network (BN), which is a directed acyclic graph composed of nodes and edges. In this context, nodes represent the features and the directed edges represented by arrows illustrate the relationships among them [26].

BNs have been successfully applied in different scenarios, such as feature selection [27-29], model prediction [30-32], and providing insights for decision-making [33-36]. Moreover, a BN is a useful tool for visualizing and interpreting complex relationships between features. This enables the identification of critical features and their impact on the outcome.

BNs can be constructed using two primary methods, namely (1) manual node definition and edge direction [37] or (2) learning through data [38,39].

In our approach, we apply the Incremental Association Markov Blanket algorithm to learn the BN from data [40]. The Markov blanket (MB) of a node X, represented by MB(X), is characterized by its parents (nodes that have arrows pointing toward X), children (nodes that receive arrows from X), and spouses (nodes with arrows leading to children of X, yet not linked to X).

To evaluate the confidence level of the BN learned from the data, we used a bootstrap approach [41] by generating 1000 samples of the BN. We then computed the probabilities of having an edge between every pair of nodes (Xi, Xj; known as strength) and the probabilities of having a directed edge from Xi to Xj and from Xj to Xi (known as direction) [42]. To construct our final BN, we selected the edges with a strength of 50% and higher [43] and set their direction to correspond with the majority of the bootstrap models.

Subsequently, we conducted another test to assess the strength of the BN by performing an independence test between every pair of nodes using mutual information [44]. We used an independence test because it can investigate whether 2 features are statistically independent, implying that the occurrence or value of 1 feature does not influence the other. If the test yielded insufficient evidence of nonindependence (*P*≥.05), then that edge was removed from the network.

Conventionally, in the literature [45,46], BN-based feature selection is predominantly performed using the MB of the outcome of interest. However, we also tested the efficacy of using all the features connected to the outcome by a path (we call this All-path-features). A feature was considered part of this path if it was possible to reach the outcome node by traversing the edges and nodes in between, regardless of the directions of the edges. Therefore, we tested 2 scenarios of the model: 1 using All-path-features in the path and another using only the MB of the outcome.

BN Parameter Learning

Once the BN structure is constructed, it is possible to assess the relationships between the features. These relationships are probabilistically expressed through conditional probability distributions (CPDs) [47]. The process of estimating the CPD is known as parameter learning. The CPD specifies the probability distribution for each node given its parents. To estimate the CPD, we used the Bayesian method [48]. The Bayesian method is a powerful tool for estimating the CPD, as it takes into account both prior knowledge and observed data.

In this study, we computed the CPD table of the outcome node given its parents. The parents represent the features on which our outcome is conditionally dependent, and the CPD table provides the probability values of having DS for different combinations of parent values. We examined the combinations of parent values that exhibited the most discriminatory power when considering individuals with DS.

The data preparation, construction, and parameter learning of the BN were implemented using R 4.2.2 with libraries bnlearn, parallel, and base. R is freely available open source software (R Foundation for Statistical Computing).

Training of Prediction Models

In the training phase of model development, particularly in a classification problem with 2 classes, it is common practice to oversample the minority class [49-51]. This is because by oversampling the minority class, we can improve the model's ability to learn from both classes and achieve better overall performance.

For the development of the models, our approach applied stochastic gradient descent (SGD), which is a widely used optimization algorithm in machine learning. The objective of SGD is to minimize an error function through an iterative process where the model's parameters are updated at each step until the algorithm converges. SGD is known for its scalability [52], stability, and robustness [53] and has shown good performance across different domains [54-56].

One of the drawbacks of using SGD is that finding the optimal combination of hyperparameters for each data set can be challenging. We used BayesSearchCV from Python's skopt package, which applies both Bayesian optimization and 5-fold cross-validation to evaluate each model's performance [57]. The selection of the optimal hyperparameter set for each data set was done by selecting the configuration that yielded the highest area under the receiver operating characteristic curve (AUC-ROC) metric.

The study focused on tuning 3 hyperparameters: the loss function, penalty term, and α coefficient. The loss function characterizes the relationship between the model's predictions and the actual values. A total of 3 types of loss functions were examined: Hinge, Modified Huber, and Log [58,59].

The penalty term corresponds to a regularization technique aimed at enhancing the model's generalization capability. In addition, the α coefficient, a positive value, controls the level of regularization applied. In this study, 3 types of regularization for the penalty term were considered: L1 [60], L2 [61], and elastic net [62].

To explore the effect of the α coefficient on regularization, a range of values from 0.000001 to 1,000,000 were used.

Evaluation of the Models

We evaluated the models using two scenarios: (1) All-path-features and (2) the MB of the outcome, as mentioned in the *Feature Selection* section. For each scenario, we used the best set of hyperparameters discussed in the *Training of Prediction Models* section.

To validate each model, we once again applied 5-fold cross-validation on top of the cross-validation for hyperparameter tuning explained in the previous section. For each fold, we recorded the threshold that optimized both sensitivity and specificity simultaneously determined by the Youden index. Afterward, by averaging the thresholds obtained from the folds, we analyzed the metrics of AUC-ROC, sensitivity, and specificity. In addition, we calculated the mean and SD of these metrics across the folds. Sensitivity, specificity, and AUC-ROC were chosen as evaluation metrics to assess the performance of the models, given their advantages in the screening process.

To assess the performance of each machine learning model, we used the test data, which was set aside during the data preparation stage. We used sensitivity, specificity, and AUC-ROC to assess the performance of the model, applying their respective thresholds calculated according to the previous paragraph. Then, we compared the results to those obtained from the training data. By comparing these results, we gained insights into the model's generalization capabilities and its performance on new and unseen instances.

The model development was implemented in Python 3.7.7 using packages SGDClassifier, CalibratedClassifierCV from sklearn, and BayesSearchCV from skopt. Python is freely available open source software (Python Software Foundation).

Application of the Models

Having identified the most important features and created a model to use these features to predict DS for each data set, the next step was to analyze how the models can help target people with DS and what benefits they offer. To this end, we used the test data, which were previously only used to evaluate the performance of the models, to illustrate the relationship between the reduced screening interviews that can be obtained by selecting screening participants based on the developed model and the sensitivity and specificity of the overall screening methodology.

To illustrate the benefits of using these models for selecting screening participants, consider a specific cohort with a DS prevalence of 10%. If we were to screen this cohort randomly, in a study with 100 participants having DS, 1000 individuals would need to be screened. From a public health perspective, 10 individuals would need to be screened for every identified individual.

To assess the effectiveness of the models, consider a scenario where we need to screen 40 individuals using a random approach (Figure 1). For a 10% DS prevalence, we would identify 4 people having DS.

Figure 1. Example of a random screening list. DS: depressive symptomatology; P(DS): probability of having depressive symptomatology.

ID	DS	P(DS)	ID	DS	P(DS)	ID	DS	P(DS)	ID	DS	P(DS)
1	No	0.02	11	No	0.06	21	No	0.43	31	No	0.71
2	No	0.06	12	No	0.22	22	No	0.38	32	No	0.05
3	No	0.12	13	No	0.22	23	No	0.24	33	No	0.14
4	No	0.22	14	No	0.04	24	No	0.16	34	Yes	0.97
5	No	0.01	15	Yes	0.53	25	No	0.08	35	No	0.28
6	No	0.09	16	No	0.44	26	No	0.08	36	No	0.37
7	No	0.08	17	No	0.22	27	Yes	0.68	37	No	0.15
8	No	0.13	18	No	0.12	28	No	0.27	38	No	0.49
9	No	0.09	19	No	0.19	29	No	0.22	39	No	0.02
10	Yes	0.86	20	No	0.18	30	No	0.05	40	No	0.48

The models developed by our approach can be used to prioritize the individuals with the highest risk of DS for screening. This is because the model provides a probability score for each individual having DS. Hence, by ordering the cohort by this probability, and screening individuals in order of decreasing the probability of DS, the screening process is expected to be much more efficient.

A ranked list example is illustrated in Figure 2. If we use our models and start to screen individuals from 1 to 10, after screening 5 individuals (based on the probability score of having DS given by our models with a threshold >0.5), we would expect to identify 4 true positives (TPs) who have DS and 1 false positive (FP) who does not have DS. To identify 4 individuals with DS using our models, it would be necessary to screen a total of 5 people. Therefore, by using our models and screening only 5 people, we achieve the same outcome as screening 40 people randomly, representing a reduction of 87.5% in screening interviews.

Figure 2. Example of a ranked list provided by the models. Class=comparing predictions with observed values. DS: depressive symptomatology; FP: false positive; P(DS): probability of having depressive symptomatology; TN: true negative; TP: true positive.

ID	Rank	DS	P(DS)	Class
34	1	Yes	0.97	TP
10	2	Yes	0.86	TP
31	3	No	0.71	FP
27	4	Yes	0.68	TP
15	5	Yes	0.53	TP
38	6	No	0.49	TN
40	7	No	0.48	TN
16	8	No	0.44	TN
21	9	No	0.43	TN
22	10	No	0.38	TN

We can measure the reduction in number of screening interviews. This value will provide us with a clear understanding of how much the model can save compared with the random screening approach and is illustrated in equation 1:

Where R is the reduction in the number of screening interviews in %, TP is the number of TP cases, FP is the number of FP cases, and DSi is the prevalence rate of DS.

Another crucial aspect of the developed models is their capability to simulate the relationship between sensitivity and the reduction in screening efforts. For example, suppose a practitioner believes that the achieved sensitivity using the default threshold of 0.5 is insufficient, it is possible to fine-tune the sensitivity by selecting a new threshold. By doing so, the models can calculate the corresponding reduction in screening interviews associated with the adjusted sensitivity value.

Results

Data Preparation

After preprocessing the training data as discussed in the previous section, we were left with 12 out of 33 features and 3854 out of 3961 observations for the PROACTIVE data set. For PNS 2013, we were left with 218 features and 77,376 observations out of the initial 1000 and 155,432, respectively. Similarly, for PNS 2019, we had 254 out of 1087 features and 111,548 out of 205,434 observations.

Despite PNS 2019 having an additional 87 features compared with PNS 2013, the features present in PNS 2013 are also included in PNS 2019. The divergence lies in the 2019 survey, wherein certain questions were broken down into multiple components. On the other hand, PROACTIVE incorporates distinct features from both PNS 2013 and 2019 data sets.

Feature Selection

We built the BN from data using 1000 bootstrapped samples. We then performed the additional independence test on all the edges as described in the *Feature Selection* section. For the PROACTIVE data set, the connection from HYPERTENSION to AGE was dropped ($P \geq .05$). For PNS 2013 and PNS 2019, no edge had to be removed. The BNs for PROACTIVE, PNS 2013, and PNS 2019, with 12, 19, and 29 features, respectively, are displayed in Figures S1-S3 in Multimedia Appendix 1.

For the PROACTIVE data set, the MB of the outcome node consisted of postural balance problems, shortness of breath, and how old people feel they are. In the PNS 2013 data set, the MB nodes were related to the ability to do usual activities, chest pain, chronic back problems, and sleep problems. Finally, in the PNS 2019 data set, the MB nodes were related to the ability to do usual activities, chest pain, verbal abuse, and sleep problems. Detailed descriptions of these features can be found in Tables S4, S5, and S6 of Multimedia Appendix 1 for PROACTIVE, PNS 2013, and PNS 2019, respectively.

BN Parameter Learning

For each data set, we analyzed the probability of having DS given their MB nodes. As detailed in Table 1 for PROACTIVE, the features of postural balance and shortness of breath each have 2 levels (“Yes” or “No”), and the feature of how old people feel they are has 4 levels, with the minimum being up to 50 years and the maximum being >70 years. In Table 1, we present the probabilities of DS for those not having a postural balance problem, not having a shortness of breath issue, and feeling up to 50 years old, against having both issues and feeling more than 70 years old. All individuals in this data set are aged ≥60 years.

Table 1. Probability of having depressive symptomatology in PROACTIVE for 2 example scenarios.

Postural balance[a]	How old people feel they are?[b]	Shortness of breath[c]	P(DS)[d]
No	Up to 50 years	No	0.08
Yes	>70 years	Yes	0.75

[a]Do you have problems with postural balance?

[b]In general (or most of the time), how old do you feel?

[c]Have you ever experienced shortness of breath while walking, climbing stairs, or with changes in temperature (eg, when it is hot or cold)?

[d]P(DS): probability of having depressive symptomatology.

For the PNS 2013 and PNS 2019 data sets, the MB features are binary, with 2 possible values: “Yes” or “No.” Thus, Tables 2 and 3 present 2 distinct scenarios, one where all features have a “Yes” value, and the other where all features have a “No” value.

Table 2. Probability of having depressive symptomatology in Pesquisa Nacional de Saúde 2013 for 2 example scenarios.

Ability to do usual activities[a]	Chest pain[b]	Chronic back problems[c]	Sleep problems[d]	P(DS)[e]
Yes	Yes	Yes	Yes	0.66
No	No	No	No	0.03

[a]In the past 2 weeks, have you been unable to perform any of your usual activities (such as work, school, playing, and household chores) due to health reasons?

[b]Do you feel chest pain or chest discomfort when walking on a hill, going up one flight of stairs, or fast walking?

[c]Do you have any chronic spinal problems such as chronic back or neck pain, lumbago, sciatica, vertebral, or disc problems?

[d]In the past 2 weeks, have you used any medication to help you sleep?

[e]P(DS): probability of having depressive symptomatology.

Table 3. Probability of having depressive symptomatology in Pesquisa Nacional de Saúde 2019 for 2 example scenarios.

Ability to do usual activities[a]	Chest pain[b]	Verbal abuse[c]	Sleep problems[d]	P(DS)[e]
Yes	Yes	Yes	Yes	0.79
No	No	No	No	0.04

[a]In the past 2 weeks, have you been unable to perform any of your usual activities (such as work, school, playing, and household chores) due to health reasons?

[b]Do you feel chest pain or chest discomfort when walking on a hill, going up one flight of stairs, or fast walking?

[c]In the past 12 months, has anyone yelled or cursed you?

[d]In the past 2 weeks, have you used any medication to help you sleep?

[e]P(DS): Probability of having depressive symptomatology.

The results presented in Table 1 for the PROACTIVE data set indicate that individuals who report postural balance problems, episodes of shortness of breath, and feeling >70 years old have a higher probability of having DS. Specifically, if a person reports all 3 of these symptoms, the probability of having DS is 0.75, which is over 8 times higher than if the person reports feeling <50 years old and having no issues with postural balance or being out of breath.

In both PNS 2013 and PNS 2019, 3 features remained consistent over the 6-year period in both data sets (ability to do usual activities, chest pain, and sleep problems). The only difference is that in PNS 2013, feature chronic back problems as shown in Table 2 is replaced by those related to verbal abuse as illustrated in Table 3. The results reveal that answering “Yes” to all 4 questions results in a probability of having DS of 0.66 for PNS 2013 and 0.79 for PNS 2019. On the other hand, if a person answers “No” to all 4 questions, the probability of having DS is only 0.03 and 0.04, respectively.

Training of Prediction Models

The results obtained from the training process of our models are presented in Tables 4-6 for the 3 data sets. These tables provide the sensitivity and specificity determined by the thresholds of 0.444, 0.412, and 0.472 for PROACTIVE, PNS 2013, and PNS 2019, respectively. Furthermore, they present the AUC-ROC metrics for two scenarios: (1) using All-path features of the BN and (2) using the MB of the outcome.

Each of the 3 tables includes sensitivity, specificity, and AUC-ROC values for 5 folds, which were obtained using a 5-fold cross-validation approach with the best set of hyperparameters, as described in the *Training of Prediction Models* section. These metrics specifically correspond to the validation sets used during the cross-validation process.

Upon analyzing the results across all 3 data sets, we observed consistent outcomes. The mean values of sensitivity, specificity, and AUC-ROC for each fold closely align with the individual values, and the SD is low.

In PROACTIVE (Table 4), the optimal parameters were determined to be α=.200, loss function=modified_huber, and penalty term=L2 for the All-path-features scenario. The MB scenario required a different set of parameters, specifically α=.0005, loss function=log, and penalty term=elastic net.

For PNS 2013 (Table 5), the optimal hyperparameters for model tuning were found to be an α value of .003, hinge loss function, and L2 penalty term for the all-features scenario. The MB scenario yielded an α value of .006, log loss function, and L1 penalty term.

The best parameters for the PNS 2019 (Table 6) model were α=.003, loss function=log, and penalty term=elastic net. The parameters for the MB scenario were α=.083, loss function=log, and penalty term=L2.

Table 4. PROACTIVE—sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC) values in 5-fold cross-validation.

Metrics	F1[a]	F2[b]	F3[c]	F4[d]	F5[e]	Values, mean (SD)
All-path-features						
Sensitivity	0.813	0.751	0.764	0.764	0.748	0.768 (0.024)
Specificity	0.640	0.598	0.634	0.642	0.647	0.632 (0.018)
AUC-ROC	0.795	0.723	0.767	0.766	0.765	0.763 (0.026)
Markov blanket						
Sensitivity	0.678	0.673	0.637	0.699	0.621	0.662 (0.029)
Specificity	0.725	0.689	0.748	0.691	0.761	0.723 (0.029)
AUC-ROC	0.771	0.713	0.758	0.758	0.759	0.751 (0.022)

[a]F1: fold 1 of cross-validation.
[b]F2: fold 2 of cross-validation.
[c]F3: fold 3 of cross-validation.
[d]F4: fold 4 of cross-validation.
[e]F5: fold 5 of cross-validation.

Table 5. Pesquisa Nacional de Saúde 2013—sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC) values in 5-fold cross-validation.

Metrics	F1[a]	F2[b]	F3[c]	F4[d]	F5[e]	Values, mean (SD)
All-path-features						
Sensitivity	0.731	0.738	0.743	0.733	0.736	0.736 (0.004)
Specificity	0.744	0.732	0.738	0.728	0.736	0.735 (0.005)
AUC-ROC	0.807	0.805	0.809	0.798	0.805	0.805 (0.004)
Markov blanket						
Sensitivity	0.712	0.726	0.722	0.717	0.720	0.756 (0.003)
Specificity	0.716	0.703	0.714	0.697	0.709	0.708 (0.007)
AUC-ROC	0.758	0.758	0.759	0.750	0.756	0.756 (0.004)

[a]F1: fold 1 of cross-validation.
[b]F2: fold 2 of cross-validation.
[c]F3: fold 3 of cross-validation.
[d]F4: fold 4 of cross-validation.
[e]F5: fold 5 of cross-validation.

Table 6. Pesquisa Nacional de Saúde 2019—sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC) values in 5-fold cross-validation.

Metrics	F1[a]	F2[b]	F3[c]	F4[d]	F5[e]	Values, mean (SD)
All-path-features						
Sensitivity	0.704	0.712	0.703	0.710	0.717	0.709 (0.005)
Specificity	0.763	0.760	0.759	0.757	0.761	0.762 (0.002)
AUC-ROC	0.80	0.804	0.80	0.804	0.810	0.804 (0.004)
Markov blanket						
Sensitivity	0.744	0.751	0.647	0.750	0.655	0.709 (0.047)
Specificity	0.707	0.704	0.773	0.710	0.771	0.733 (0.032)
AUC-ROC	0.765	0.767	0.766	0.769	0.772	0.768 (0.003)

[a]F1: fold 1 of cross-validation.

[b]F2: fold 2 of cross-validation.

[c]F3: fold 3 of cross-validation.

[d]F4: fold 4 of cross-validation.

[e]F5: fold 5 of cross-validation.

Evaluation of the Models

The evaluation results of the models on the test data are presented in Table 7. In addition, we provide the metrics obtained from analyzing the entire training data and the mean of the 5-fold cross-validation. It is important to note that the whole data evaluation approach differs from the 5-fold cross-validation method. Instead of comparing the average metrics across the 5 folds in the validation set as if they were separate models, we assess the overall performance on the complete training data set.

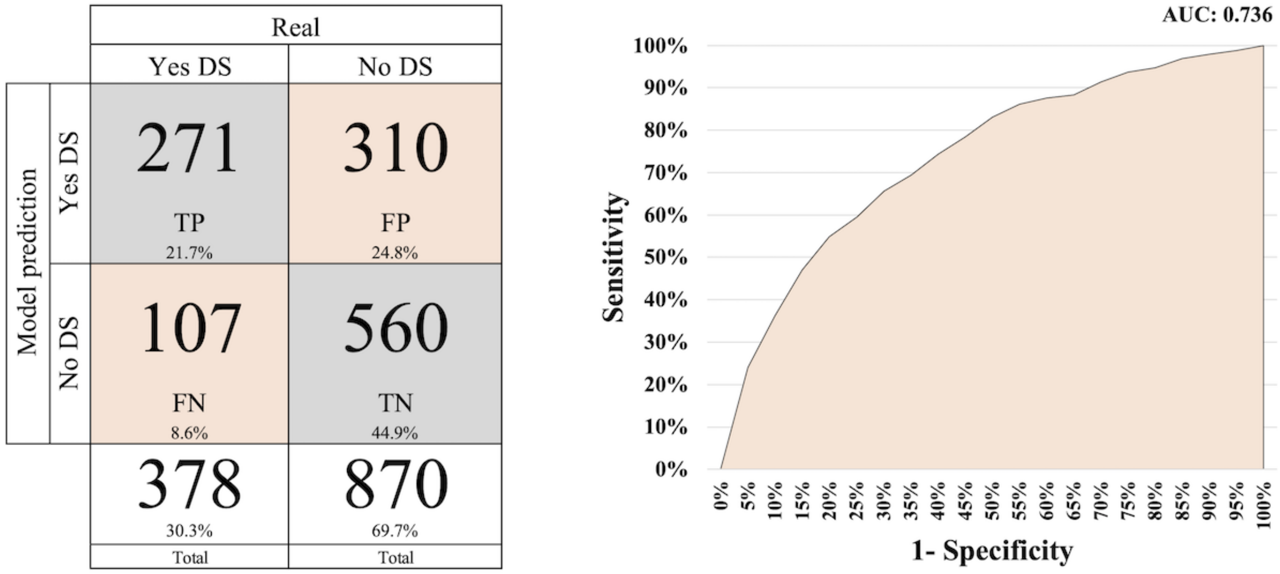
Table 7. Sensitivity, specificity, and area under the receiver operating characteristic curve on test data using All-path-features.

	PROACTIVE	Pesquisa Nacional de Saúde 2013	Pesquisa Nacional de Saúde 2019
Sensitivity			
Training	0.766	0.735	0.708
Cross-validation	0.768	0.736	0.709
Test	0.717	0.741	0.718
Specificity			
Training	0.633	0.737	0.762
Cross-validation	0.632	0.735	0.762
Test	0.644	0.737	0.766
Area under the receiver operating characteristic curve			
Training	0.765	0.806	0.804
Cross-validation	0.763	0.805	0.804
Test	0.736	0.801	0.809

The sensitivity specificity and AUC-ROC metrics obtained from the whole data evaluation, illustrated in Table 7, exhibit consistency with those obtained through the 5-fold cross-validation approach. Moreover, these metrics are not considerably different from the sensitivity, specificity, and AUC-ROC observed in the test data.

Figures 3-5 display the AUC-ROC graph and confusion matrices for PROACTIVE, PNS 2013, and PNS 2019, respectively. The confusion matrix is based on the threshold of each model applied to the test data.

Figure 3. Confusion matrix and area under the receiver operating characteristic curve for PROACTIVE. DS: depressive symptomatology; FN: false negative; FP: false positive; TN: true negative; TP: true positive.



In Figure 3, the confusion matrix reveals that out of a total of 581 cases above the threshold (TPs+FPs), 271 cases actually have DS. In addition, there are 107 cases below the threshold that have DS (false negatives). Moreover, the model for the PROACTIVE data set accurately classifies 831 cases (66.6%: TPs+true negatives) out of a total of 1248 cases.

Comparing Figures 4 and 5 with Figure 3, we observe better results. Both figures exhibit higher AUC-ROC values, with 0.801 for PNS 2013 and 0.809 for PNS 2019. In terms of correctly classifying cases, the models achieved a success rate of 73.7% for PNS 2013 and 76.1% for PNS 2019.

Figure 4. Confusion matrix and area under the receiver operating characteristic curve for Pesquisa Nacional de Saúde 2013. DS: depressive symptomatology; FN: false negative; FP: false positive; TN: true negative; TP: true positive.

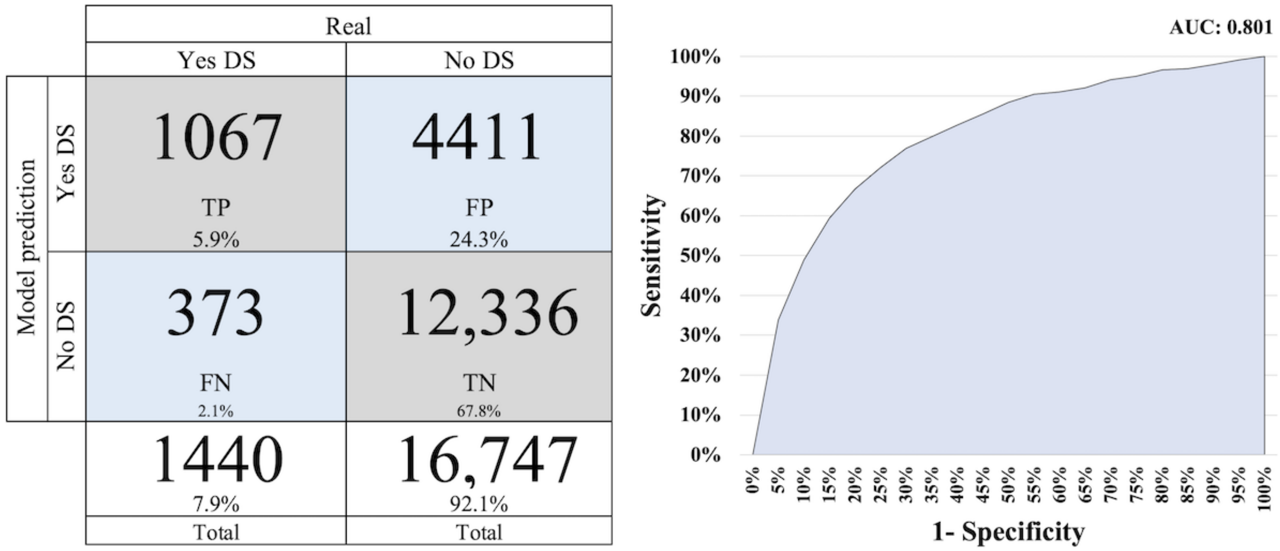
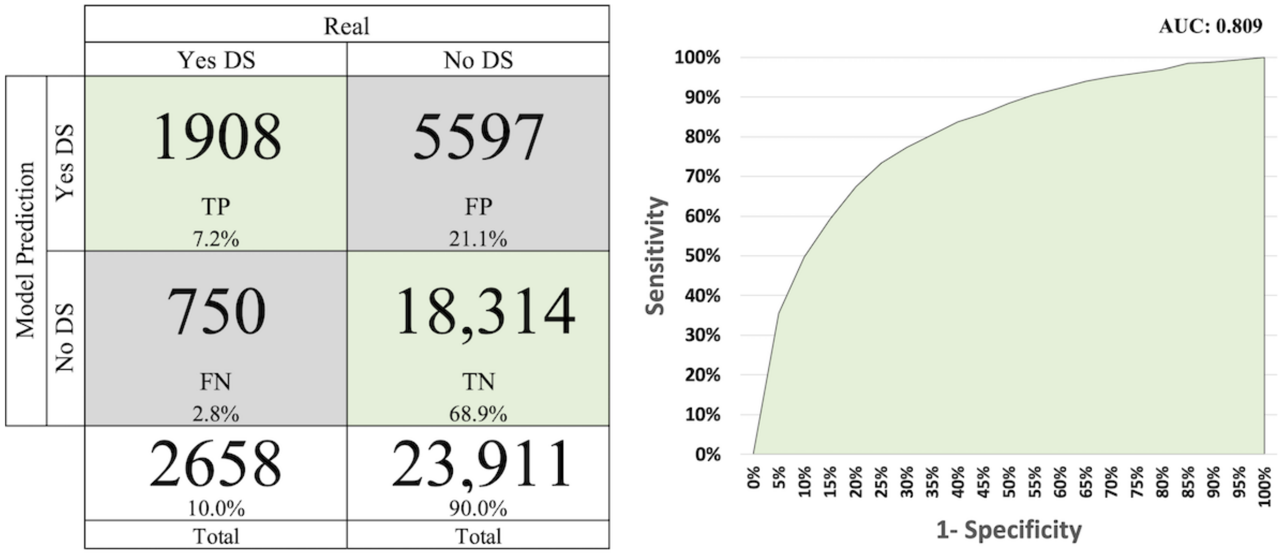


Figure 5. Confusion matrix and area under the receiver operating characteristic curve for Pesquisa Nacional de Saúde 2019. DS: depressive symptomatology; FN: false negative; FP: false positive; TN: true negative; TP: true positive.



Application of the Models

Our approach offers a high degree of flexibility in terms of adjusting the threshold to suit a desired sensitivity value. To achieve this, we can leverage the test data to simulate the trade-off between sensitivity and screening interviews reduction using equation 1 outlined in the *Methods* section.

The prevalence of DS in the PROACTIVE, PNS 2013, and PNS 2019 data sets is 30.3%, 7.9%, and 10% as illustrated in Figure 3, Figure 4, and Figure 5, respectively. The prevalence of DS in the PROACTIVE data set is more than 3 times that in the PNS data sets. Hence, it is reasonable to expect that both PNS models will exhibit superior performance with respect to screening interviews reduction in comparison to PROACTIVE. This is because when using a random screening approach, a higher prevalence of DS within a particular cohort leads to more efficient identification of affected individuals, as opposed to a cohort with a lower prevalence.

Figures 6-8 illustrate the receiver operating characteristic curves, where the bars on the x-axis represent the reduction in screening interviews as a percentage. These graphs demonstrate the trade-off between the sensitivity and the reduction in screening interviews achieved by using our models.

As an example, according to Table 7, the calculated threshold in the PROACTIVE data set corresponds to a sensitivity of 0.717. In Figure 6, this sensitivity reduces screening interviews by 35%. Furthermore, even when aiming for higher sensitivities, such as 78% or 92%, using the model developed for this data set can still result in reducing interviews by 29% or 17%, respectively.

The key finding is that for any sensitivity above 0.640, using this model consistently outperforms a normal random approach (the screening interviews reduction is higher than 1-sensitivity). This holds even with the high prevalence of DS in this cohort, as demonstrated by Figure 6.

In the case of both PNS models, it is apparent from Figures 7 and 8 that they outperform the PROACTIVE model in Figure 6. Although both models exhibit considerable improvements, the PNS 2013 model demonstrates a slightly better performance than the PNS 2019 model. Specifically, the results indicate that at a sensitivity of 88%, the PNS 2013 model achieves a 40% reduction in screening interviews, while the PNS 2019 model achieves a 39% reduction. The reduction at the same sensitivity level of 88% achieved by the PROACTIVE model is 22%. Furthermore, for both PNS data sets, using our models is always superior to using a random approach.

Figure 6. PROACTIVE—trade-off between sensitivity versus reducing screening interviews. ROC: receiver operating characteristic.

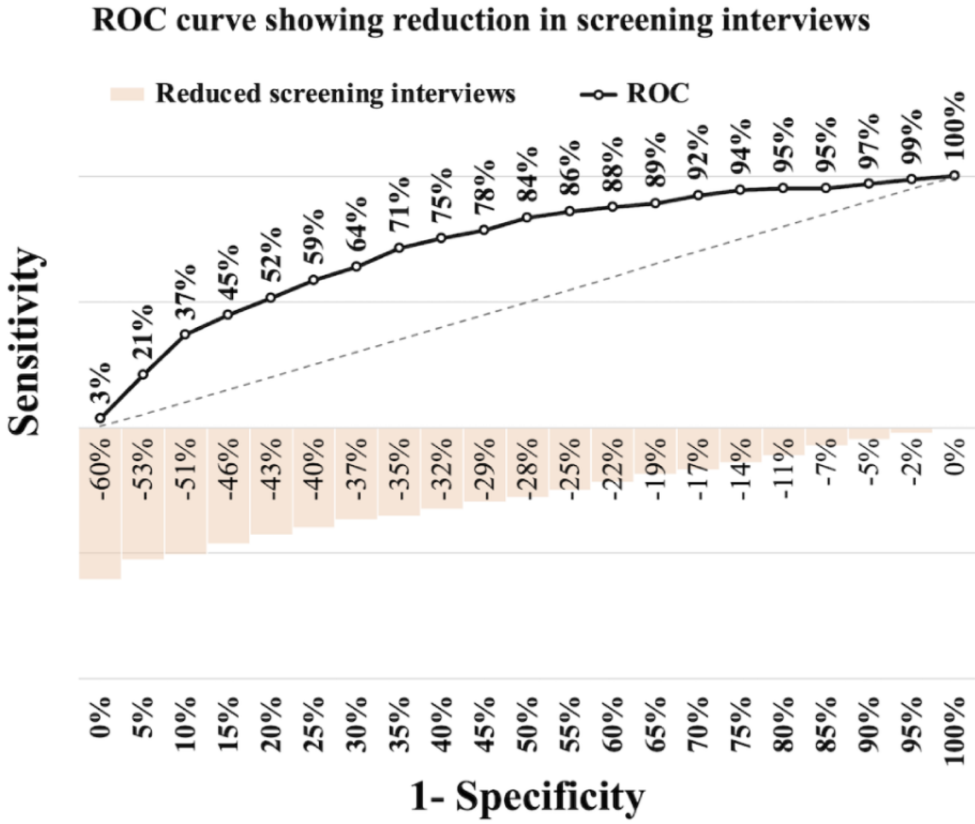


Figure 7. Pesquisa Nacional de Saúde 2013—trade-off between sensitivity versus reducing screening interviews. ROC: receiver operating characteristic.

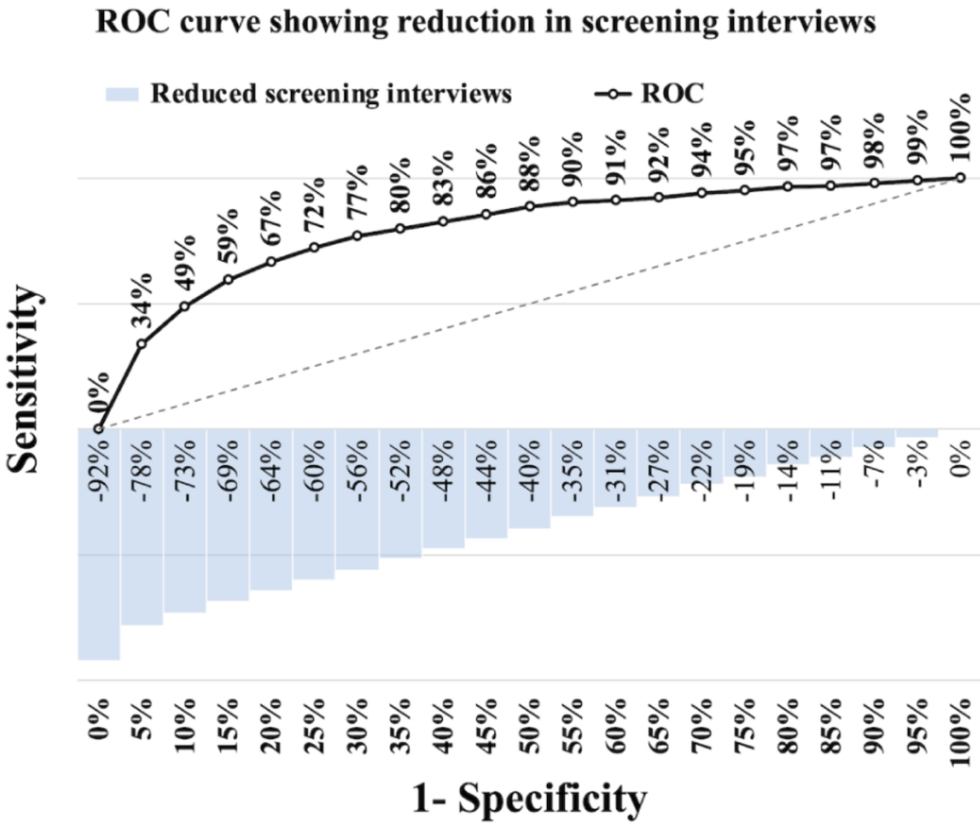
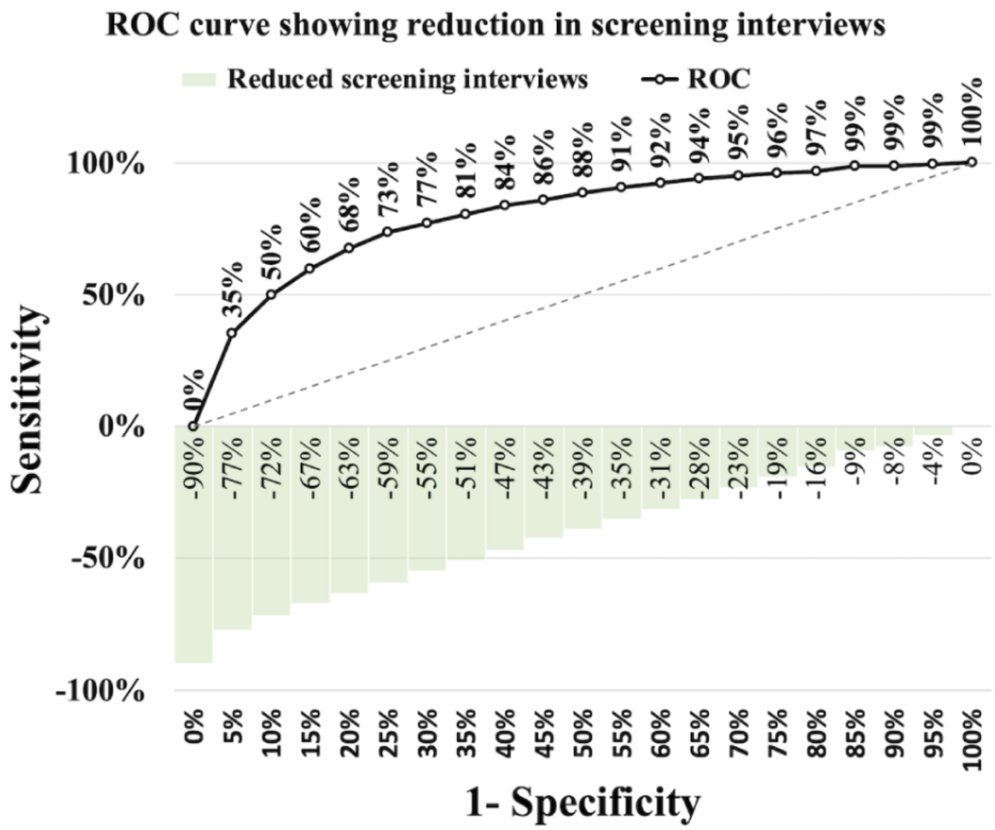


Figure 8. Pesquisa Nacional de Saúde 2019—trade-off between sensitivity versus reducing screening interviews. ROC: receiver operating characteristic.



Discussion

Overview

To enhance the development of our models, we conducted a comparison between 2 scenarios regarding feature selection. In the first scenario, we used All-path-features nodes from our BN, while in the second scenario, we focused on using only the MB of the outcome. By analyzing and contrasting these 2 scenarios, we aim to gain a comprehensive understanding of the impact that including or excluding these specific nodes has on our model.

The results of our analysis consistently showed that using All-path-features from the BN yielded superior outcomes compared with using only the MB of the outcome. Furthermore, we used the test data in the All-path-features scenario to evaluate the performance of the models for each data set. This approach ensures that the model's performance is assessed in a fair manner using previously unseen data, reflecting its real-world performance.

Using the optimal threshold that maximizes both sensitivity and specificity, determined by the Youden index, can be beneficial. This approach ensures more reliable and accurate diagnoses by minimizing the errors of missing a positive case (false negatives) or incorrectly diagnosing a negative case as positive (FPs).

The PNS 2013 and PNS 2019 data sets stand as valuable pillars of Brazil's national health data landscape. The models have consistently highlighted 3 features—the ability to do usual activities, chest pain, and sleep problems. This result across a 6-year span underscores these attributes' robustness as good indicators for detecting DS, thereby offering valuable insights for health care services. However, it is important to reapply the methodology as new data sets emerge to check for patterns and new insights.

Our assessment revealed that both the models trained on the PNS 2013 and PNS 2019 data sets outperformed the PROACTIVE data set model. This can be attributed to the larger sample sizes and lower prevalence rates of DS in the PNS data sets. It is worth noting that the PROACTIVE study exclusively concentrated on individuals aged ≥60 years. Larger data sets provide a more diverse representation of the population, leading to improved generalization of the model's predictions. In addition, the lower prevalence of DS in these data sets implies that a random screening approach would require screening a larger number of individuals to identify those with DS. In this context, the model's ability to target individuals with a higher probability of having DS can markedly reduce screening interviews.

Despite our emphasis on developing a methodology using general health and socioeconomic data without relying on features related to depression, our approach outperformed studies using similar nondepression-related features. For instance, a study conducted with university undergraduates in Bangladesh used machine learning models to predict depression. The predictors included basic information such as academic year and cumulative grade point average, without using the PHQ-9 or any depression-related information. The models achieved an AUC-ROC ranging from 0.694 to 0.802, sensitivity ranging from 0.32 to 0.53, and specificity ranging from 0.86 to

0.87 [63]. Another study, conducted in Japan, used machine learning to predict depressive symptoms using sociodemographic and biological metabolite information, also without relying on the PHQ-9 or questionnaires assessing depressive symptoms. Their models achieved an AUC-ROC ranging from 0.53 to 0.68 but did not report results for sensitivity and specificity [64].

Principal Findings

Our models exhibit almost the same sensitivity, specificity, and AUC-ROC values when applied to both the training and test data across all 3 data sets. This finding suggests that our approach is capable of generalizing effectively to unseen data, demonstrating its robustness and reliability.

It is worth noting that even when using only the MB, the results obtained were still satisfactory, making it a viable option for reducing the number of questions required to predict DS. Furthermore, the probabilistic measures extracted from the MB of the outcome demonstrated strong discriminatory power. This emphasizes that the features within the MB hold the utmost importance in the model across all 3 data sets.

The output of the models offers a visual representation of the relationship between sensitivities and the associated reduction in screening interviews. This gives health care providers the flexibility to adjust the desired sensitivity of the models according to their specific requirements. Such adaptability enhances the utility of the model as a proof-of-concept in clinical settings, making it particularly beneficial in environments with limited resources.

Limitations

While a PHQ-9 score of 10 or higher can serve as a useful indicator for DS, it should be interpreted in conjunction with other clinical information and within the context of an individual's unique circumstances. If available in the data sets, it would be straightforward with our methodology to predict different outcomes beyond PHQ-9 scores. However, the methodology may require some adjustments in the algorithm.

While our proposed methodology demonstrated good performance in all 3 data sets, it is important to note that there is a possibility of the model not performing well on certain data sets. This could happen if the features in the data set do not have a strong correlation with the outcome or if the prevalence of DS in a specific cohort is high.

The consistent prominence of the same 3 essential features in both PNS 2013 and PNS 2019 is noteworthy, but this information holds relevance primarily within Brazilian primary care settings. Across diverse cultures, different features might emerge as the most important indicators. Furthermore, it is important to consider the potential for these features to dynamically shift in significance over time, depending on the readily available data.

We acknowledge that having access to the readily available data may require obtaining permissions. Despite the data being accessible, various legal and ethical considerations may necessitate obtaining explicit permissions or approvals before its use.

Although the use of BN for feature selection yielded consistent results, it is worth noting that constructing the model using bootstrap can be very hardware demanding. In addition, the feasibility of this approach depends on the size of the data set in terms of the number of features, the range of values of the features, and the number of observations.

Conclusions

We presented a data-driven proof-of-concept methodology for identifying individuals with DS using alternative features beyond common mental health screening questionnaires. This approach was tested on 3 distinct data sets, yielding consistent results that indicate a strong generalization of the methodology for targeting individuals who could benefit from further screening.

Using BNs, we were able to identify the most influential features (ie, the MB of the outcome) and extract insights using probabilistic measures through parameter learning. Our analysis revealed that for the PROACTIVE data set, a study that screened Brazilians aged ≥60 years, the most influential features were related to postural balance, shortness of breath, and how old people feel they are. For the PNS data sets, which screened a national sample of Brazilians aged ≥18 years, the influential features of the 2013 data set were related to the ability to do usual activities, chest pain, chronic back problems, and sleep problems. In the PNS 2019 data set, the same features were found to be important as in PNS 2013, except for chronic back problems, which were replaced by a feature related to verbal abuse.

Finally, it has been demonstrated through empirical analysis that using the proposed models results in a considerable reduction of the screening interviews (up to 52%) while maintaining a sensitivity of 80%.

Acknowledgments

This publication has emanated from research supported in part by a grant from Science Foundation Ireland (grant 18/CRT/6049l; author EM). The authors would like to acknowledge funding from the São Paulo Research Foundation (FAPESP 2018/19343-9 and 2022/05107-7; CAN) and CNPq-Brazil (307579/2019-0; MS). The PROACTIVE study was funded by the São Paulo Research Foundation (FAPESP 2017/50094-2) and the Joint Global Health Trials initiative jointly funded by the Department of Health and Social Care, the Foreign, Commonwealth & Development Office, the Medical Research Council, and Wellcome (MRC MR/R006229/1).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary tables and figures.
[DOCX File , 367 KB - mental_v11i1e52045_app1.docx]

References

1. Mitchell AJ, Vaze A, Rao S. Clinical diagnosis of depression in primary care: a meta-analysis. Lancet 2009 Aug 22;374(9690):609-619. [doi: 10.1016/S0140-6736(09)60879-5] [Medline: 19640579]
2. Lopes CS, Hellwig N, E Silva GA, Menezes PR. Inequities in access to depression treatment: results of the Brazilian National Health Survey - PNS. Int J Equity Health 2016 Nov 17;15(1):154 [FREE Full text] [doi: 10.1186/s12939-016-0446-1] [Medline: 27852278]
3. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
4. Squires M, Tao X, Elangovan S, Gururajan R, Zhou X, Acharya UR, et al. Deep learning and machine learning in psychiatry: a survey of current progress in depression detection, diagnosis and treatment. Brain Inform 2023 Apr 24;10(1):10 [FREE Full text] [doi: 10.1186/s40708-023-00188-6] [Medline: 37093301]
5. Zavlis O, Matheou A, Bentall R. Identifying the bridge between depression and mania: a machine learning and network approach to bipolar disorder. Bipolar Disord 2023 Nov;25(7):571-582 [FREE Full text] [doi: 10.1111/bdi.13316] [Medline: 36869637]
6. Xie X, Samaei A, Guo J, Liu WK, Gan Z. Data-driven discovery of dimensionless numbers and governing laws from scarce measurements. Nat Commun 2022 Dec 08;13(1):7562 [FREE Full text] [doi: 10.1038/s41467-022-35084-w] [Medline: 36476735]
7. Sadeque F, Xu D, Bethard S. UArizona at the CLEF eRisk 2017 pilot task: linear and recurrent models for early depression detection. CEUR Workshop Proc 2017 Sep;1866 [FREE Full text] [Medline: 29075167]
8. Zulfiker MS, Kabir N, Biswas AA, Nazneen T, Uddin MS. An in-depth analysis of machine learning approaches to predict depression. Curr Res Behav Sci 2021 Nov;2:100044. [doi: 10.1016/j.crbeha.2021.100044]
9. Gomes SR, von Schantz M, Leocadio-Miguel M. Predicting depressive symptoms in middle-aged and elderly adults using sleep data and clinical health markers: a machine learning approach. Sleep Med 2023 Feb;102:123-131. [doi: 10.1016/j.sleep.2023.01.002] [Medline: 36641929]
10. Grzenda A, Speier W, Siddarth P, Pant A, Krause-Sorio B, Narr K, et al. Machine learning prediction of treatment outcome in late-life depression. Front Psychiatry 2021 Oct 20;12:738494 [FREE Full text] [doi: 10.3389/fpsyt.2021.738494] [Medline: 34744829]
11. Razavi R, Gharipour A, Gharipour M. Depression screening using mobile phone usage metadata: a machine learning approach. J Am Med Inform Assoc 2020 Apr 01;27(4):522-530 [FREE Full text] [doi: 10.1093/jamia/ocz221] [Medline: 31977041]
12. Nemesure MD, Heinz MV, Huang R, Jacobson NC. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. Sci Rep 2021 Jan 21;11(1):1980 [FREE Full text] [doi: 10.1038/s41598-021-81368-4] [Medline: 33479383]
13. Wang F, Cheung CW, Wong SS. Use of pain-related gene features to predict depression by support vector machine model in patients with fibromyalgia. Front Genet 2023 Mar 29;14:1026672 [FREE Full text] [doi: 10.3389/fgene.2023.1026672] [Medline: 37065490]
14. Ebrahimzadeh E, Fayaz F, Rajabion L, Seraji M, Aflaki F, Hammoud A, et al. Machine learning approaches and non-linear processing of extracted components in frontal region to predict rTMS treatment response in major depressive disorder. Front Syst Neurosci 2023 Mar 9;17:919977 [FREE Full text] [doi: 10.3389/fnsys.2023.919977] [Medline: 36968455]
15. Huang Y, Zhu C, Feng Y, Ji Y, Song J, Wang K, et al. Comparison of three machine learning models to predict suicidal ideation and depression among Chinese adolescents: a cross-sectional study. J Affect Disord 2022 Dec 15;319:221-228. [doi: 10.1016/j.jad.2022.08.123] [Medline: 36122602]
16. Li Q, Dong F, Gai Q, Che K, Ma H, Zhao F, et al. Diagnosis of major depressive disorder using machine learning based on multisequence MRI neuroimaging features. J Magn Reson Imaging 2023 Nov;58(5):1420-1430. [doi: 10.1002/jmri.28650] [Medline: 36797655]
17. Scazufca M, Nakamura CA, Seward N, Moreno-Agostino D, van de Ven P, Hollingworth W, et al. A task-shared, collaborative care psychosocial intervention for improving depressive symptomatology among older adults in a socioeconomically deprived area of Brazil (PROACTIVE): a pragmatic, two-arm, parallel-group, cluster-randomised controlled trial. Lancet Healthy Longev 2022 Oct;3(10):e690-e702 [FREE Full text] [doi: 10.1016/S2666-7568(22)00194-5] [Medline: 36202129]

18. Van de Ven P, Araya R, de Paula Couto MCP, Henrique MG, Meere D, Vilela Mendes A, et al. Investigating software requirements for systems supporting task-shifted interventions: usability study. J Med Internet Res 2019 Nov 12;21(11):e11346 [FREE Full text] [doi: 10.2196/11346] [Medline: 31714246]
19. Lopes CS, Gomes NL, Junger WL, Menezes PR. Trend in the prevalence of depressive symptoms in Brazil: results from the Brazilian National Health Survey 2013 and 2019. Cad Saude Publica 2022;38Suppl 1(Suppl 1):e00123421 [FREE Full text] [doi: 10.1590/0102-311X00123421] [Medline: 35544918]
20. Edwards C, Mukherjee S, Simpson L, Palmer LJ, Almeida OP, Hillman DR. Depressive symptoms before and after treatment of obstructive sleep apnea in men and women. J Clin Sleep Med 2015 Sep 15;11(9):1029-1038 [FREE Full text] [doi: 10.5664/jcsm.5020] [Medline: 25902824]
21. Hammash MH, Hall LA, Lennie TA, Heo S, Chung ML, Lee KS, et al. Psychometrics of the PHQ-9 as a measure of depressive symptoms in patients with heart failure. Eur J Cardiovasc Nurs 2013 Oct 21;12(5):446-453. [doi: 10.1177/1474515112468068] [Medline: 23263270]
22. Whooley MA, de Jonge P, Vittinghoff E, Otte C, Moos R, Carney RM, et al. Depressive symptoms, health behaviors, and risk of cardiovascular events in patients with coronary heart disease. JAMA 2008 Nov 26;300(20):2379-2388 [FREE Full text] [doi: 10.1001/jama.2008.711] [Medline: 19033588]
23. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O. A survey on missing data in machine learning. J Big Data 2021 Oct 27;8(1):140 [FREE Full text] [doi: 10.1186/s40537-021-00516-9] [Medline: 34722113]
24. Enders CK. Missing data: an update on the state of the art. Psychol Methods 2023 Mar 16 (forthcoming). [doi: 10.1037/met0000563] [Medline: 36931827]
25. Nagula PK, Alexakis C. A novel machine learning approach for predicting the NIFTY50 index in India. Int Adv Econ Res 2022 Dec 06;28(3-4):155-170. [doi: 10.1007/s11294-022-09861-8]
26. Pearl J. From Bayesian networks to causal networks. In: Coletti G, Dubois D, Scozzafava R, editors. Mathematical Models for Handling Partial Knowledge in Artificial Intelligence. Boston, MA: Springer; 1995.
27. Inza I, Larrañaga P, Etxeberria R, Sierra B. Feature subset selection by Bayesian network-based optimization. Artif Intell 2000 Oct;123(1-2):157-184. [doi: 10.1016/S0004-3702(00)00052-7]
28. Kuschner KW, Malyarenko DI, Cooke WE, Cazares LH, Semmes OJ, Tracy ER. A Bayesian network approach to feature selection in mass spectrometry data. BMC Bioinformatics 2010 Apr 08;11(1):177 [FREE Full text] [doi: 10.1186/1471-2105-11-177] [Medline: 20377906]
29. Wang Z, Wang Z, Gu X, He S, Yan Z. Feature selection based on Bayesian network for chiller fault diagnosis from the perspective of field applications. Appl Therm Eng 2018 Jan;129:674-683. [doi: 10.1016/j.applthermaleng.2017.10.079]
30. Marcot BG, Penman TD. Advances in Bayesian network modelling: integration of modelling technologies. Environ Model Softw 2019 Jan;111:386-393. [doi: 10.1016/j.envsoft.2018.09.016]
31. Rebolledo M, Eiben AE, Bartz-Beielstein T. Bayesian networks for mood prediction using unobtrusive ecological momentary assessments. In: Proceedings of the 24th International Conference on Applications of Evolutionary Computation. 2021 Presented at: EvoApplications 2021; April 7-9, 2021; Virtual Event. [doi: 10.1007/978-3-030-72699-7_24]
32. Sihag G, Delcroix V, Grislin E, Siebert X, Piechowiak S, Puisieux F. Prediction of risk factors for fall using Bayesian networks with partial health information. In: Proceedings of the IEEE Globecom Workshops. 2020 Presented at: GC Wkshps 2020; December 7-11, 2020; Taipei, Taiwan. [doi: 10.1109/gcwkshps50303.2020.9367534]
33. Ruiz-Tagle A, Lewis AD, Schell CA, Lever E, Groth KM. BaNTERA: a Bayesian network for third-party excavation risk assessment. Reliab Eng Syst Saf 2022 Jul;223:108507. [doi: 10.1016/j.ress.2022.108507]
34. Ruskey B, Rosenberg E. Minimizing risk in Bayesian supply chain networks. Comput Ind Eng 2022 Jul;169:108134. [doi: 10.1016/j.cie.2022.108134]
35. Skaanning C, Jensen FV, Kjærulff U. Printer troubleshooting using Bayesian networks. In: Proceedings of the 13th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems. 2000 Presented at: IEA/AIE 2000; June 19-22, 2000; New Orleans, LA. [doi: 10.1007/3-540-45049-1_45]
36. Su C, Andrew A, Karagas MR, Borsuk ME. Using Bayesian networks to discover relations between genes, environment, and disease. BioData Min 2013 Mar 21;6(1):6 [FREE Full text] [doi: 10.1186/1756-0381-6-6] [Medline: 23514120]
37. Lucas P. Expert knowledge and its role in learning Bayesian networks in medicine: an appraisal. In: Proceedings of the 8th Conference on Artificial Intelligence in Medicine in Europe. 2001 Presented at: AIME 2001; July 1-4, 2001; Cascais, Portugal. [doi: 10.1007/3-540-48229-6_24]
38. Beretta S, Castelli M, Gonçalves I, Henriques R, Ramazzotti D. Learning the structure of Bayesian Networks: a quantitative assessment of the effect of different algorithmic schemes. arXiv Preprint posted online on August 3, 2018 [FREE Full text] [doi: 10.1155/2018/1591878]
39. Kitson NK, Constantinou AC. Learning Bayesian networks from demographic and health survey data. J Biomed Inform 2021 Jan;113:103588 [FREE Full text] [doi: 10.1016/j.jbi.2020.103588] [Medline: 33217542]
40. Tsamardinos I, Aliferis CF, Statnikov AR. Algorithms for large scale Markov blanket discovery. In: Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference. 2003 Presented at: FLAIRS 2003; May 12-14, 2003; St. Augustine, FL.

41. Friedman N, Goldszmidt M, Wyner A. Data analysis with Bayesian networks: a bootstrap approach. arXiv Preprint posted online on January 23, 2013 [FREE Full text]
42. Imoto S, Kim SY, Shimodaira H, Aburatani S, Tashiro K, Kuhara S, et al. Bootstrap analysis of gene networks based on Bayesian networks and nonparametric regression. Genome Inform 2002;13:369-370.
43. Myte R, Gylling B, Häggström J, Schneede J, Magne Ueland P, Hallmans G, et al. Untangling the role of one-carbon metabolism in colorectal cancer risk: a comprehensive Bayesian network analysis. Sci Rep 2017 Feb 24;7:43434 [FREE Full text] [doi: 10.1038/srep43434] [Medline: 28233834]
44. Nicholson AE, Jitnah N. Using mutual information to determine relevance in Bayesian networks. In: Proceedings of the 5th Pacific Rim International Conference on Artificial Intelligence. 1998 Presented at: PRICAI'98; November 22-27, 1998; Singapore, Singapore. [doi: 10.1007/bfb0095287]
45. Hruschka ERJ, Hruschka E, Ebecken NF. Feature selection by Bayesian networks. In: Proceedings of the 17th Conference of the Canadian Society for Computational Studies of Intelligence. 2004 Presented at: Canadian AI 2004; May 17-19, 2004; London, ON. [doi: 10.1007/978-3-540-24840-8_26]
46. Wang Y, Gao X, Ru X, Sun P, Wang J. Identification of gene signatures for COAD using feature selection and Bayesian network approaches. Sci Rep 2022 May 24;12(1):8761 [FREE Full text] [doi: 10.1038/s41598-022-12780-7] [Medline: 35610288]
47. Maekawa E, Glavin D, Grua EM, Nakamura CA, Scazufca M, Araya R, et al. A machine learning approach to optimize the assessment of depressive symptomatology. Procedia Comput Sci 2022;206:111-120 [FREE Full text] [doi: 10.1016/j.procs.2022.09.090]
48. Ji Z, Xia Q, Meng G. A review of parameter learning methods in Bayesian network. In: Proceedings of the 11th International Conference on Advanced Intelligent Computing Theories and Applications. 2015 Presented at: ICIC 2015; August 20-23, 2015; Fuzhou, China. [doi: 10.1007/978-3-319-22053-6_1]
49. Xie Z, Jiang L, Ye T, Li X. A synthetic minority oversampling method based on local densities in low-dimensional space for imbalanced learning. In: Proceedings of the 20th International Conference on Database Systems for Advanced Applications. 2015 Presented at: DASFAA 2015; April 20-23, 2015; Hanoi, Vietnam. [doi: 10.1007/978-3-319-18123-3_1]
50. Malhotra R, Kamal S. An empirical study to investigate oversampling methods for improving software defect prediction using imbalanced data. Neurocomputing 2019 May;343:120-140. [doi: 10.1016/j.neucom.2018.04.090]
51. Wei J, Huang H, Yao L, Hu Y, Fan Q, Huang D. New imbalanced bearing fault diagnosis method based on Sample-characteristic Oversampling TechniquE (SCOTE) and multi-class LS-SVM. Appl Soft Comput 2021 Mar;101:107043. [doi: 10.1016/j.asoc.2020.107043]
52. Bottou L. Large-scale machine learning with stochastic gradient descent. In: Proceedings of the 19th International Conference on Computational Statistics. 2010 Presented at: COMPSTAT'2010; August 22-27, 2010; Paris, France. [doi: 10.1007/978-3-7908-2604-3_16]
53. Hardt M, Recht B, Singer Y. Train faster, generalize better: stability of stochastic gradient descent. arXiv Preprint posted online on September 3, 2015 [FREE Full text]
54. Jha RS, Jha NN, Lele MM. Stochastic gradient descent algorithm for the predictive modelling of grate combustion and boiler dynamics. ISA Trans 2023 May;136:571-589. [doi: 10.1016/j.isatra.2022.10.036] [Medline: 36402597]
55. Quan J, Yan B, Sang X, Zhong C, Li H, Qin X, et al. Multi-depth computer-generated hologram based on stochastic gradient descent algorithm with weighted complex loss function and masked diffraction. Micromachines (Basel) 2023 Mar 06;14(3):605 [FREE Full text] [doi: 10.3390/mi14030605] [Medline: 36985013]
56. Vidyashree KP, Rajendra AB. An improvised sentiment analysis model on Twitter data using stochastic gradient descent (SGD) optimization algorithm in stochastic gate neural network (SGNN). SN Comput Sci 2023;4(2):190 [FREE Full text] [doi: 10.1007/s42979-022-01607-x] [Medline: 36748096]
57. Victoria AH, Maragatham G. Automatic tuning of hyperparameters using Bayesian optimization. Evol Syst 2020 May 25;12(1):217-223. [doi: 10.1007/s12530-020-09345-2]
58. Wang Q, Ma Y, Zhao K, Tian Y. A comprehensive survey of loss functions in machine learning. Ann Data Sci 2020 Apr 12;9(2):187-212. [doi: 10.1007/s40745-020-00253-5]
59. Mosli R, Li R, Yuan B, Pan Y. Automated malware detection using artifacts in forensic memory images. In: Proceedings of the IEEE Symposium on Technologies for Homeland Security (HST). 2016 Presented at: HST 2016; May 10-11, 2016; Waltham, MA. [doi: 10.1109/ths.2016.7568881]
60. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 1996 Jan;58(1):267-288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]
61. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 1970;12(1):55-67. [doi: 10.1080/00401706.1970.10488634]
62. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol 2005 Apr;67(2):301-320 [FREE Full text] [doi: 10.1111/j.1467-9868.2005.00503.x]
63. Choudhury AA, Khan MR, Nahim NZ, Tulon SR, Islam S, Chakrabarty A. Predicting depression in Bangladeshi undergraduates using machine learning. In: Proceedings of the IEEE Region 10 Symposium (TENSYMP). 2019 Presented at: TENSYMP 2019; June 07-09, 2019; Kolkata, India. [doi: 10.1109/tensymp46218.2019.8971369]

64. Takahashi Y, Ueki M, Yamada M, Tamiya G, Motoike IN, Saigusa D, et al. Improved metabolomic data-based prediction of depressive symptoms using nonlinear machine learning with feature selection. Transl Psychiatry 2020 May 19;10(1):157 [FREE Full text] [doi: 10.1038/s41398-020-0831-9] [Medline: 32427830]

Abbreviations

AUC-ROC: area under the receiver operating characteristic curve
BN: Bayesian network
CPD: conditional probability distribution
DS: depressive symptomatology
FP: false positive
MB: Markov blanket
PHQ-9: 9-item Patient Health Questionnaire
PNS: Pesquisa Nacional de Saúde
SGD: stochastic gradient descent
TP: true positive

Edited by J Torous; submitted 21.08.23; peer-reviewed by S Pandey, Y Liu; comments to author 04.03.24; revised version received 02.04.24; accepted 17.04.24; published 04.07.24.

Please cite as:
Maekawa E, Grua EM, Nakamura CA, Scazufca M, Araya R, Peters T, van de Ven P
Bayesian Networks for Prescreening in Depression: Algorithm Development and Validation
JMIR Ment Health 2024;11:e52045
URL: https://mental.jmir.org/2024/1/e52045
doi:10.2196/52045
PMID:

Original Paper

Engagement, Acceptability, and Effectiveness of the Self-Care and Coach-Supported Versions of the Vira Digital Behavior Change Platform Among Young Adults at Risk for Depression and Obesity: Pilot Randomized Controlled Trial

Lauren S Weiner[1], PhD; Ryann N Crowley[1,2], MS; Lisa B Sheeber[3], PhD; Frank H Koegler[4], PhD; Jon F Davis[4], PhD; Megan Wells[1], BA; Carter J Funkhouser[5,6], PhD; Randy P Auerbach[5,6,7], PhD; Nicholas B Allen[1,2], PhD

[1]Ksana Health, Eugene, OR, United States
[2]Center for Digital Mental Health, University of Oregon, Eugene, OR, United States
[3]Oregon Research Institute, Eugene, OR, United States
[4]Integrated Physiology Research, Global Drug Discovery, Novo Nordisk Research Center Seattle, Seattle, WA, United States
[5]Department of Psychiatry, Columbia University, New York, NY, United States
[6]Division of Child and Adolescent Psychiatry, New York State Psychiatric Institute, New York, NY, United States
[7]Sackler Institute for Developmental Psychobiology, Columbia University, New York, NY, United States

Corresponding Author:
Lauren S Weiner, PhD
Ksana Health
2288 Alder Street
Eugene, OR, 97405
United States
Phone: 1 541 912 2883
Email: lauren.weiner@ksanahealth.com

Abstract

Background: Adolescence and early adulthood are pivotal stages for the onset of mental health disorders and the development of health behaviors. Digital behavioral activation interventions, with or without coaching support, hold promise for addressing risk factors for both mental and physical health problems by offering scalable approaches to expand access to evidence-based mental health support.

Objective: This 2-arm pilot randomized controlled trial evaluated 2 versions of a digital behavioral health product, Vira (Ksana Health Inc), for their feasibility, acceptability, and preliminary effectiveness in improving mental health in young adults with depressive symptoms and obesity risk factors.

Methods: A total of 73 participants recruited throughout the United States were randomly assigned to use Vira either as a self-guided product (Vira Self-Care) or with support from a health coach (Vira+Coaching) for 12 weeks. The Vira smartphone app used passive sensing of behavioral data related to mental health and obesity risk factors (ie, activity, sleep, mobility, and language patterns) and offered users personalized insights into patterns of behavior associated with their daily mood. Participants completed self-reported outcome measures at baseline and follow-up (12 weeks). All study procedures were completed via digital communications.

Results: Both versions of Vira showed strong user engagement, acceptability, and evidence of effectiveness in improving mental health and stress. However, users receiving coaching exhibited more sustained engagement with the platform and reported greater reductions in depression (Cohen d=0.45, 95% CI 0.10-0.82) and anxiety (Cohen d=0.50, 95% CI 0.13-0.86) compared to self-care users. Both interventions also resulted in reduced stress (Vira+Coaching: Cohen d=–1.05, 95% CI –1.57 to –0.50; Vira Self-Care: Cohen d=–0.78, 95% CI –1.33 to –0.23) and were perceived as useful and easy to use. Coached users also reported reductions in sleep-related impairment (Cohen d=–0.51, 95% CI –1.00 to –0.01). Moreover, participants increased their motivation for and confidence in making behavioral changes, with greater improvements in confidence among coached users.

Conclusions: An app-based intervention using passive mobile sensing to track behavior and deliver personalized insights into behavior-mood associations demonstrated feasibility, acceptability, and preliminary effectiveness for reducing depressive symptoms and other mental health problems in young adults. Future directions include (1) optimizing the interventions, (2) conducting a fully powered trial that includes an active control condition, and (3) testing mediators and moderators of outcome effects.

Trial Registration: ClinicalTrials.gov NCT05638516; https://clinicaltrials.gov/study/NCT05638516

(JMIR Ment Health 2024;11:e51366) doi:10.2196/51366

KEYWORDS

depression; behavioral activation; digital health; mental health; behavior change; mobile sensing; anxiety; health coaching; mobile phone

Introduction

Background

Adolescence and early adulthood are peak periods for the emergence of major depressive disorder [1-3] and other mental health problems [4,5]. In addition, health behaviors, both positive (eg, exercise and healthy diet) and negative (eg, substance use and unsafe sex practices), often become established during adolescence and frequently persist into adulthood [6].

Mental health disorders in adolescents and young adults often co-occur with risk factors for chronic physical diseases [7] and are strongly linked to poor physical health across the life span [8-10]. Although most research on the co-occurrence of mental and physical health problems has been conducted with adults [7], some studies have suggested that the initial comorbidity between mental and physical health begins to emerge during adolescence [7,11]. Depressive disorders in adolescence have been associated with later-onset physical diseases, including arthritis, allergies, heart disease, diabetes, digestive system diseases, and epilepsy [7]. Individuals with mental health diagnoses lose 10 years from their lives, on average, due to all-cause mortality [12], and the predictive power of mental health disorders on these outcomes is similar to that of other well-known biomedical risk factors, such as smoking and obesity [10].

Moreover, mental health problems in young people often precede and predict the onset of risk factors for chronic physical health conditions [13]. This is especially true for obesity, a risk factor for a range of chronic diseases [14]. The National Longitudinal Study of Adolescent Health in the United States found that adolescents' depressive symptoms were positively associated with risk for obesity 1 year later, even among those not obese at baseline [15]. Similarly, the Coronary Artery Risk Development in Youth Adults study found that young adults aged 18 to 30 years with higher levels of self-reported depressive symptoms experienced greater increases in their BMI and waist-to-hip ratio [16]. Accordingly, adolescents and young adults experiencing depression are an ideal target group for a risk reduction approach.

Behavioral activation (BA) is an evidence-based treatment for depression [17,18]. It has also been shown to improve a range of other depression-related phenomena, such as anxiety [19], sleep [20], emotional and social support [21,22], obesity [23], and general well-being, among those without clinical diagnoses [24]. In BA, people learn techniques to monitor their mood and daily activities, observe the connection between them, and then develop a plan to increase the number of pleasant activities and positive interactions with their environment. Importantly, the types of positive behavior change targeted by BA (eg, increasing physical activity and increasing social connection) can not only improve mood but are also strong inverse correlates of risk for chronic disease [25-27]. Thus, BA is an intervention that may reduce risk for chronic disease in two ways: (1) by reducing mental health–related risk factors and (2) by instigating positive behavior changes that can directly impact risk factors for obesity and related disorders.

BA can also be easily adapted to a brief intervention format [28] and can be delivered digitally [29], increasing its potential for scalability. Scalability is important because the traditional model of delivery (eg, face-to-face treatment) cannot meet the rising mental health needs of the population [30]. A recent meta-analysis revealed that digital BA interventions, used as self-guided tools or with support from a therapist or lay coach, can improve depressive symptoms, anxiety, and other aspects of quality of life [31]. Thus, using digital methods to deliver interventions such as BA is a promising approach to expanding access to evidence-based mental health support.

One of the most pressing issues for digital interventions is establishing evidence of effectiveness, as well as user adherence, typically a prerequisite to effectiveness [32]. Self-guided or stand-alone digital mental health apps, which are fully technological in their delivery, show great promise in terms of their scalability [33]. However, some meta-analyses have raised questions about whether self-guided apps are effective in improving mental health [34,35]. Other reviews have concluded that interventions incorporating human support (eg, coaching, manual adherence prompts, or integration with professional face-to-face care) yield greater effectiveness [36,37] and adherence [38,39] than self-guided apps. If smartphone apps need human support to be effective, this resource limitation must be factored into the implementation of digital interventions. By contrast, if a self-guided app can show high adherence and effectiveness without human support, then the scalability of the approach is significantly enhanced.

One form of human support particularly relevant to community-based interventions is coaching that is designed to bolster the use of and engagement with the app [32]. With appropriate training, this type of coaching can be delivered successfully by nonspecialists, scheduled or on demand, and across various modalities (eg, phone, text, and social media) [40]. Previous studies have found that coaching can contribute to the effectiveness of digital mental health interventions across many use cases [41-43] and that users are interested in coaching support when using an app [44,45].

Goals of This Study

The purpose of this 2-arm pilot randomized controlled trial was to assess the feasibility, acceptability, and preliminary effectiveness of 2 versions of a digital BA product (Vira, Ksana Health Inc) in young adults with depressive symptoms and obesity risk factors. Users were randomly assigned to use Vira as a self-guided product (Vira Self-Care) or Vira with support from a trained health coach (Vira+Coaching) for 12 weeks. The Vira smartphone app includes functions that leverage the passive sensing of behavioral data related to mental health (ie, data on activity, sleep, mobility, and language patterns) and provides feedback in the form of personalized insights about patterns of behavior associated with their daily mood. These functions target positive behavior changes that can both reduce mental health symptoms (including depression and anhedonia) and directly impact risk factors for obesity and related disorders (eg, increased physical activity, improved sleep quality, increased mobility, and improved mental health [46-49]). The primary aims were to examine the following: (1) feasibility of and engagement with the platform, defined as objectively measured daily active use and retention, as well as self-reported use of Vira features; (2) acceptability of the platform, operationalized as ratings on items adapted from the Technology Acceptance Model (TAM) [50] and user feedback; and (3) preliminary evidence of clinical effectiveness, assessed through change in depressive symptoms from baseline to 3 months within and between study groups. Secondary aims were to evaluate the effectiveness of the 2 interventions on additional, self-reported aspects of mood and health (ie, anxiety, perceived stress, sleep-related impairment, sleep disturbance, and emotional support) by exploring preintervention to postintervention changes in these outcomes within and between study groups.

Methods

Study Design and Overview

This fully internet–based, 2-arm, pilot randomized controlled trial compared the feasibility, acceptability, and preliminary effectiveness of 2 versions of the Vira behavior change platform: a self-guided version (Vira Self-Care) versus a version with support from a health coach (Vira+Coaching). Participants were young adults with elevated depressive symptoms and obesity risk factors. The intervention period was 12 weeks long and was preceded and followed by web-based baseline and follow-up assessments, respectively. Randomization occurred after baseline assessment. Data were collected from November 2022 through May 2023. The trial was prospectively registered on ClinicalTrials.gov (NCT05638516).

Ethical Considerations

All procedures were reviewed and approved by the Oregon Research Institute Institutional Review Board (Optimizing Vira). All coaches and study staff were trained in the ethical treatment of human participants in research. All participants received detailed information about the study and provided informed consent. Participants were compensated for assessment completion (US $30 gift card at baseline; US $40 at follow-up), regardless of app installation or use. Data were deidentified for analysis.

Eligibility

Participants were eligible if they (1) were aged 18 to 25 years, (2) resided in the United States, (3) demonstrated English fluency and literacy, (3) had access to an Android (Google LLC) or iOS (Apple, Inc) smartphone, (4) reported elevated depressive symptoms (8-item Patient Health Questionnaire [PHQ-8] [51] score≥10), and (5) were overweight (BMI≥25 kg/m^2) or reported a parental history of overweight or obesity. Participants were ineligible if they previously participated in the study.

Procedures

Recruitment, Screening, and Informed Consent

Recruitment was conducted across the United States between November 2022 and January 2023, led by a market research firm, KJT Group Inc. The study was advertised to members of KJT's existing databases and consumer research panels who have opted in to participate in the research. Research panels are advantageous in providing access to large population pools, enabling researchers to reach particular segments of the population and thus control sample composition [52]. In addition to eligibility requirements, recruitment was targeted at a nationally representative sample with regard to demographic characteristics.

Interested individuals completed a web-based screening questionnaire to assess eligibility. The screening questionnaire consisted of questions on demographic characteristics, self-reported height, weight, parental history of being overweight, and depressive symptoms (PHQ-8). Eligible individuals were presented with a web-based consent form and were contacted by a KJT staff member via phone to verify their eligibility and address questions related to the consent or study procedures. Contact information for eligible, verified, and consented individuals was sent to Ksana Health via a secure file sharing platform. No formal power calculations were conducted to determine the sample size for this pilot study.

Randomization and Onboarding

After completing the baseline assessment, participants were randomly allocated to either Vira Self-Care or Vira+Coaching. Participants were randomly allocated using a random number generation function (RANDBETWEEN) in Excel (Microsoft Corp). The random allocation sequence was generated by a biostatistician not involved in study enrollment or measurement. The biostatistician and project director assigned participants to conditions using the random allocation sequence. Assignment was completed in blocks of eligible participants. Upon assignment, each eligible participant was assigned a value from the random number generator, and allocation was determined based on gender and the random number. Per block, condition assignment was balanced by gender (equal; when possible, +1 or –1 was tracked across batches). Following randomization, participants in both conditions were emailed brief written directions on how to install and use the Vira app and subsequently completed a 5-minute phone call with the study coordinator to verify installation.

Intervention Conditions

Participants in both conditions had access to the Vira app for 12 weeks. During this period, they were prompted by the app to complete a single daily mood question with a 5-point rating scale: rate your enjoyment during the previous day, ranging from *not at all* to *super enjoyable!* They were also prompted to complete a daily, 3-item food questionnaire (results not presented here).

Vira Self-Care

Participants in the Vira Self-Care condition used the Vira app as a self-guided intervention. The Vira app, installed on the participant's smartphone, passively collected data indicative of risk-relevant behavioral patterns and psychological states from phone sensors (ie, measures of physical activity, sleep patterns, mobility, and language patterns reflecting mood states and cognition). Upon installation, participants were asked to grant the app permission to collect various data, which were then processed on their device using validated algorithms and displayed to the user on their *Today* (home) screen (Figure 1). The passive data streams collected by Vira have been shown to be related to changes in mental health [53-55]. The Vira app also prompted users to reflect on their sense of enjoyment over the previous day through one daily question. During an initial 10-day period, the app assessed the relationship between the user's patterns of behavior and their day-to-day variations in mood and well-being using passive mobile sensing. After this period, the app continued to collect data and offered the user up to 2 personalized insights per week about behavioral factors (eg, sleep, movement, location, or communication patterns) that may relate to their mood. For example, Vira could suggest that "you seem to enjoy the days when you drive more" or "you tend to enjoy the day more when you wake up earlier." The Vira app guided users to reflect on these insights and create a behavior change plan by exploring more detailed information about the insight (ie, insight details) and accessing in-app and external resources, such as links to articles about health behaviors and behavior change. The user experience for Vira insights is illustrated in Figure 2. In line with BA, pursuing activities that brought participants meaning or joy could break the cycle of thoughts and behaviors associated with poor mental health and improve their overall energy, mood, and quality of life [56]. Participants could also view their passive sensing data for the previous day (eg, bed and wake time, time away from home, time in transit, and movement time) and see a graphical display of their check-in ratings for the previous week (ie, weekly mood graph) within the Vira app.

Figure 1. Weekly averages of the user's passive sensing data and their most recent Vira insights are displayed on the Today screen of the Vira app.

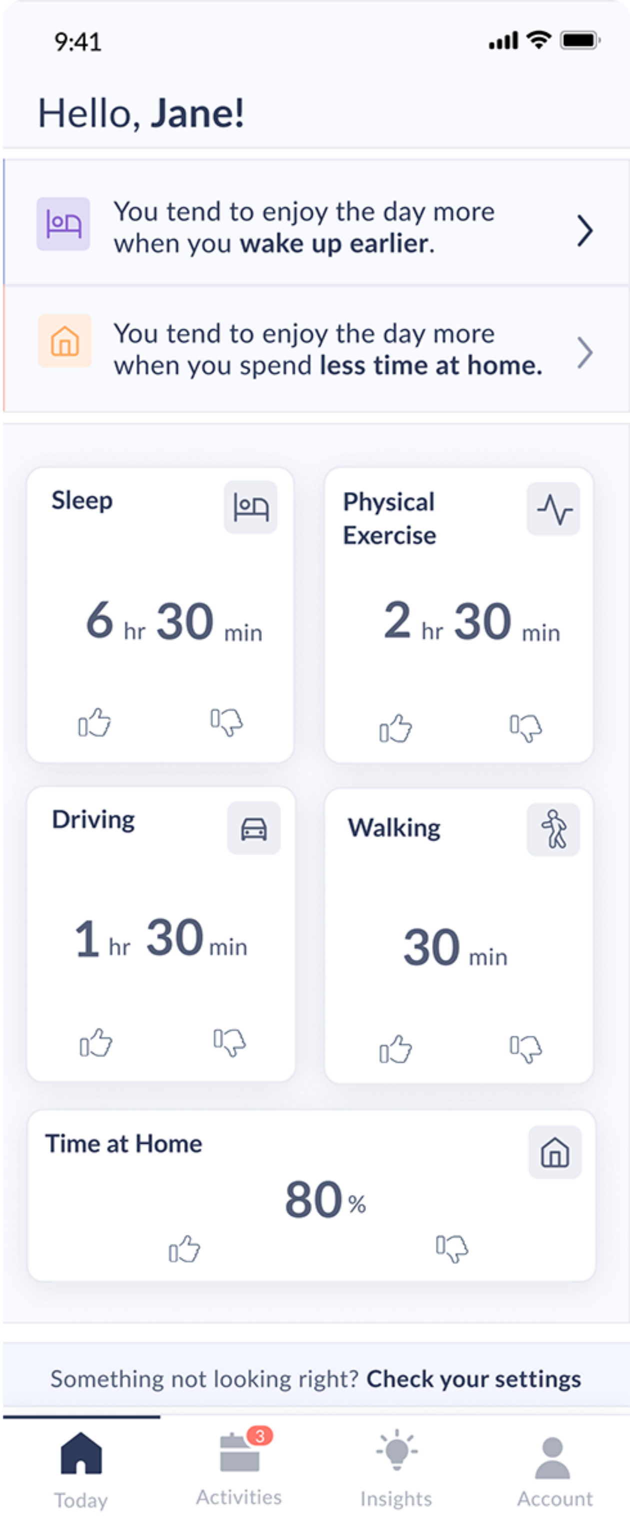


Figure 2. User experience flow of the Vira insights feature.

Vira+Coaching

Participants in the Vira+Coaching condition used the same version of the Vira app as participants in the Vira Self-Care arm and additionally received text-based support and nudges within the Vira app related to their behavior change goals from a trained health coach. Table 1 lists differences between the coaching and self-guided versions of the Vira app experience. Coaches were trained in motivational interviewing techniques and followed a manualized approach to coaching. A total of 5 health coaches with diverse educational and professional backgrounds and a minimum of a bachelor's degree (in any discipline) participated in the study. Of the 5 coaches, 2 (40%) had previous experience delivering behavioral interventions, but none were professional behavioral health providers (ie, none had degrees or training in a health profession, such as clinical psychology, psychiatry, or social work).

Table 1. Features of the coaching and self-care Vira app experiences.

Feature	Self-care version	Coached version
Prompt to complete daily mood check-in	✓	✓
Prompt to complete daily food questionnaire	✓	✓
In-app display of weekly data from passive mobile sensing	✓	✓
Personalized insights and in-app resources	✓	✓
Personalized nudges related to behavior change goals		✓
Interpersonal support from a human coach		✓

The primary role of the health coach was to support the participant in using Vira effectively and encourage engagement with the app. Through text conversations, coaches supported participants with usability or technical issues and encouraged consistent use of the app. They also facilitated engagement with BA-informed features of the Vira app. For instance, coaches supported participants in setting goals and creating action plans based on Vira's personalized insights about the patterns of behaviors that may be impacting a user's mood. Coaches encouraged participants to engage in activities that they found pleasurable or rewarding, such as physical, social, or self-care activities, and provided supportive accountability toward meeting their goals.

Coaches accessed a practitioner platform (Vira Pro) to monitor objective behavioral and self-report mood data weekly between sessions, schedule reminders (nudges), and provide encouragement and accountability to participants. In addition, coaches texted with participants through Signal (Signal Technology Foundation), a secure texting platform outside of the Vira app. Coaches were trained to follow a coaching manual and set a schedule for their coaching interactions; specifically, coaches were instructed to schedule weekly, synchronized text chats and to reach out to each participant at least once between weekly coaching sessions. Within these constraints, coaches were instructed to personalize the intervention for each participant based on participant preferences (eg, preference for more vs fewer reminders or nudges to practice techniques).

Participants scheduled a synchronized text chat with their coach approximately once per week. The initial coaching text session lasted approximately 45 to 60 minutes, with additional variability based on the participant's response times and engagement in the conversation. The initial session focused on introductions, rapport building, participant's motivations for joining, expectations for the Vira app and the coach's role, and ways to optimize program participation. Subsequent weekly text chats lasted approximately 10 to 30 minutes and focused on helping participants stay on track with using Vira by identifying facilitators of and barriers to ongoing use and providing performance feedback to support progress toward behavior change goals. At the end of each session, coaches scheduled nudges through Vira Pro. Nudges arrived in the participant's smartphone at scheduled times between weekly sessions to support their behavior change plan.

Coaches also monitored each participant's progress in Vira Pro at least once per week between scheduled weekly text chats. Midweek monitoring, which could include additional nudges or texts, focused on providing accountability toward behavior change goals and encouraging consistent use of Vira. In the last week of the intervention, participants completed a final text chat. Final session topics included encouraging reflection on and the reinforcement of progress to date; strategies for maintaining behavior changes after coaching and access to Vira were no longer available; and providing logistical information about completing the study, including uninstalling the app and completing the follow-up assessment.

Assessments

Consented participants were invited to complete a battery of web-based questionnaires at baseline and again at follow-up after the 12-week intervention period (ie, 12 weeks after installing the Vira app). Assessments required 10 to 20 minutes to complete at baseline and 20 to 30 minutes to complete at follow-up. Participants received up to 4 reminders to complete the questionnaires at each time point. Participants who did not complete the follow-up assessment within 4 weeks of the end of the intervention period were considered lost to follow-up.

Measures

Participant Characteristics

At baseline, participants self-reported their age, gender, race, ethnicity, education, and employment status. They also provided information about obesity risk factors, including self-reported height and weight and parental history of overweight.

App Feasibility and Engagement

Feasibility of and engagement with the Vira platform were measured through metrics derived from objective and self-reported use data. Objective use metrics included (1) the daily active use and (2) retention. Daily active use was operationalized as the percentage of users who completed the daily mood check-in and the daily food questionnaire per day. Retention was defined as the furthest day from installation on which active use was recorded for each participant (values range from 0 to 90) and is displayed as the percentage of the sample retained on a specific intervention day. Self-reported use metrics included the frequency of viewing one’s activity data, mood graph, insight details, and insight resources in the Vira app.

Acceptability

The TAM

Acceptability of the Vira platform was measured with 11 items based on the TAM [50]. The model centers on 2 specific beliefs shown to influence the acceptance of technology: perceived ease of use and perceived usefulness. Perceived ease of use was measured with a 10-item scale on which participants rated their experience using Vira on a 7-point Likert scale from 1 (strongly disagree) to 7 (strongly agree). Perceived usefulness was assessed with a single item about the usefulness of the app, using the same Likert scale as perceived ease of use items. Higher ratings indicated higher acceptability.

Follow-Up Feedback Questionnaire

At follow-up, participants were asked to rate the extent to which participating in the Vira program impacted them in three different ways: (1) changed them or their behavior, (2) increased their motivation to improve their mental health and well-being, and (3) increased their confidence in improving their mental health and well-being. Items were assessed on a 6-point Likert scale from 1 (strongly disagree) to 6 (strongly agree).

Effectiveness

The PHQ-8

The PHQ-8 is an 8-item self-report measure assessing the severity of depressive symptoms [51,57]. The PHQ-8 has high internal consistency (Cronbach α=0.89) and construct validity, with a score ≥10 used to define current depression [51].

Generalized Anxiety Disorder-7

The Generalized Anxiety Disorder-7 (GAD-7) is a 7-item self-report measure of the severity of anxiety symptoms, including excessive worry, restlessness, and difficulty concentrating. The GAD-7 has shown high internal consistency and convergent validity with other anxiety measures [58].

Perceived Stress Scale-10

The Perceived Stress Scale-10 (PSS-10) is a 10-item questionnaire that measures the degree to which various situations are perceived as stressful, capturing both subjective feelings and the perceived ability to cope with stress [59]. The PSS-10 has shown good reliability and validity [60].

Patient-Reported Outcomes Measurement Information System Measures

Sleep disturbance, sleep-related impairment, and emotional support were assessed via Patient-Reported Outcomes Measurement Information System (PROMIS) short form self-report measures for each construct (Sleep Disturbance, Sleep-Related Impairment, and Emotional Support, respectively) [61-63]. The PROMIS sleep disturbance and sleep-related impairment measures have demonstrated excellent measurement properties across various patient groups and interventions [64]. Each PROMIS measure yields a standardized *t* score with a mean of 50 and an SD of 10. Higher scores indicate higher levels of the construct.

Analytic Plan

Participant Characteristics

Descriptive statistics and frequencies were used to examine participant characteristics within and across study arms. Chi-square and 2-sample 2-tailed *t* tests were used to examine between-arm differences in baseline characteristics. Overall study attrition and differential study attrition (ie, after assessment completion) between study arms were also calculated.

Feasibility and Engagement

Analyses of feasibility and engagement were restricted to users who downloaded the Vira app and connected to a coach for coached users or a study coordinator for self-care users. Metrics were evaluated per study arm and are reported using frequencies and descriptive statistics. Linear regression and Fisher exact tests were used to evaluate the effects of study arm on objectively measured days of active use and self-reported use of Vira features, respectively. Of the 35 self-care users, 4 (11%) reported never seeing data on the Vira app home screen and were excluded from analyses of that variable. Overall, 4 (10%) of the 38 coached users and 2 (6%) of the 35 self-care users reported never receiving insights in the Vira app and were excluded from analyses of insight features. Data were analyzed using the R software (version 4.1.3; R Foundation for Statistical Computing).

App Acceptability

App acceptability was assessed through items adapted from the TAM and the follow-up feedback questionnaire. Analyses were restricted to users who downloaded the Vira app and completed the follow-up assessment. The effects of study arm on acceptability ratings from the adapted TAM were evaluated using separate linear regression models for each outcome (perceived ease of use and perceived usefulness). For the analysis of the follow-up feedback items (ie, the extent to which Vira changed participants or their behavior, increased their motivation, and improved their confidence), response options were dichotomized into 2 categories representing the top half (strongly agree, agree, and slightly agree) and bottom half (slightly disagree, disagree, and strongly disagree) of the Likert scale. Separate Fisher exact tests were used to evaluate the effect of study arm on each feedback item. Data were analyzed using the R software.

Effectiveness

Descriptive statistics and frequencies were calculated at each time point (baseline and 12 weeks) for each outcome: PHQ-8, GAD-7, PSS-10, PROMIS sleep disturbance, PROMIS sleep-related impairment, and PROMIS emotional support. An intention-to-treat approach was used to avoid the overestimation of intervention effects on patient-reported outcomes. Multiple imputation was used to account for missing outcome data at the 12-week time point. Little's Missing Completely at Random test [65] was used to assess whether data were missing at random or missing completely at random.

Separate linear mixed-effects regression models were used to examine between-group differences over time for each outcome. Each model included fixed-effect terms for group (coaching vs self-care), time point (baseline and 12 weeks), and the time-by-group interaction. To account for the correlation between repeated measures of the same individuals over time, each model included a participant-level random intercept. Differences in outcome scores within each group over time were explored using linear mixed-effects models. Each model included a fixed-effect term for time point (baseline and 12 weeks) and a participant-level random intercept. Data were analyzed using the *lme4* package (version 1.1.233) [66] in the R software.

Results

Participant Characteristics and Study Flow

A total of 73 individuals completed the baseline assessment and were randomly assigned to either the Vira+Coaching (n=38, 52%) or Vira Self-Care (n=35, 48%) arm. Figure 3 presents the CONSORT (Consolidated Standards of Reporting Trials) diagram. Overall, 35 (92%) of the 38 coached participants and 31 (91%) of the 35 self-care participants downloaded the Vira app. Of the total 73 participants, 18 (25%; coached: 8/18, 44%; self-care: 10/35, 29%) were lost to follow-up, resulting in an overall attrition rate of 25% (18/73; coached=8/38, 21%; self-care=10/35, 29%) and a differential attrition rate of 7.5% (no significant difference between study arms; *P*=.59). Demographic characteristics of study participants at baseline, stratified by arm, are provided in Table 2. There was a significant difference in average age at baseline between the coached and self-care groups (*P*=.006). No other significant between-group differences in demographic characteristics emerged (*P*>.05).

Figure 3. CONSORT (Consolidated Standards of Reporting Trials) diagram.

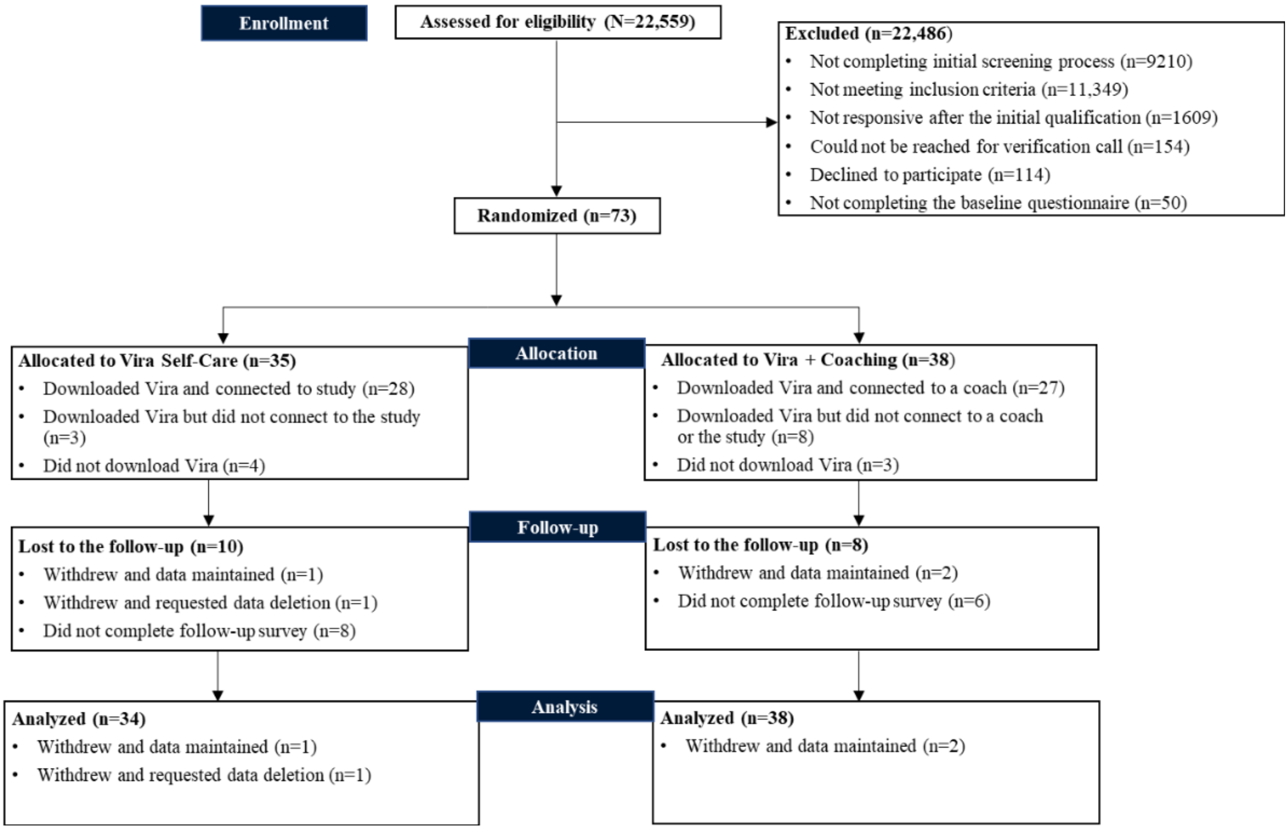


Table 2. Baseline characteristics by study arm.

Demographic characteristics	Vira+Coaching (n=38)	Vira Self-Care (n=35)	Overall (N=73)
Age (y), mean (SD)	21.2 (2.2)	22.6 (1.0)	21.9 (2.2)
BMI (kg/m^2), mean (SD)	30.3 (8.7)	28.4 (7.2)	29.4 (8.0)
Overweight status, n (%)			
Self-overweight	25 (66)	20 (57)	45 (62)
Parental history of overweight	32 (84)	29 (83)	61 (84)
Self and parental history of overweight	19 (50)	14 (40)	22 (45)
Gender, n (%)			
Cisgender female	25 (66)	23 (66)	48 (66)
Cisgender male	9 (24)	9 (26)	18 (25)
Transgender female	1 (3)	1 (3)	2 (3)
Transgender male	0 (0)	1 (3)	1 (1)
Nonconforming	3 (8)	1 (3)	4 (5)
Race, n (%)[a]			
African American	10 (26)	8 (23)	18 (25)
Asian American	3 (8)	4 (11)	7 (10)
White	28 (74)	21 (60)	49 (67)
Other	1 (3)	1 (3)	2 (3)
Hispanic, n (%)	5 (13)	6 (17)	11 (15)
Operating system, n (%)			
Android	18 (47)	15 (43)	33 (45)
iOS	20 (53)	20 (57)	40 (55)
Depressive symptoms (PHQ-8[b]), mean (SD)	14.8 (5.1)	14.0 (4.3)	14.4 (4.7)
Anxiety symptoms (GAD-7[c]), mean (SD)	13.9 (5.6)	13.7 (4.7)	13.8 (5.2)

[a]The totals across the categories are higher than the group sizes because participants could self identify with more than one racial identity.

[b]PHQ-8: 8-item Patient Health Questionnaire.

[c]GAD-7: Generalized Anxiety Disorder-7.

App Feasibility and Engagement

The percentage of active users per day over the 90-day intervention period is displayed in Figure 4. Among participants who downloaded the Vira app and connected it to the study, those allocated to Vira+Coaching had a median of 63 (IQR 15-79.5) active days (63/90, 70% of intervention days), and Vira Self-Care users had a median of 45.5 (IQR 21.8-75) active days (45.5/90, 51% of intervention days); the number of active days did not significantly differ between groups (β=–12.67, SE=8.05; t_{50}=–1.58, P=.12). Overall, 30% (8/27) of users engaged with the app on <30% (26/90) of intervention days in the Vira+Coaching group, whereas 32% (9/28) of users engaged with the app on <30% (26/90) of intervention days in the Vira Self-Care group. Retention rates among both coached and self-care users declined across the first 3 weeks of the intervention period. The retention of coached participants remained relatively stable for the remainder of the 90-day intervention period, whereas the retention of self-care participants continued to decline over time (Figure 5).

Figure 4. Daily active use over the 90-day intervention period by study arm. Users who downloaded Vira and connected it to the trial (55/73, 75%) are included. Actively engaged=completed daily mood check-in or food questionnaire.

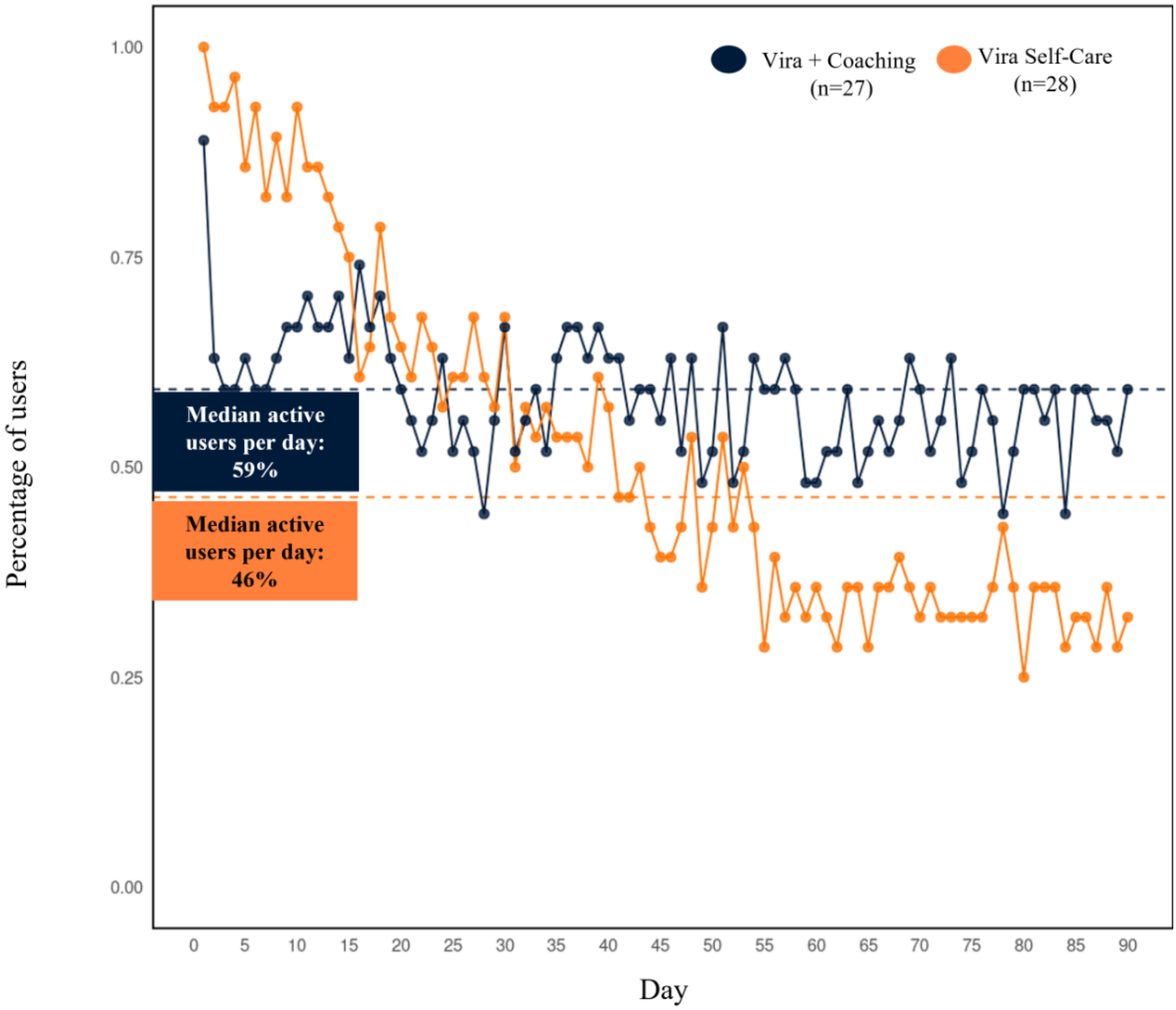
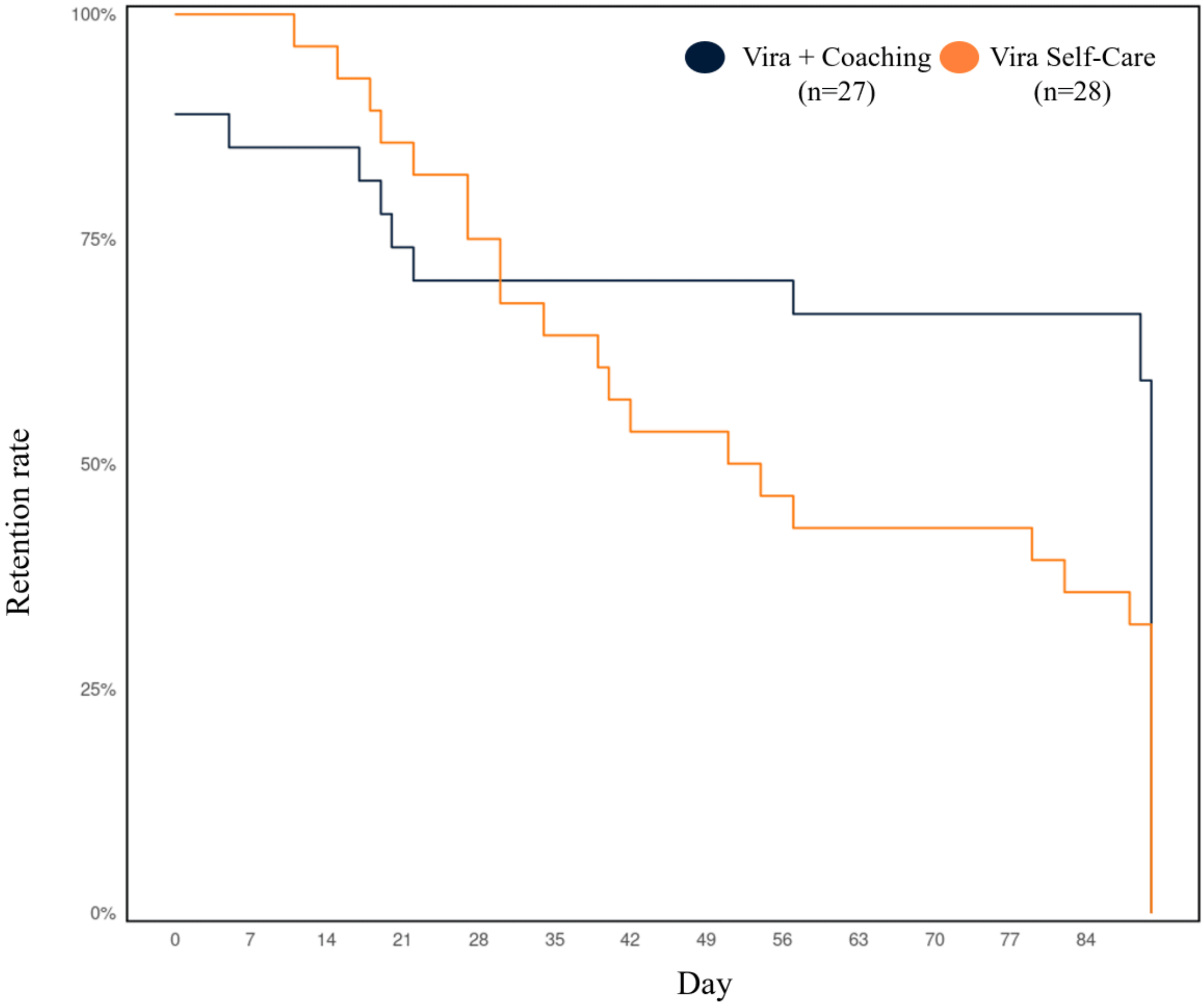
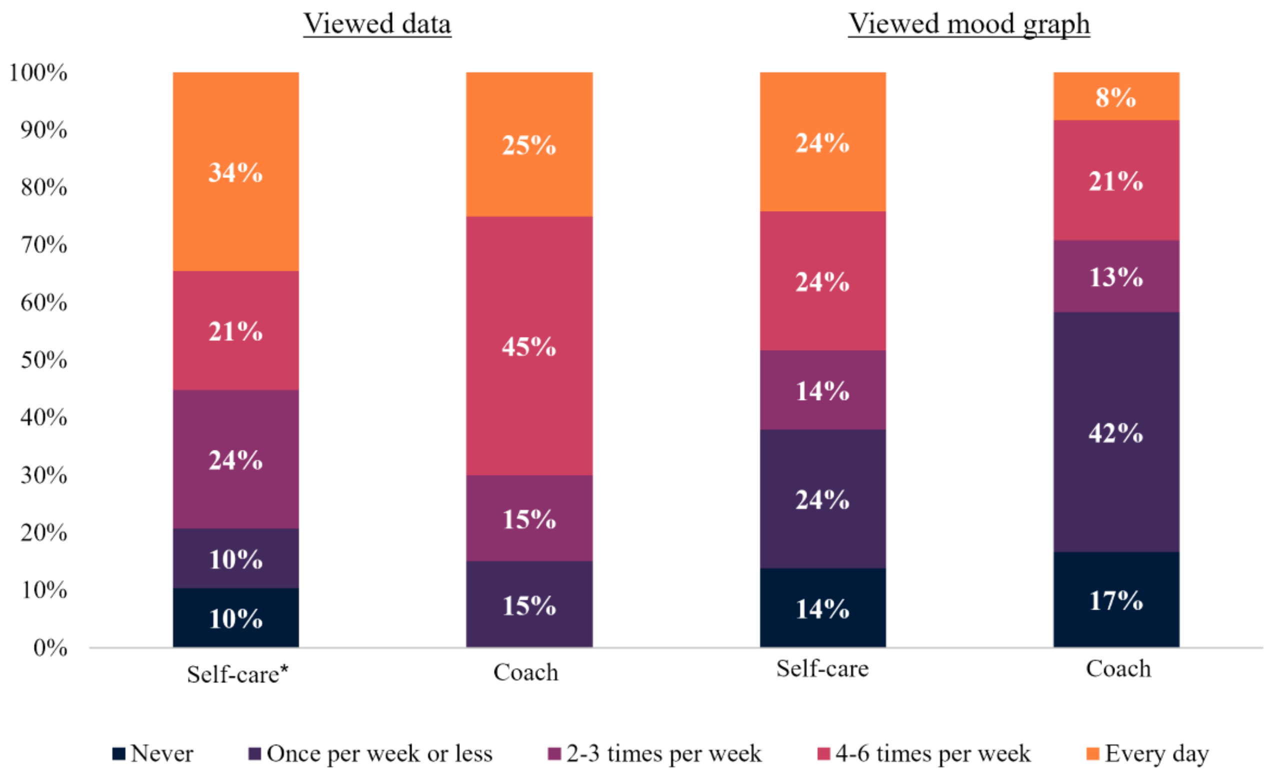


Figure 5. User retention by the length of time in study. Users who downloaded Vira and connected it to the trial (55/73, 75%) are included. Retention=last active day of use across the 90-day intervention period. Values range from 0 to 90, with app installation occurring on day 0.



Participants in both the coached and self-care groups reported high engagement with Vira features (Figure 6). There were no significant between-group differences in the frequency of viewing one's activity data (*P*=.30) or the weekly mood graph (*P*=.52). Users also reported high engagement with insights features. All coached users and 96% (21/22) of self-care users who reported receiving a Vira insight endorsed viewing an insight's details at least once (*P*=.46). A greater proportion of coached than self-care users reported viewing psychoeducational resources upon receiving an insight at least once, although this difference was not statistically significant (Vira+Coaching: 22/25, 88%; Vira Self-Care: 14/22, 64%; *P*=.08; odds ratio [OR] 4.06, 95% CI 0.80-27.82).

Figure 6. Self-reported engagement with Vira features. Users who downloaded Vira, connected it to the trial, and completed the postintervention assessment (coached: 29/38, 76%; self-care: 24/35, 69%) are included. A total of 4 self-care users reported never seeing data on the Today screen and were excluded from the analysis of this variable.



App Acceptability

Coached and self-care participants reported similar ratings of acceptability, such that there were no significant group effects on perceived ease of use (β=.13, SE=1.29; t_{51}=0.10, *P*=.92; Cohen *d*=–0.03, 95% CI –0.57 to 0.51), or perceived usefulness (β=–.19, SE=0.39; t_{51}=–0.49, *P*=.63; Cohen *d*=0.13, 95% CI –0.41 to 0.68). Ratings for individual perceived ease of use and perceived usefulness items are displayed in Table 3. Moreover, a similar proportion of users across groups felt that participating in the Vira program changed themselves or their behavior (Vira+Coaching: 20/29, 69%; Vira Self-Care: 17/24, 71%; *P*>.99; OR 0.92, 95% CI 0.23-3.47) and increased their motivation to improve their mental health (Vira+Coaching: 23/29, 79%; Vira Self-Care: 17/24, 71%; *P*=.53; OR 1.56, 95% CI 0.37-6.78). A greater proportion of coached users than self-care users reported increased confidence in their ability to improve their mental health (Vira+Coaching: 24/29, 83%; Vira Self-Care: 13/24, 54%; *P*=.04; OR 3.95, 95% CI 1.00-17.88).

Table 3. Mean Vira acceptability ratings by study group. Includes users who downloaded Vira and completed Technology Acceptance Model items at follow-up (53/73, 73%).

	Vira+Coaching, mean (SD)	Vira Self-Care, mean (SD)
Perceived ease of use		
Easy to use	6.1 (1.1)	6.0 (1.4)
Easy to become skillful	5.8 (1.4)	5.5 (1.6)
Easy to understand	6.1 (1.1)	6.2 (0.9)
Mentally effortful	2.7 (1.7)	3.2 (1.6)
Easy to remember how to use	5.9 (1.1)	5.8 (1.5)
Rigid and inflexible	2.9 (1.8)	2.8 (1.8)
Confusing	2.3 (1.7)	2.2 (1.4)
Frustrating	2.6 (2.0)	2.5 (1.8)
Easy to learn	5.9 (1.3)	6.0 (1.2)
Cumbersome	3.4 (1.8)	3.5 (1.5)
Perceived usefulness	5.5 (1.5)	5.3 (1.3)

Effectiveness

Effectiveness data were imputed for 17 participants who were missing follow-up assessments. Little's test revealed that data were missing at random and not missing completely at random (χ^2_{11}=23.2; *P*=.02). Changes in patient-reported outcomes from baseline to 12 weeks within and between the coaching and self-care arms are presented in Table 4. Reductions in depressive symptoms were significantly greater in the coaching arm than in the self-care arm, with a large effect size for the coached group (Cohen *d*=–1.07, 95% CI –1.59 to –0.55) versus a small, nonsignificant effect size for the self-care group (Cohen *d*=–0.32, 95% CI –0.85 to 0.21; Table 4). Coached participants also reported significantly greater reductions in anxiety symptoms (GAD-7) than self-care participants, with a large effect size for the coached group (Cohen *d*=–1.02, 95% CI –1.53 to –0.50) versus a small, nonsignificant effect size for the self-care group (Cohen *d*=–0.12, 95% CI –0.65 to 0.41; Table 4). No further group × time differences emerged for perceived stress, sleep-related impairment, sleep disturbance, or emotional support.

Table 4. Change and differences in patient-reported outcomes at baseline and follow-up (12 weeks) by study arm (N=73).

Variable and group	Within-group change over time					Difference in change between groups		
	Baseline, mean (SD)	Follow-up, mean (SD)	Change	*P* value	Effect size (95% CI)	Estimate (SE)	*P* value	Effect size, (95% CI)
PHQ-8[a]						4.01 (1.58)	.01	0.45 (0.10 to 0.82)
Coaching	14.76 (0.88)	9.58 (0.98)	–5.19 (1.21)	<.001	–1.07 (–1.59 to –0.55)			
Self-care	14.15 (0.74)	13.00 (0.85)	–1.15 (0.97)	.24	–0.32 (–0.85 to 0.21)			
GAD-7[b]						4.39 (1.61)	.007	0.50 (0.13 to 0.86)
Coaching	13.92 (0.94)	9.04 (1.04)	–4.88 (1.19)	<.001	–1.02 (–1.54 to –0.50)			
Self-care	13.56 (0.83)	13.08 (0.95)	–0.48 (1.04)	.65	–0.12 (–0.65 to 0.41)			
PSS-10[c]						2.68 (1.58)	.09	0.31 (–0.05 to 0.67)
Coaching	26.11 (0.91)	20.86 (1.02)	–5.24 (1.24)	<.001	–1.05 (–1.57 to –0.53)			
Self-care	27.09 (0.90)	24.53 (0.99)	–2.56 (0.88)	.005	–0.78 (–1.33 to –0.23)			
PROMIS[d] Sleep Disturbance 4a						1.91 (1.28)	.14	0.27 (–0.09 to 0.63)
Coaching	53.47 (0.55)	52.63 (0.62)	–0.85 (0.71)	.24	–0.30 (–0.78 to –0.20)			
Self-care	53.14 (0.94)	54.22 (1.07)	1.07 (1.10)	.33	0.26 (–0.27 to 0.79)			
PROMIS Sleep-Related Impairment 4a						2.58 (2.87)	.37	0.16 (–0.19 to 0.58)
Coaching	63.06 (1.66)	58.45 (1.87)	–4.61 (2.27)	.046	–0.51 (–1.00 to –0.01)			
Self-care	64.46 (1.41)	62.58 (1.58)	–1.87 (1.56)	.23	–0.32 (–0.85 to 0.21)			
PROMIS Emotional Support 4a						2.00 (2.41)	.41	0.15 (–0.21 to 0.51)
Coaching	46.40 (1.57)	49.02 (1.68)	2.62 (1.38)	.06	0.48 (–0.02 to 0.97)			
Self-care	44.52 (1.55)	49.09 (1.78)	4.58 (2.01)	.03	0.61 (0.07 to 1.15)			

[a]PHQ-8: 8-item Patient Health Questionnaire.

[b]GAD-7: Generalized Anxiety Disorder-7.

[c]PSS-10: Perceived Stress Scale-10.

[d]PROMIS: Patient-Reported Outcomes Measurement Information System.

Within-group models indicated that participants in both the coaching and self-care arms reported significant reductions in perceived stress over time, with medium-to-large effect sizes for both groups (Vira+Coaching: Cohen *d*=–1.05, 95% CI –1.57 to –0.50; Vira Self-Care: Cohen *d*=–0.78, 95% CI –1.33 to –0.23). Furthermore, the coaching arm reported significant reductions in sleep-related impairment from baseline to 12 weeks with a medium effect size (Cohen *d*=–0.51, 95% CI –1.00 to –0.01). Within the self-care group, there was a significant increase in emotional support from baseline to 12 weeks (Cohen *d*=0.48, 95% CI –0.02 to 0.97). There were no significant within-group differences in change over time for sleep disturbance. Table 4 provides full results.

Discussion

Principal Findings

This is the first pilot randomized controlled trial to examine the feasibility, acceptability, and effectiveness of a BA-focused digital health product in young adults with elevated depressive symptoms and risk for obesity. Both versions of the app demonstrated high engagement and acceptability. Users reported that the Vira platform was useful and easy to use, which likely fostered their motivation and confidence to enact behavioral changes. Users receiving coaching alongside the Vira app exhibited more sustained engagement than those who used the app on their own. The intervention provided by the Vira platform was effective, with those receiving coaching reporting larger reductions in depression and anxiety than self-care users. Both intervention conditions resulted in reductions in stress, and coached users reported less sleep-related impairment.

Engagement and Acceptability

Both coached and self-care users showed strong and sustained engagement with Vira. Coached users engaged on a greater number of intervention days than self-care users (median 63, IQR 15-79.5 vs 45.5, IQR 21.8-75), and there was a higher median proportion of active users per day in the coaching group than in the self-care group, although differences between conditions were not statistically significant. The engagement rates in this study exceeded those observed in many trials of app-based interventions, although use is likely to be lower outside of a trial context. For example, Baumel and Kane [67] found that the median daily use rate for real-world depression apps was 4.8%. A subsequent review of studies that proactively recruited users and included pre-to-post assessment comparisons (ie, a clinical trial context) revealed that the median digital mental health intervention use was 4.06 times higher than the real-world use of the same intervention [68]. The objectively measured engagement rates observed in both the self-care and coaching conditions in this study exceeded both these industry benchmarks. Participants in both conditions also self-reported frequent use of Vira product features, and there were similar ratings of perceived usefulness and ease of use in the coaching and self-care groups. Taken together, these findings provide compelling evidence of Vira's feasibility and use.

The completion of the daily check-in item and food questionnaire was used as the objective metric of daily engagement because these were the only user actions that were expected to be performed daily and that could be objectively tracked in the app at the time of the study. While there is no universally accepted and comprehensive definition of engagement with mental health interventions, it has been suggested that engagement is a multifaceted construct that includes user interactions with an app's user interface and user experience and engagement with the behavior change intervention components and active ingredients specifically designed to influence behavioral determinants, which in turn impact behavior change [69-71]. Therefore, we also analyzed user engagement with Vira features designed to support BA, such as viewing personalized behavioral insights and resources. However, these data were collected via self-report, and participants were not queried about all BA-related app interactions (eg, viewing a nudge or reminder related to their behavior change goal). Since the study was performed, additional automated tracking functionality has been added to the Vira platform. Future studies of Vira will benefit from a more robust analysis of objective engagement metrics representing user engagement with both user interface and user experience components and behavior change intervention components that support BA and behavior change. In particular, engagement metrics aligned with behavior change intervention components will allow further exploration of mechanisms of action, both for behavior change and improvements in clinical outcomes, and support continued optimization of the Vira platform and its implementation.

Notably, self-care users demonstrated higher engagement early in the intervention, but their engagement was less sustained than that of coached users. The early advantage for self-care users is likely attributable to the lower friction associated with onboarding to an app-only experience versus having to coordinate with a coach to get started. Efficient access to digital mental health apps without waiting time is particularly important for high-risk populations, given that depression is associated with both (1) high levels of negative affect that lead to people searching for things to try in periods of escalating distress and (2) difficulties with sustained motivation and energy [72,73]. This finding also speaks to the relative benefits of the 2 approaches. Self-care users appear to have found it easier to navigate their initial engagement, which they could manage independently, whereas coached users found it easier to sustain their engagement across the 12-week intervention period.

In terms of acceptability, participants across conditions reported that they found the intervention to be useful and acceptable, with few differences between the conditions. Users reported that the platform helped them change their behavior, increased their confidence in improving their mental health and motivation to improve their mental health, and helped them learn about their behavior. In the coaching condition, nearly all participants reported that working with a coach motivated them to change their behavior and improve their well-being.

Effectiveness

The primary effectiveness outcome for this study was depressive symptoms as measured by the PHQ-8. Coached users experienced greater reductions in depressive and anxiety symptoms than self-care users after 12 weeks of use. Moreover, the magnitude of changes observed for these outcomes in the coached group is on par with established minimally clinically important differences [74]. The between-group effect sizes reported in this study for depressive symptoms (Cohen d=0.45) and anxiety symptoms (Cohen d=0.50) are consistent with meta-analyses of mobile apps for mental health issues and a meta-analysis of 28 studies of traditional (ie, nondigital) BA treatments, with some variability based on the comparator (eg, waitlist control, active control, or inactive control [19,75]).

With regard to other secondary outcomes, both coaching and self-care groups reported reductions in perceived stress and sleep-related impairment (ie, alertness, sleepiness, and tiredness during usual waking hours). By contrast, sleep disturbance (ie, sleep quality, depth, and perceived difficulties related to getting and staying asleep) did not significantly change across the intervention period. We also observed that both groups experienced an improvement in their experiences of emotional support, although this improvement was only statistically significant for self-care users. Overall, these effect sizes are useful for powering and guiding future trials to further understand how the Vira platform impacts mental health outcomes.

The findings that coached users showed more sustained engagement (albeit not a significantly greater number of active days) and greater improvements in depressive and anxiety symptoms than self-care users are consistent with literature suggesting that human support enhances the adherence to [38] and efficacy of [34] digital mental health interventions. In a similar trial comparing coached versus unguided (ie, self-care) treatment with a digital mental health app (with or without weekly app recommendations), Mohr and colleagues [42] found that coaching resulted in more app downloads but not a greater number of app sessions or increased use. Participants in both the coached and self-guided arms of that trial showed significant reductions in depressive (9-item Patient Health Questionnaire) and anxiety (GAD-7) symptoms over time, with significantly stronger effects for the coached group in anxiety (although not depressive) symptoms. Moreover, participants who received weekly app recommendations (similar to the nudges sent by coaches in the Vira+Coaching condition) showed greater improvement in depressive symptoms than those who were told to browse the available apps. This study is also not the first to evaluate a digital platform that uses mobile sensing data to personalize treatment components. Frank and colleagues [76] reported a trial with Cue, a digital intervention that uses mobile sensing data to provide participants with personalized microinterventions based on social rhythm therapy. Their trial found that the combination of Cue with treatment as usual was not more effective than digital monitoring alone with treatment as usual for treating depression, but a subgroup analysis did show a stronger effect of the Cue intervention in a subgroup with moderately severe to severe depression. Currey and Torous [77] reported on a study using the MindLAMP app, which used mobile sensing data to provide algorithm-recommended personalized app interventions. Participants were allocated to conditions where they received (1) an automated email signed by a human research assistant encouraging them to complete an additional activity based on their data, (2) a similar activity suggestion from a study bot (Marvin), or (3) no additional activity suggestions. Outcome evaluations, however, did not detect any differences between the 3 conditions in terms of the number of activities completed or changes in depressive or anxious symptoms. Combining these previous findings with those presented here suggests that 2 aspects of the coached version of Vira, the presence of a human coach combined with the use of nudges toward specific behavior change recommendations, may be responsible for the stronger effects observed in the coached condition in this study. This has implications for future developments of the self-care version of Vira (as well as other groups developing self-guided digital support tools). For example, a feature that allows self-care users to schedule their own in-app nudges or reminders may enhance effectiveness [42] and could also increase scalability by reducing or eliminating the need for a human coach.

Although both intervention conditions likely require fewer resources than traditional face-to-face interventions, the human coaching required some resources, including training and supervision of coaches, time spent in weekly text sessions, and time spent reviewing Vira between sessions and sending personalized behavior nudges or reminders. With the increasing accuracy of transformer-based machine-learning techniques, including via retrieval-augmented generation for large language models [78], developing and testing software features that would support a personalized, fully virtual coached version of Vira is an important next step. Using reinforcement learning with human feedback, a fully virtual coach model could be trained to use a motivational interviewing approach to understand each user's needs and support personalized goal setting. The virtual coach model could also be trained to recommend specific nudges or reminders based on the user's passively sensed behavioral data (ie, data on activity, sleep, mobility, and language patterns) and chat conversations with the user. Once developed, the acceptability and effectiveness of a fully virtual coach would need to be compared to those of a human coach. While the support functions of interacting with a human (eg, feeling cared for, respected, and validated) may not be fully replicable by an algorithm, early evidence suggests that fully virtual solutions are acceptable and capable of establishing therapeutic bonds. For example, formative research with a fully automated, cognitive behavioral therapy–based conversational agent showed that users established similar working alliances as patients receiving traditional outpatient cognitive behavioral therapy [79]. Another study comparing experiences with a human coach with those with a virtual chatbot coach found that the chatbot could create a coach-like experience for participants and had the advantage of being persistent and fostering greater participant autonomy by more consistently offering choices and options during the coaching chat session [80].

Moreover, a recent systematic review and meta-analysis of the effectiveness of chatbots on lifestyle behaviors found that chatbot interventions are efficacious for many of the lifestyle behaviors targeted by Vira, such as physical activity and sleep duration [81], although larger studies with more rigorous designs, objective behavioral measures, and longer follow-up are needed. There are a limited number of studies examining whether virtual coach solutions lead to similar improvements in clinical outcomes to those caused by human-supported digital interventions. Finally, further research should seek to understand care providers' attitudes toward human or virtual coaches and their willingness to incorporate such interventions into care plans. Although this decentralized research study was conducted in a community-based sample of volunteers, we anticipate that most people using Vira will ultimately be referred by a health care provider, regardless of whether on a clinic waiting list, as part of a stepped care model, or in an integrative behavioral health care model. Thus, in these settings, it will be critical to understand providers' thoughts and concerns about digital and human coaching to ensure broad uptake and reach.

Limitations

A few methodological limitations must be considered. Due to the pilot nature of this trial, it did not benefit from a control condition, which would have enabled us to rule out placebo, Hawthorne, and regression to the mean effects. Although results revealed that users improved in several self-reported outcomes after 12 weeks of using Vira, the relatively small sample size likely reduced statistical power to detect smaller effect sizes. Further, it is unclear whether positive effects would be sustained beyond 12 weeks, necessitating future studies with longer follow-up periods that could also enable researchers to examine preventive effects.

The small sample size, specificity of the eligibility criteria (ie, risk for obesity and elevated depressive symptoms), and small absolute numbers of men and some racial and ethnic minority groups point to the need for future studies with larger, more representative samples and broader eligibility criteria to validate the present findings and provide data regarding the generalizability of the findings. Notably, there were a limited number of male participants in the sample (n=9 in each group, representing ~25% of the sample). To ensure the intervention was reaching participants who could most benefit from a BA program, participants were required to report clinically significant depressive symptoms (eg, PHQ-8 score≥10). The lower rate of male participants in this sample is consistent with the higher prevalence of major depression among both adult and adolescent female individuals than among male individuals [82], as well as evidence suggesting higher rates of digital mental health app use among female individuals [83]. Nonetheless, the demographics of the current sample highlight both the known difficulties in recruiting and engaging male individuals in mental health research [84] and the importance of developing and implementing recruitment strategies that target underserved populations. In addition, friction in the experience of onboarding with a coach often led to delays between when coached users downloaded Vira and when they started the intervention. This may have impacted retention. Furthermore, as the intervention consisted of multiple components, most of which were not objectively tracked, we could not quantitatively determine which specific features were effective. Identifying the most effective components using more sophisticated experimental designs [85] is an important future direction. Although using an intent-to-treat analytic approach was a strength, 23% (17/73) of follow-up assessment data were imputed, which may have introduced potential bias and affected the reliability of the results. Finally, it should be acknowledged that the lead investigators in this study (albeit not all investigators) are employed by the company that designed and commercially licenses the Vira platform. Independent evaluations will be important in the future, and others should seek to replicate and extend these findings. The Vira platform is generally available for research and clinical apps to facilitate these evaluations.

Strengths

This study also had several strengths. The fully decentralized trial design supported the recruitment of a heterogeneous community sample, reflecting the complexities and diversity of everyday life and further enhancing the generalizability and external validity of the findings. The sample also included a fairly balanced proportion of Android and iOS users. In addition, the study used validated patient-reported outcome measures, enhancing the reliability and validity of the data collected, as well as the extent to which these findings can be replicated and placed within the broader literature. Finally, despite the importance of this issue for optimizing digital interventions, there are relatively few studies that directly compare coached versus self-guided versions of the same digital intervention within a unified study protocol. For instance, a recent exploratory analysis by Chang and colleagues [86] generated effect sizes for a digital intervention by analyzing and comparing the results of 2 distinct studies of the same tool (unguided vs coached versions); however, inclusion criteria differed between the 2 trials, and there were high rates of missingness in the study of the unguided version of the intervention, presenting potential bias. This study design, which directly compared coached versus self-guided versions, provides insights into who can most benefit from a coaching intervention, supporting a more efficient and targeted allocation of resources to meet individuals' unique needs. For example, these results suggest that a coached intervention may be more effective than a self-guided intervention for individuals seeking treatment for elevated depression or anxiety symptoms. By contrast, a self-guided intervention may be as effective as or potentially more efficient than a coached intervention for individuals whose primary concern is stress, sleep disturbance or associated impairment, or emotional support. These findings can facilitate stratified care by informing providers' decisions regarding whether to refer a patient to a self-guided or coached intervention.

Conclusions

This study found that both versions of the Vira app showed strong engagement compared to benchmarks, and users who received coaching along with the Vira app showed more sustained engagement than those who used the app for self-care. Both versions of Vira were found to be acceptable and showed initial effectiveness in improving various dimensions of behavioral health and quality of life for young adults with depressive symptoms and obesity risk factors; however, users who received coaching reported greater reductions in depressive and anxiety symptoms than self-care users. Both intervention conditions were perceived as useful and easy to use and increased participants' motivation and confidence for making behavioral changes. Overall, the combination of Vira with coaching was found to be more engaging and effective, but an enhancement to the self-care version of the app may enhance these outcomes for self-guided users. This study represents a unique contribution to the literature, with indicators of feasibility, acceptability, and effectiveness sufficient to support a fully powered trial in a larger sample. To our knowledge, this is the first trial of an app-based intervention using passive mobile sensing to track behavior and deliver personalized insights. Future directions include conducting a fully powered controlled trial with longer follow-up, optimizing the interventions, developing and testing a virtual coach-supported version, and studying salient mediators and moderators of outcome effects using more objective engagement metrics.

Acknowledgments

The authors would like to thank Kristal Hartman, Karen Burgard, and Tony Scripa for their contributions as health coaches. The authors also wish to acknowledge Danielle DeLuise for her support throughout the project.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

LSW, RNC, LBS, FHK, JFD, and NBA designed the study; LSW and MW performed data curation; RNC performed formal data analysis; LSW and NBA drafted the paper; LSW, RNC, FHK, JFD, and NBA interpreted results. All authors read and approved the final version of the paper and agreed with the order of presentation of the authors.

Conflicts of Interest

LSW, RNC, MW, and NBA are employed by and hold equity in Ksana Health, which developed and commercialized the Vira platform. LBS was employed by Ksana Health at the time the project was initiated. FHK was employed by Novo Nordisk during the study, and JFD is currently employed by Novo Nordisk, which provided funding for the study. RPA is an unpaid adviser to Ksana Health. Ksana Health has a Scientific Integrity Policy designed to prevent the company's interest in bringing Ksana Health products to market from influencing or appearing to influence the science conducted using Ksana Health's products. Ksana Health follows open science practices to ensure the transparency, integrity, and reproducibility of its work. This includes preregistration before data collection and data, code, and material sharing upon request. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
CONSORT-EHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 3622 KB - mental_v11i1e51366_app1.pdf]

References

1. Paus T, Keshavan M, Giedd JN. Why do many psychiatric disorders emerge during adolescence? Nat Rev Neurosci 2008 Dec;9(12):947-957 [FREE Full text] [doi: 10.1038/nrn2513] [Medline: 19002191]
2. Kessler RC, Avenevoli S, Green J, Gruber MJ, Guyer M, He Y, et al. National comorbidity survey replication adolescent supplement (NCS-A): III. Concordance of DSM-IV/CIDI diagnoses with clinical reassessments. J Am Acad Child Adolesc Psychiatry 2009 Apr;48(4):386-399 [FREE Full text] [doi: 10.1097/CHI.0b013e31819a1cbc] [Medline: 19252450]
3. Avenevoli S, Swendsen J, He JP, Burstein M, Merikangas KR. Major depression in the national comorbidity survey-adolescent supplement: prevalence, correlates, and treatment. J Am Acad Child Adolesc Psychiatry 2015 Jan;54(1):37-44.e2 [FREE Full text] [doi: 10.1016/j.jaac.2014.10.010] [Medline: 25524788]
4. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005 Jun;62(6):593-602. [doi: 10.1001/archpsyc.62.6.593] [Medline: 15939837]
5. Merikangas KR, He JP, Burstein M, Swanson SA, Avenevoli S, Cui L, et al. Lifetime prevalence of mental disorders in U.S. adolescents: results from the National Comorbidity Survey Replication--Adolescent Supplement (NCS-A). J Am Acad Child Adolesc Psychiatry 2010 Oct;49(10):980-989 [FREE Full text] [doi: 10.1016/j.jaac.2010.05.017] [Medline: 20855043]
6. Patton GC, Sawyer SM, Santelli JS, Ross DA, Afifi R, Allen NB, et al. Our future: a Lancet commission on adolescent health and wellbeing. Lancet 2016 Jun 11;387(10036):2423-2478 [FREE Full text] [doi: 10.1016/S0140-6736(16)00579-1] [Medline: 27174304]
7. Tegethoff M, Stalujanis E, Belardi A, Meinlschmidt G. Chronology of onset of mental disorders and physical diseases in mental-physical comorbidity - a national representative survey of adolescents. PLoS One 2016 Oct 21;11(10):e0165196 [FREE Full text] [doi: 10.1371/journal.pone.0165196] [Medline: 27768751]
8. Bruffaerts R, Demyttenaere K, Kessler RC, Tachimori H, Bunting B, Hu C, et al. The associations between preexisting mental disorders and subsequent onset of chronic headaches: a worldwide epidemiologic perspective. J Pain 2015 Jan;16(1):42-52 [FREE Full text] [doi: 10.1016/j.jpain.2014.10.002] [Medline: 25451620]
9. Iacovides A, Siamouli M. Comorbid mental and somatic disorders: an epidemiological perspective. Curr Opin Psychiatry 2008 Jul;21(4):417-421. [doi: 10.1097/YCO.0b013e328303ba42] [Medline: 18520749]
10. Niles AN, O'Donovan A. Comparing anxiety and depression to obesity and smoking as predictors of major medical illnesses and somatic symptoms. Health Psychol 2019 Feb;38(2):172-181 [FREE Full text] [doi: 10.1037/hea0000707] [Medline: 30556708]

11. Tegethoff M, Belardi A, Stalujanis E, Meinlschmidt G. Association between mental disorders and physical diseases in adolescents from a nationally representative cohort. Psychosom Med 2015 Apr;77(3):319-332. [doi: 10.1097/PSY.0000000000000151] [Medline: 25851547]
12. Walker ER, McGee RE, Druss BG. Mortality in mental disorders and global disease burden implications: a systematic review and meta-analysis. JAMA Psychiatry 2015 Apr;72(4):334-341 [FREE Full text] [doi: 10.1001/jamapsychiatry.2014.2502] [Medline: 25671328]
13. Byrne ML, O'Brien-Simpson NM, Mitchell SA, Allen NB. Adolescent-onset depression: are obesity and inflammation developmental mechanisms or outcomes? Child Psychiatry Hum Dev 2015 Dec;46(6):839-850. [doi: 10.1007/s10578-014-0524-9] [Medline: 25666100]
14. Tisi R. Obesity and chronic disease in younger people. Br J Gen Pract 2012 Mar;62(596):123; author reply 123 [FREE Full text] [doi: 10.3399/bjgp12X629973] [Medline: 22429413]
15. Goodman E, Whitaker RC. A prospective study of the role of depression in the development and persistence of adolescent obesity. Pediatrics 2002 Sep;110(3):497-504. [doi: 10.1542/peds.110.3.497] [Medline: 12205250]
16. Needham BL, Epel ES, Adler NE, Kiefe C. Trajectories of change in obesity and symptoms of depression: the CARDIA study. Am J Public Health 2010 Jun;100(6):1040-1046 [FREE Full text] [doi: 10.2105/AJPH.2009.172809] [Medline: 20395582]
17. Malik K, Ibrahim M, Bernstein A, Venkatesh RK, Rai T, Chorpita B, et al. Behavioral activation as an 'active ingredient' of interventions addressing depression and anxiety among young people: a systematic review and evidence synthesis. BMC Psychol 2021 Oct 07;9(1):150 [FREE Full text] [doi: 10.1186/s40359-021-00655-x] [Medline: 34615559]
18. Martin F, Oliver T. Behavioral activation for children and adolescents: a systematic review of progress and promise. Eur Child Adolesc Psychiatry 2019 Apr;28(4):427-441 [FREE Full text] [doi: 10.1007/s00787-018-1126-z] [Medline: 29476253]
19. Stein AT, Carl E, Cuijpers P, Karyotaki E, Smits JA. Looking beyond depression: a meta-analysis of the effect of behavioral activation on depression, anxiety, and activation. Psychol Med 2021 Jul;51(9):1491-1504. [doi: 10.1017/S0033291720000239] [Medline: 32138802]
20. Young J. Behavioral activation as an adjunct treatment to CBT-I. Sleep Med Clin 2023 Mar;18(1):31-38. [doi: 10.1016/j.jsmc.2022.10.002] [Medline: 36764784]
21. Campbell SB, Fortney J, Simpson TL, Jakupcak M, Wagner A. Change in social support while participating in behavioral activation for PTSD. Psychol Trauma 2019 Nov;11(8):905-908 [FREE Full text] [doi: 10.1037/tra0000470] [Medline: 31107046]
22. Choi NG, Pepin R, Marti CN, Stevens CJ, Bruce ML. Improving social connectedness for homebound older adults: randomized controlled trial of tele-delivered behavioral activation versus tele-delivered friendly visits. Am J Geriatr Psychiatry 2020 Jul;28(7):698-708 [FREE Full text] [doi: 10.1016/j.jagp.2020.02.008] [Medline: 32238297]
23. Pagoto S, Bodenlos JS, Schneider KL, Olendzki B, Spates CR, Ma Y. Initial investigation of behavioral activation therapy for co-morbid major depressive disorder and obesity. Psychotherapy (Chic) 2008 Sep;45(3):410-415. [doi: 10.1037/a0013313] [Medline: 22122500]
24. Mazzucchelli TG, Kane RT, Rees CS. Behavioral activation interventions for well-being: a meta-analysis. J Posit Psychol 2010 Mar;5(2):105-121 [FREE Full text] [doi: 10.1080/17439760903569154] [Medline: 20539837]
25. Barnes CM, Drake CL. Prioritizing sleep health: public health policy recommendations. Perspect Psychol Sci 2015 Nov;10(6):733-737. [doi: 10.1177/1745691615598509] [Medline: 26581727]
26. Chekroud AM, Trugerman A. The opportunity for exercise to improve population mental health. JAMA Psychiatry 2019 Nov 01;76(11):1206-1207. [doi: 10.1001/jamapsychiatry.2019.2282] [Medline: 31483446]
27. Crowell SE, Puzia ME, Yaptangco M. The ontogeny of chronic distress: emotion dysregulation across the life span and its implications for psychological and physical health. Curr Opin Psychol 2015 Jun;3:91-99. [doi: 10.1016/j.copsyc.2015.03.023]
28. Dahne J, Kustanowitz J, Lejuez CW. Development and preliminary feasibility study of a brief behavioral activation mobile application (behavioral apptivation) to be used in conjunction with ongoing therapy. Cogn Behav Pract 2018 Feb;25(1):44-56 [FREE Full text] [doi: 10.1016/j.cbpra.2017.05.004] [Medline: 29725198]
29. Huguet A, Rao S, McGrath PJ, Wozney L, Wheaton M, Conrod J, et al. A systematic review of cognitive behavioral therapy and behavioral activation apps for depression. PLoS One 2016 May 02;11(5):e0154248 [FREE Full text] [doi: 10.1371/journal.pone.0154248] [Medline: 27135410]
30. Kazdin AE. Addressing the treatment gap: a key challenge for extending evidence-based psychosocial interventions. Behav Res Ther 2017 Jan;88:7-18. [doi: 10.1016/j.brat.2016.06.004] [Medline: 28110678]
31. Han A, Kim TH. Effects of internet-delivered behavioral activation on individuals with depressive symptoms: a systematic review and meta-analysis. J Psychiatr Res 2022 Aug;152:104-118 [FREE Full text] [doi: 10.1016/j.jpsychires.2022.05.031] [Medline: 35717866]
32. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021 Oct;20(3):318-335 [FREE Full text] [doi: 10.1002/wps.20883] [Medline: 34505369]
33. Schueller SM, Torous J. Scaling evidence-based treatments through digital mental health. Am Psychol 2020 Nov;75(8):1093-1104 [FREE Full text] [doi: 10.1037/amp0000654] [Medline: 33252947]

34. Cuijpers P, Noma H, Karyotaki E, Cipriani A, Furukawa TA. Effectiveness and acceptability of cognitive behavior therapy delivery formats in adults with depression: a network meta-analysis. JAMA Psychiatry 2019 Jul 01;76(7):700-707 [FREE Full text] [doi: 10.1001/jamapsychiatry.2019.0268] [Medline: 30994877]
35. Weisel KK, Fuhrmann LM, Berking M, Baumeister H, Cuijpers P, Ebert DD. Standalone smartphone apps for mental health-a systematic review and meta-analysis. NPJ Digit Med 2019 Dec 2;2:118 [FREE Full text] [doi: 10.1038/s41746-019-0188-8] [Medline: 31815193]
36. Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry 2019 Oct;18(3):325-336 [FREE Full text] [doi: 10.1002/wps.20673] [Medline: 31496095]
37. Werntz A, Amado S, Jasman M, Ervin A, Rhodes JE. Providing human support for the use of digital mental health interventions: systematic meta-review. J Med Internet Res 2023 Feb 06;25:e42864 [FREE Full text] [doi: 10.2196/42864] [Medline: 36745497]
38. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: a systematic and meta-analytic review. J Consult Clin Psychol 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
39. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
40. Meyer A, Wisniewski H, Torous J. Coaching to support mental health apps: exploratory narrative review. JMIR Hum Factors 2022 Mar 08;9(1):e28301 [FREE Full text] [doi: 10.2196/28301] [Medline: 35258468]
41. Graham AK, Greene CJ, Kwasny MJ, Kaiser SM, Lieponis P, Powell T, et al. Coached mobile app platform for the treatment of depression and anxiety among primary care patients: a randomized clinical trial. JAMA Psychiatry 2020 Sep 01;77(9):906-914 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.1011] [Medline: 32432695]
42. Mohr DC, Schueller SM, Tomasino KN, Kaiser SM, Alam N, Karr C, et al. Comparison of the effects of coaching and receipt of app recommendations on depression, anxiety, and engagement in the IntelliCare platform: factorial randomized controlled trial. J Med Internet Res 2019 Aug 28;21(8):e13609 [FREE Full text] [doi: 10.2196/13609] [Medline: 31464192]
43. Hildebrandt T, Michaeledes A, Mayhew M, Greif R, Sysko R, Toro-Ramos T, et al. Randomized controlled trial comparing health coach-delivered smartphone-guided self-help with standard care for adults with binge eating. Am J Psychiatry 2020 Feb 01;177(2):134-142. [doi: 10.1176/appi.ajp.2019.19020184] [Medline: 32008396]
44. Aschbrenner KA, Naslund JA, Tomlinson EF, Kinney A, Pratt SI, Brunette MF. Adolescents' use of digital technologies and preferences for mobile health coaching in public mental health settings. Front Public Health 2019 Jul 02;7:178 [FREE Full text] [doi: 10.3389/fpubh.2019.00178] [Medline: 31312629]
45. Chikersal P, Belgrave D, Doherty G, Enrique A, Palacios JE, Richards D, et al. Understanding client support strategies to improve clinical outcomes in an online mental health intervention. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI '20; April 25-30, 2020; Honolulu, HI. [doi: 10.1145/3313831.3376341]
46. Noetel M, Sanders T, Gallardo-Gómez D, Taylor P, Del Pozo Cruz B, van den Hoek D, et al. Effect of exercise for depression: systematic review and network meta-analysis of randomised controlled trials. BMJ 2024 Feb 14;384:e075847 [FREE Full text] [doi: 10.1136/bmj-2023-075847] [Medline: 38355154]
47. Oftedal S, Kolt GS, Holliday EG, Stamatakis E, Vandelanotte C, Brown WJ, et al. Associations of health-behavior patterns, mental health and self-rated health. Prev Med 2019 Jan;118:295-303. [doi: 10.1016/j.ypmed.2018.11.017] [Medline: 30476503]
48. Shin J, Bae SM. A systematic review of location data for depression prediction. Int J Environ Res Public Health 2023 May 29;20(11):5984 [FREE Full text] [doi: 10.3390/ijerph20115984] [Medline: 37297588]
49. Butryn ML, Webb V, Wadden TA. Behavioral treatment of obesity. Psychiatr Clin North Am 2011 Dec;34(4):841-859 [FREE Full text] [doi: 10.1016/j.psc.2011.08.006] [Medline: 22098808]
50. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319-340. [doi: 10.2307/249008]
51. Kroenke K, Strine TW, Spitzer RL, Williams JBW, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
52. Moss AJ, Hauser D, Rosenzweig C, Jaffe SN, Robinson J, Litman L. Using market-research panels for behavioral science: an overview and tutorial. PsyArXiv Preprint posted online on November 11, 2022 [FREE Full text] [doi: 10.31234/osf.io/wj3bk]
53. Li LY, Trivedi E, Helgren F, Allison GO, Zhang E, Buchanan SN, et al. Capturing mood dynamics through adolescent smartphone social communication. J Psychopathol Clin Sci 2023 Nov;132(8):1072-1084. [doi: 10.1037/abn0000855] [Medline: 37498714]
54. Funkhouser CJ, Trivedi E, Li LY, Helgren F, Zhang E, Sritharan A, et al. Detecting adolescent depression through passive monitoring of linguistic markers in smartphone communication. J Child Psychol Psychiatry 2024 Jul;65(7):932-941. [doi: 10.1111/jcpp.13931] [Medline: 38098445]

55. McNeilly EA, Mills KL, Kahn LE, Crowley R, Pfeifer JH, Allen NB. Adolescent social communication through smartphones: linguistic features of internalizing symptoms and daily mood. Clin Psychol Sci 2023 Nov;11(6):1090-1107 [FREE Full text] [doi: 10.1177/21677026221125180] [Medline: 38149299]
56. Mazzucchelli T, Kane R, Rees C. Behavioral activation treatments for depression in adults: a meta-analysis and review. Clin Psychol Sci Pract 2009 Dec;16(4):383-411. [doi: 10.1111/j.1468-2850.2009.01178.x]
57. Shin C, Lee SH, Han KM, Yoon HK, Han C. Comparison of the usefulness of the PHQ-8 and PHQ-9 for screening for major depressive disorder: analysis of psychiatric outpatient data. Psychiatry Investig 2019 Apr;16(4):300-305 [FREE Full text] [doi: 10.30773/pi.2019.02.01] [Medline: 31042692]
58. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
59. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385. [doi: 10.2307/2136404]
60. Lee EH. Review of the psychometric evidence of the perceived stress scale. Asian Nurs Res (Korean Soc Nurs Sci) 2012 Dec;6(4):121-127 [FREE Full text] [doi: 10.1016/j.anr.2012.08.004] [Medline: 25031113]
61. Yu L, Buysse DJ, Germain A, Moul DE, Stover A, Dodds NE, et al. Development of short forms from the PROMIS™ sleep disturbance and Sleep-Related Impairment item banks. Behav Sleep Med 2011 Dec 28;10(1):6-24 [FREE Full text] [doi: 10.1080/15402002.2012.636266] [Medline: 22250775]
62. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol 2010 Nov;63(11):1179-1194 [FREE Full text] [doi: 10.1016/j.jclinepi.2010.04.011] [Medline: 20685078]
63. Hahn EA, DeWalt DA, Bode RK, Garcia SF, DeVellis RF, Correia H, et al. New English and Spanish social health measures will facilitate evaluating health determinants. Health Psychol 2014 May;33(5):490-499 [FREE Full text] [doi: 10.1037/hea0000055] [Medline: 24447188]
64. Buysse DJ, Yu L, Moul DE, Germain A, Stover A, Dodds NE, et al. Development and validation of patient-reported outcome measures for sleep disturbance and sleep-related impairments. Sleep 2010 Jun;33(6):781-792 [FREE Full text] [doi: 10.1093/sleep/33.6.781] [Medline: 20550019]
65. Little RJ. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 1988;83(404):1198-1202. [doi: 10.1080/01621459.1988.10478722]
66. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw 2015;67(1):1-48. [doi: 10.18637/jss.v067.i01]
67. Baumel A, Kane JM. Examining predictors of real-world user engagement with self-guided eHealth interventions: analysis of mobile apps and websites using a novel dataset. J Med Internet Res 2018 Dec 14;20(12):e11491 [FREE Full text] [doi: 10.2196/11491] [Medline: 30552077]
68. Baumel A, Edan S, Kane JM. Is there a trial bias impacting user engagement with unguided e-mental health interventions? A systematic comparison of published reports and real-world usage of the same programs. Transl Behav Med 2019 Nov 25;9(6):1020-1033. [doi: 10.1093/tbm/ibz147] [Medline: 31689344]
69. Lipschitz JM, Pike CK, Hogan TP, Murphy SA, Burdick KE. The engagement problem: a review of engagement with digital mental health interventions and recommendations for a path forward. Curr Treat Options Psychiatry 2023 Sep;10(3):119-135 [FREE Full text] [doi: 10.1007/s40501-023-00297-3] [Medline: 38390026]
70. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017 Jun;7(2):254-267 [FREE Full text] [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
71. Cole-Lewis H, Ezeanochie N, Turgiss J. Understanding health behavior technology engagement: pathway to measuring digital behavior change interventions. JMIR Form Res 2019 Oct 10;3(4):e14052 [FREE Full text] [doi: 10.2196/14052] [Medline: 31603427]
72. Kornfield R, Zhang R, Nicholas J, Schueller SM, Cambo SA, Mohr DC, et al. "Energy is a finite resource": designing technology to support individuals across fluctuating symptoms of depression. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI '20; April 25-30, 2020; Honolulu, HI. [doi: 10.1145/3313831.3376309]
73. Pretorius C, McCashin D, Kavanagh N, Coyle D. Searching for mental health: a mixed-methods study of young people's online help-seeking. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI '20; April 25-30, 2020; Honolulu, HI URL: https://dl.acm.org/doi/10.1145/3313831.3376328 [doi: 10.1145/3313831.3376328]
74. Kroenke K. Enhancing the clinical utility of depression screening. CMAJ 2012 Feb 21;184(3):281-282 [FREE Full text] [doi: 10.1503/cmaj.112004] [Medline: 22231681]
75. Lecomte T, Potvin S, Corbière M, Guay S, Samson C, Cloutier B, et al. Mobile apps for mental health issues: meta-review of meta-analyses. JMIR Mhealth Uhealth 2020 May 29;8(5):e17458 [FREE Full text] [doi: 10.2196/17458] [Medline: 32348289]

76. Frank E, Wallace ML, Matthews MJ, Kendrick J, Leach J, Moore T, et al. Personalized digital intervention for depression based on social rhythm principles adds significantly to outpatient treatment. Front Digit Health 2022 Sep 2;4:870522 [FREE Full text] [doi: 10.3389/fdgth.2022.870522] [Medline: 36120713]
77. Currey D, Torous J. Digital phenotyping data to predict symptom improvement and mental health app personalization in college students: prospective validation of a predictive model. J Med Internet Res 2023 Feb 09;25:e39258 [FREE Full text] [doi: 10.2196/39258] [Medline: 36757759]
78. Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, et al. Retrieval-augmented generation for large language models: a survey. arXiv Preprint posted online on December 18, 2023 [FREE Full text]
79. Darcy A, Daniels J, Salinger D, Wicks P, Robinson A. Evidence of human-level bonds established with a digital conversational agent: cross-sectional, retrospective observational study. JMIR Form Res 2021 May 11;5(5):e27868 [FREE Full text] [doi: 10.2196/27868] [Medline: 33973854]
80. Mitchell EG, Maimone R, Cassells A, Tobin JN, Davidson P, Smaldone AM, et al. Automated vs. human health coaching: exploring participant and practitioner experiences. Proc ACM Hum Comput Interact 2021 Apr;5(CSCW1):99 [FREE Full text] [doi: 10.1145/3449173] [Medline: 36304916]
81. Singh B, Olds T, Brinsley J, Dumuid D, Virgara R, Matricciani L, et al. Systematic review and meta-analysis of the effectiveness of chatbots on lifestyle behaviours. NPJ Digit Med 2023 Jun 23;6(1):118 [FREE Full text] [doi: 10.1038/s41746-023-00856-1] [Medline: 37353578]
82. Nolen-Hoeksema S. Sex differences in unipolar depression: evidence and theory. Psychol Bull 1987;101(2):259-282. [doi: 10.1037//0033-2909.101.2.259]
83. Tong HL, Maher C, Parker K, Pham TD, Neves AL, Riordan B, et al. The use of mobile apps and fitness trackers to promote healthy behaviors during COVID-19: a cross-sectional survey. PLOS Digit Health 2022 Aug 18;1(8):e0000087. [doi: 10.1371/journal.pdig.0000087] [Medline: 36812578]
84. Balcombe L, De Leo D. The potential impact of adjunct digital tools and technology to help distressed and suicidal men: an integrative review. Front Psychol 2022 Jan 4;12:796371. [doi: 10.3389/fpsyg.2021.796371] [Medline: 35058855]
85. Nahum-Shani I, Dziak JJ, Wetter DW. MCMTC: a pragmatic framework for selecting an experimental design to inform the development of digital interventions. Front Digit Health 2022 Mar 9;4:798025 [FREE Full text] [doi: 10.3389/fdgth.2022.798025] [Medline: 35355685]
86. Chang S, Alon N, Torous J. An exploratory analysis of the effect size of the mobile mental health application, mindLAMP. Digit Health 2023 Jul 07;9:20552076231187244 [FREE Full text] [doi: 10.1177/20552076231187244] [Medline: 37434734]

Abbreviations

BA: behavioral activation
CONSORT: Consolidated Standards of Reporting Trials
GAD-7: Generalized Anxiety Disorder-7
OR: odds ratio
PHQ-8: 8-item Patient Health Questionnaire
PHQ-9: 9-item Patient Health Questionnaire
PROMIS: Patient-Reported Outcomes Measurement Information System
PSS-10: Perceived Stress Scale-10
TAM: Technology Acceptance Model

Edited by J Torous; submitted 28.07.23; peer-reviewed by X Yan, A Parks, M Subotic-Kerry; comments to author 30.01.24; revised version received 01.03.24; accepted 15.06.24; published 19.09.24.

Please cite as:
Weiner LS, Crowley RN, Sheeber LB, Koegler FH, Davis JF, Wells M, Funkhouser CJ, Auerbach RP, Allen NB
Engagement, Acceptability, and Effectiveness of the Self-Care and Coach-Supported Versions of the Vira Digital Behavior Change Platform Among Young Adults at Risk for Depression and Obesity: Pilot Randomized Controlled Trial
JMIR Ment Health 2024;11:e51366
URL: https://mental.jmir.org/2024/1/e51366
doi:10.2196/51366
PMID:

Original Paper

The Most Effective Interventions for Classification Model Development to Predict Chat Outcomes Based on the Conversation Content in Online Suicide Prevention Chats: Machine Learning Approach

Salim Salmi[1], MSc; Saskia Mérelle[1], PhD; Renske Gilissen[1], PhD; Rob van der Mei[2], PhD; Sandjai Bhulai[3], PhD

[1]Research Department, 113 Suicide Prevention, Amsterdam, Netherlands

[2]Department of Stochastics, Centrum Wiskunde & Informatica, Amsterdam, Netherlands

[3]Department of Mathematics, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

Corresponding Author:
Salim Salmi, MSc
Research Department
113 Suicide Prevention
Paasheuvelweg 25
Amsterdam, 1105 BP
Netherlands
Phone: 31 640673474
Email: s.salmi@113.nl

Abstract

Background: For the provision of optimal care in a suicide prevention helpline, it is important to know what contributes to positive or negative effects on help seekers. Helplines can often be contacted through text-based chat services, which produce large amounts of text data for use in large-scale analysis.

Objective: We trained a machine learning classification model to predict chat outcomes based on the content of the chat conversations in suicide helplines and identified the counsellor utterances that had the most impact on its outputs.

Methods: From August 2021 until January 2023, help seekers (N=6903) scored themselves on factors known to be associated with suicidality (eg, hopelessness, feeling entrapped, will to live) before and after a chat conversation with the suicide prevention helpline in the Netherlands (113 Suicide Prevention). Machine learning text analysis was used to predict help seeker scores on these factors. Using 2 approaches for interpreting machine learning models, we identified text messages from helpers in a chat that contributed the most to the prediction of the model.

Results: According to the machine learning model, helpers' positive affirmations and expressing involvement contributed to improved scores of the help seekers. Use of macros and ending the chat prematurely due to the help seeker being in an unsafe situation had negative effects on help seekers.

Conclusions: This study reveals insights for improving helpline chats, emphasizing the value of an evocative style with questions, positive affirmations, and practical advice. It also underscores the potential of machine learning in helpline chat analysis.

(JMIR Ment Health 2024;11:e57362) doi:10.2196/57362

KEYWORDS

suicide; suicidality; suicide prevention; helpline; suicide helpline; classification; interpretable AI; explainable AI; conversations; BERT; bidirectional encoder representations from transformers; machine learning; artificial intelligence; large language models; LLM; natural language processing

Introduction

Driving Factors for This Study

Worldwide, helplines have been set up to answer thousands of people with suicidal thoughts every day. With technology advancing and the internet having become a big presence in daily life, helplines can now often also be contacted online through chat services. An important question that is yet to be answered regarding helplines is what counselling approach is effective to take. Helplines are often anonymous, thereby making it difficult for evidence-based research, and little is still known. Several studies have been conducted on the Crisis Text Line to identify the characteristics of help seekers and their perception of the helpline's effectiveness [1,2]. Furthermore, Gould et al [3] examined call reports of help seekers calling helplines in the National Suicide Prevention Lifeline network. In studies by Mokkenstorm et al [4] and Mishara et al [5], helpline chat logs were annotated and analyzed for gathering empirical evidence. A downside of these approaches is that manual annotation of chat logs is often time-consuming work, and not a lot of available data are left unused.

In recent years, significant advancements have been made in the field of natural language analysis. Deep learning models such as transformers have enabled more effective use of big data [6]. Furthermore, bidirectional encoder representations from transformers (BERT) models [7] have made transfer learning a more viable practice. These models provide an opportunity to perform a large-scale analysis of helpline chat data. By training a model and using interpretation methods to view salient conversation features, it is possible to receive an indication of what this model thinks is important to do or not to do in a helpline chat conversation. This can be used to possibly support current findings or lead to new insights.

113 Suicide Prevention, the national suicide prevention helpline in the Netherlands, uses prechat and postchat conversation questionnaires to assess the help seeker's mental well-being [8], that is, questions related to associated suicide risk factors such as hopelessness, entrapment, perceived burdensomeness, and thwarted belongingness. In this study, we used these risk factor data as an indication of the positive outcome of the intervention. In this way, we were able to gather large amounts of labeled data without the need for manual annotation. From these data, a classification model could be trained to predict chat outcomes based on the content of the chat conversation. We assumed a lower score indicated that the help seeker was less distressed, and therefore, the conversation was labeled as positive and negative otherwise. A better understanding of what contributes to a positive conversation could help inform helplines and possibly result in actionable recommendations for helpline policy.

However, this approach requires addressing 2 main challenges. First, similar to the sentiment analysis of large documents, a decent level of accuracy can be difficult to achieve [9]. The main limitations of transformers are very long-range dependencies because self-attention scales quadratically in the length of the sequence. The $O(n^2)$ time and memory complexity means that the text must be truncated before it can be encoded by the model, and truncation means information will be lost. The long length of the conversations in a crisis line easily leads to a lot of loss of information. Second, the model should be interpretable, such that insights can be gained from the relation of the text content to the classification, that is, which parts of the conversation have more impact on the model's output.

Both challenges can be addressed using hierarchical models [10,11]. By first passing a subset of the sequence through the model, a representation for that subset can be learned. A text-based chat conversation can quite easily be segmented into individual messages or groups of concurrent messages to use as the first level in the hierarchy. At a second level of hierarchy, another sequence-based model can use the input of the first level to produce the final prediction.

In the domain of text analysis for health care, several applications of transformers have been used to gain insights into health care text data. Gao et al [12] found that pretrained BERT models did not outperform simpler methods for medical document classification. The simpler methods consisted of a convolutional neural network and a hierarchical self-attention network, which had similar performance while having fewer learnable parameters. Ilias and Askounis [13] used local interpretable model-agnostic explanations (LIME) to find influential words of BERT classifications of dementia transcripts.

In this research, we trained and compared a hierarchical document classification approach with the goal of gaining insights into the quality of helpline text-based chat conversations. The first level of the hierarchy leverages pretrained BERT network to obtain embedding representations of chat messages. We theorized that a shallow second level, for example, the message level, would allow for an easier extraction of salient chat messages for interpretation. We tested several different approaches for the message level encoding: a baseline mean pooling layer, a weighted average model with conversation participant masking, a long short-term memory (LSTM) model [10], and a transformer encoder model [10,11]. We compared these models to 3 models that did not use a hierarchical approach. Afterward, we ranked salient messages for improved and not improved scores after helpline conversations by using the best performing model. The salient chat messages where then labeled based on motivational interviewing concepts by 2 experts from the helpline.

Background

Many state-of-the-art language models that have been developed in recent years rely on transformers. First introduced by Vaswani et al [6], transformers leverage the self-attention mechanism to create long-range connections in a sequential input. This mechanism uses scaled dot-product attention (equation 1).

$$\text{Attention}\ (Q,K,V) = \text{softmax}\ (QK^T/\sqrt{d_k})V \quad \textbf{(1)}$$

The projections of the input sequence Q and K are used to compute a weighted average of the final projection V. To prevent the weights from getting too large, in turn causing the gradients to become too small, they are scaled by the dimension d_k of the input sequence. In the original paper, the transformer was developed as an encoder-decoder network for the task of language translation. Devlin et al [7] adapted the encoder section of the transformer to create high-quality embeddings. Dubbed BERT, this network, or variants thereof were applied to obtain state-of-the-art performance on many natural language processing tasks [14]. Many pretrained variants of these networks have been made available. For Dutch language tasks, there are 2 variants trained: Bertje, based on the BERT network and RobBERT, based on the RoBERTa network. Due to the nature of the attention mechanism, a straightforward assumption to make is that the attention weights directly relate to token importance. However, this assumption has been frequently questioned. Serrano and Smith [15] found that attention weights only noisily predict importance.

Jain and Wallace [16] argue attention weights do not provide explanations, while Wiegreffe and Pinter [17] in turn challenged their claims. They argue that there is a time and a place for it and provided tests to determine when attention can be used in such a way. A frequently suggested alternative to attention weight as an importance metric is gradient-based saliency [18,19]. However, even saliency maps have limitations [19]. LIME [20] is a general method for explainability that can also be applied to natural language processing. LIME generates explanations for complex models by locally approximating their behavior with simpler models.

Several adaptations of the transformer method have been proposed to deal with the issue of classification of long documents. Longformer got around the $O(n^2)$ time and memory complexity by using windowed attention, combined with a limited number of global task-specific tokens [21]. An alternate approach is used with hierarchical networks [9,11]. By first computing a fixed representation for a smaller section of the sequence, these representations are used as input for another sequence-based approach. In this method, a sequence has to be split up in some way. Often, paragraphs are used as the delimiter; however, in the case of conversations, a message or utterance could also be appropriate. Lu et al [11] used a hierarchical BERT classification method to also extract salient sentences from documents for improved explainability of the model decisions. In this paper, we use this concept for helpline conversations. As we previously mentioned, saliency from model parameters is a contested approach in literature. Because of this, we included a weighted average approach in our comparison of the hierarchical models as a shallow alternative.

Methods

Task Definition

We modelled the problem of predicting the outcome of a text-based helpline chat conversation as a binary classification task. We compared the scores of the questionnaire before and after the conversation. The classification outcome was defined as whether the help seeker's score on the questionnaire for suicide risk factors improved or did not improve.

Data Collection

The data consisted of chat conversations of a suicide prevention helpline. Between August 2021 and January 2023, help seekers (N=6903) of this helpline were asked to fill in a short questionnaire on suicide risk factors before and after the conversation with a counsellor. Table 1 [22-28] lists all the items in the questionnaire. To find the class for each conversation from the questionnaire output, we performed the following: we summed the scores of all the items for the prechat and postchat questionnaires. If the summed score for the postchat questionnaire is strictly lower than that of the prechat questionnaire, then we label a conversation as improved, and we label a conversation as not improved otherwise. Conversations that already started at the best possible value for the questionnaire before having the conversation were left out of the data set. Conversations in the suicide prevention helpline also included a triage, where the help seeker was screened for safety. The triage part of the conversation was left out of the data set as well. Without the triage, the average number of messages per conversation was 64.64 (SD 34.81). Due to a large class imbalance between improved and not improved pre-post scores for conversations, we rebalanced the data. Randomly, samples from the larger positive class were removed so that it matched the size of the negative class. The resulting initial data set consisted of 6000 chat conversations. We used an 80:20 split for train and validation sets. We used 903 conversations as the test set, which were obtained during training and validating. This test set was also used for the explainability approaches.

Table 1. Items of the prechat and postchat questionnaires.

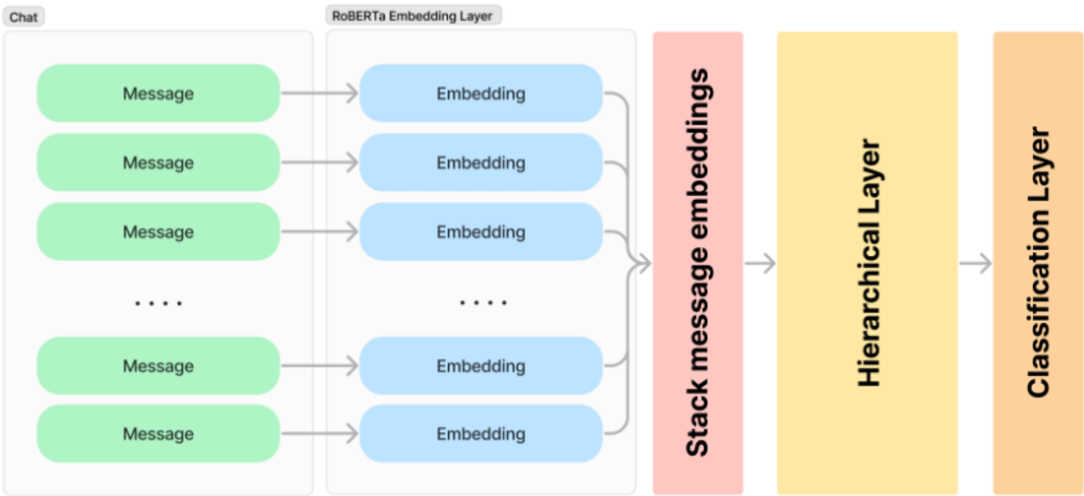
Variable	Item	Reference
Suicidal ideation	I feel the urge to kill myself	[22]
Unbearable psychache	I can't take my pain anymore	[23]
Hopelessness	I feel hopeless	[24]
Defeat	I feel that I have given up	[25]
Entrapment	I feel trapped	[26]
Perceived burdensomeness	I am a burden to others	[24]
Thwarted belongingness	I feel like I do not belong	[27]
Desire to live	I have the desire to live	[24]
Capability for suicide	I could kill myself if I wanted to	[28]

Hierarchical Approach

The individual messages were embedded using a pretrained RoBERTa network called RobBERT [29]. This network was subsequently fine-tuned on the chat conversations by using a triplet-loss strategy. The models that are described in the remainder of this section used this network to embed individual chat messages first. A message embedding was created using a mean pooling layer, resulting in a matrix M of dimension $b \times l \times d$, where l is the length of the longest sequence, d the embedding size of the pretrained network, and b the batch size.

The message embeddings for each conversation were then stacked for the hierarchical step. In the hierarchical step, we convert M to matrix C of dimension $n \times d$, which is then used in the final binary classification layer. Figure 1 shows a visual overview of this hierarchical approach.

Figure 1. Overview of the hierarchical classification of chats.



Weighted Average

To improve explainability, we used a simpler adaptation of the attention mechanism. The weighted average is described in equation 2.

$$\text{Weighted average } (C) = \text{softmax } ((CW_k^T + b_k)^T)(CW_v^T + b_v) \quad \textbf{(2)}$$

Here, C is of dimension $n \times d$, where n is the number of messages and d the embedding size. W_k and W_v are learnable weight matrices of dimensions $1 \times d$ and $d \times d$, respectively. This approach can also be described as a simplified version of dot product attention, where only a single class token attends to the sequence. This removes the need for the projections Q and K. This weighted average results in a d dimensional vector, which is used as input for the final feedforward layer for classification. Because we were also interested in the speech of the counsellor in particular, 1 additional adaptation we made was the inclusion of participant masking. Each weighted average is conditional on the sender. So, in a conversation, each weighted average only considered the messages of each participant. This was done by using multiple weighted averages and masking the logits of the weights for the weighted average, which corresponded to each participant. As is common in transformer models, we also used multiple heads, which meant the model created multiple weighted averages. The final heads were then concatenated and projected to a classification output.

Before the message embeddings were combined into the weighted average, the weights were first masked. We created 2 masks: one for only the counsellor message and one for only the help seeker messages. This resulted in the weighted average only being an average of either the counsellor or help seeker. The counsellor and help seeker each had the same number of heads.

Other Hierarchical Models

We also applied the same hierarchical method of embedding the chat messages and hierarchically classifying these shorter inputs with 3 other methods. We applied a 4-layer LSTM [30] on message embeddings. We also applied 4 transformer embedding layers. A trainable class vector was concatenated to each sequence, which was pooled as the output. The final method applied a simple average of all message embeddings over the sequence dimension. The outputs of these models were fed in the same feedforward layer as the weighted average method.

Baseline Models

We applied several additional preprocessing steps for the baseline models. All words were also lowercased, lemmatized, and all special characters and punctuations were removed as well as stop words. During tokenization, we limited the number of tokens to 2000. We vectorized the chat conversations by using TF-IDF (term frequency–inverse document frequency) [31]. Finally, each embedded conversation was trained on a support vector machine [32]. Furthermore, the Dutch BERT model, RobBERT, was used as another baseline model. Because it has a maximum length of 512 tokens for the text input, the chats were truncated at the maximum length. Two RobBERT models were fine-tuned—one where the start of the conversation was truncated and one where the end was truncated.

Explainability

To gain insights into the workings of the network, we employed 2 techniques. First, we used the weights of the weighted average model. The assumption was that messages with a higher weight were of higher importance to the final decision and therefore, more important to the result of the conversation. As a second technique, we applied LIME [20] to the models. For this approach, we left out counsellor messages one at a time to compute the difference in loss. A larger difference indicates more importance to the classification.

The process was as follows: first, we did a feedforward pass of the chat conversations in the test set through the model. Second, we recorded the resulting logits. We selected the chats with a logit score of less than 0.2 and higher than 0.8 that were also correctly classified. Third, we used LIME on just the hierarchical part of the model (yellow part in Figure 1), where each message is a feature, to obtain feature importance scores for each message. We also recorded the weights W_k from the weighted average approach. We then selected all the messages that were at least 1 SD above the mean as impactful messages for each of the 2 explainability methods. Each chat that was selected in step 2 was annotated with which the messages were deemed important by the model. We then prompted 2 clinical psychologists operating in the helpline. As this approach is exploratory, we did not formulate a hypothesized set of possible behaviors beforehand for the experts to choose from. However, the 2 experts were trained in motivational interviewing, which was the main paradigm used in the helpline for the chat conversations [8]. Through this lens, they were asked to annotate each impactful message and indicate the behavior conform to this method. From these labels, the most frequently occurring labels were compiled.

Ethics Approval

This study protocol was performed in accordance with the relevant guidelines. This study was reviewed and approved by the Medical Research Ethics Committee of Amsterdam Universitair Medische Centra (registration: 2021.0447).

Results

Model Performance

Table 2 shows the performance scores on a held-out test set. The hierarchical weighted average model was the best performing model with an accuracy of 0.683 and the highest F_1-score of 0.688. This was closely followed by the hierarchical LSTM model, which had an accuracy of 0.672. The results for the hierarchical transformer model and hierarchical average did not perform as well, with accuracies of 0.638 and 0.640, respectively. The support vector machine model had an accuracy of 0.638, which was lower than that of the hierarchical models. The 2 truncated versions of the BERT model had accuracies of 0.570 and 0.629 for the truncated end and truncated start models, respectively. This suggests that the information in the truncated text was most likely insufficient compared to the hierarchical models that do not have the ability to attend to words from different messages, but overall have more information available. Overall, our results suggest that the hierarchical LSTM and hierarchical weighted average outperformed other models for the task classifying suicide helpline chat conversations.

Table 2. Model performance on the test set of the suicide chat classification task.

Model	Accuracy	Precision	Recall	F_1-score
Support vector machine	0.638	0.635	0.632	0.633
BERT[a] truncated end	0.570	0.556	0.699	0.620
BERT truncated start	0.629	0.605	0.743	0.667
Hierarchical average	0.640	0.621	0.721	0.667
Hierarchical weighted average	0.683	0.679	0.697	0.688
Hierarchical LSTM[b]	0.672	0.674	0.668	0.671
Hierarchical transformer	0.638	0.612	0.754	0.676

[a]BERT: bidirectional encoder representations from transformers.

[b]LSTM: long short-term memory.

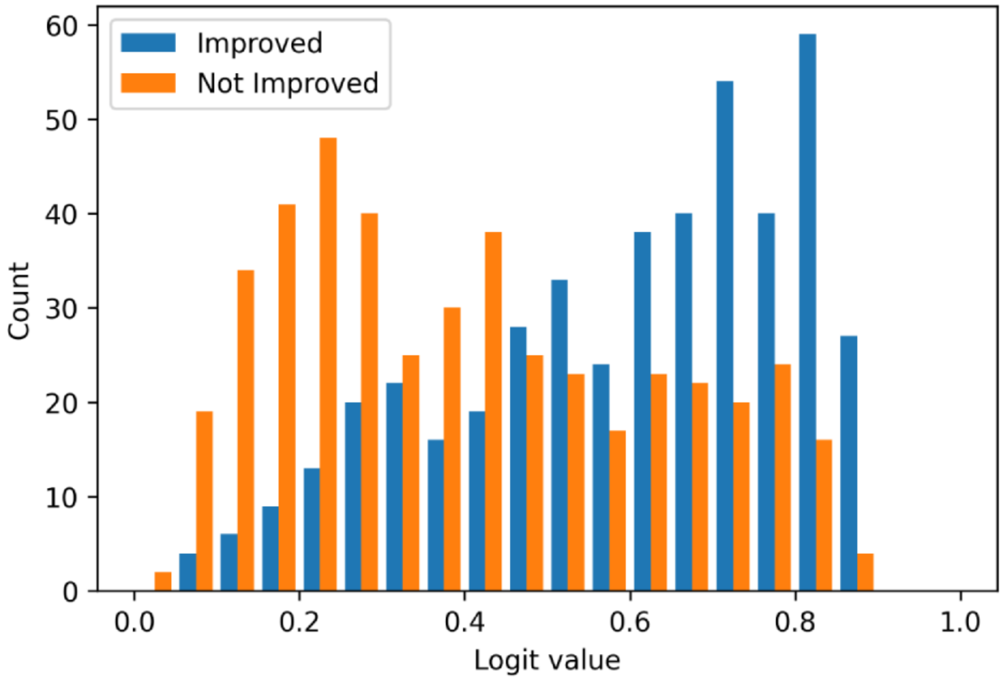
Model Explanations

The weighted average model was overall the best performer in terms of accuracy and F_1-score. Because it was more interpretable than the hierarchical LSTM and the BERT networks, it was the obvious choice to extract explanations. The explanations were compiled from a test set only, and a subselection of the data was made, with only the correctly classified samples and where the model was confident in its output. This confidence was measured through the logit output of the model. A logit value close to 1 corresponds to the classification of a chat conversation that resulted in an improved score and closer to 0 for the opposite case. Figure 2 shows a histogram of the logit outputs of the model for this test data set. Two peaks can be seen for the correctly classified samples. This shows that while there are still chats that are difficult to categorize, there is a clear set of chats where the model is confident for either class. We can also see that the model was slightly more confident for the not-improved chat conversation than for improved ones. We chose values below 0.2 and above 0.8 as subsets to extract the explanations.

Using the weights from the hierarchical weighted average model, we compiled the most influential messages from the counselors.

The messages selected as influential were messages with a weight 1 SD above the mean. The weights used for this purpose were from the heads that were masked for the help seeker and thus only contained nonzero values for counselor messages. Furthermore, LIME was used in combination with the weighted average model to obtain explanations. This section describes the outcomes based on observations by the authors and 2 senior psychologists from the helpline.

Figure 2. Histogram of logits for the hierarchical weighted model.



Conversations Without an Improved Suicide Risk Score

For conversations that did not improve, we identified 3 distinct situations that emerged from the influential messages. The first and the most common situation was when a conversation ended prematurely. In such cases, the counselor would typically try to redirect the help seeker to alternative channels for assistance, such as a general practitioner, a different helpline, or emergency services. Alternatively, in some cases, the counselor suggested to contact the helpline at another time or to apply for a web-based therapy service. The second situation involved messages where the counselor was unable to respond promptly to the help seeker due to a high volume of ongoing conversations. The counselor apologized to the help seeker for the delay and sometimes mentioned that the helpline was particularly busy and that the counselor was dealing with multiple help seekers simultaneously. The third situation included the counsellor not connecting properly with the help seeker. This was expressed in the use of macros and lists. The macros would often include a standardized set of options for the help seeker to consider or sometimes a set of websites and resources to visit. Sometimes this was also expressed as the counsellor not properly listening to the help seeker. Table 3 shows examples of 5 messages from conversations that did not improve the score on the postchat questionnaire.

Table 3. Messages that were the most influential according to the model in the 5 conversations for which the model was the most confident that the conversation had not improved the score on the postchat questionnaire.

Message	Details
Example 1	*…Okay help seeker. To cope with these thoughts at this moment, there are a few options: seeking distraction, contacting someone in your environment, expressing emotions in a pleasant way (creatively, through sports, writing, etc), and relaxation exercises (eg, mindfulness, breathing exercises).*
Example 2	*…At [url], there are tips on how to handle moments like these. Maybe you can take a look to see if there's something that feels right for you to do now?*
Example 3	*…Our chat is exclusively for people with suicidal thoughts. I suggest you contact other helplines if you want to talk about what's on your mind now.*
Example 4	*…That's not the case, help seeker. That confusion must be troubling and occupying your thoughts. I wish I had an answer for you, but I don't have one right away. I notice that you are sharing more, so perhaps it's a good idea to contact the listening line. I wish you a lot of strength, help seeker!*
Example 5	*…I notice that I'm not quite sure what you would like to discuss during this conversation. I want to help you, but I don't really know what you want to talk about.*

Conversations With an Improved Suicide Risk Score

The salient messages from conversations with improved scores had a wider range of responses compared to those that did not improve scores. However, we identified 2 frequently recurring situations in conversations that showed an improved score. In the first situation, the counselor provided positive reinforcement to the help seeker. During these conversations, the counselor would typically use supportive language, such as showing empathy, offering praise, and expressing happiness for the help seeker. In the second situation, the counselor expressed involvement. For example, the counselor would think along with the help seeker and provide concrete solutions to the help seeker. These solutions could include specific actions or resources tailored to the help seeker's individual situation. The counselor would provide the help seeker with practical steps that could be taken or resources specific to the help seeker's situations. Lastly, the 2 less often recurring situations were situations where the counsellor would ask open-ended questions as well as show respect for the autonomy of the help seeker by asking what they wanted to do. Table 4 shows examples of 5 messages from conversations that improved the score on the postchat questionnaire.

Table 4. Messages that were the most influential according to the model in the 5 conversations for which the model was the most confident that the conversation improved the score on the postchat questionnaire.

Message	Details
Example 1	…*I find it incredibly admirable that you have clarified this for yourself and that you are going to discuss it with your therapist.*
Example 2	…*You're amazing for taking this on right away.*
Example 3	…*Okay, it's good that you want to try, help seeker. Can I assist you further with something else, or is there something specific you'd like to talk about?*
Example 4	…*I see that you mentioned you are not in treatment. You said that a lot has happened over the past few months that is troubling you and that you tried to drown yourself tonight. I think it's important that you talk to someone about this so you can get the help you need because you deserve it, dear help seeker.*
Example 5	…*Are you familiar with relaxation exercises? These can usually help with panic attacks as well.*

Discussion

Principal Findings

This study compares the performance of different models for classifying suicide helpline chat conversations and found that the hierarchical weighted average model had the best performance. This study also extracted explanations from the hierarchical weighted average model and identified 3 distinct situations for conversations that did not improve and 2 clear recurring situations for conversations that showed an improved score. The results showed that the model had an easier time determining when a conversation would not lead to an improvement in the risk factors. This was also apparent in the explanations where clear and easy distinctions in the output could be made, whereas this was not as easy to do in the case of positive examples.

The research by Mishara et al [5] found that collaborative problem-solving significantly predicted positive outcomes in helpline calls. In line with these findings, our study shows that messages with positive reinforcement and concrete solutions contributed to positive outcomes in chat conversations. Furthermore, Côté and Mishara [33] found through qualitative analysis that reinforcing a strength or a positive action was a significant predictor for increased scores on a pre-post questionnaire in a text message helpline setting. This is in line with our finding that positive reinforcement was a frequently occurring impactful message. In a qualitative study, Gilat and Rosenau [34] analyzed volunteers' perspective of effective methods in helpline conversations. Among their findings, they identified practical advice as an effective strategy. Building rapport was another aspect of note that their study identified. Because building rapport is highly specific to the individual, this might be something our method was not able to generalize and pick up. However, positive reinforcement could also have been a contributor to building rapport. Overall, these findings highlight the potential of using machine learning models to analyze suicide helpline chat conversations and provide insights into the most influential messages. This allows helplines to be more informed and possibly enable them to improve helpline quality.

Limitations

Although this study sheds light on influential messages in suicide chat conversations, there are 3 key limitations to be considered. First, there are general limitations of machine learning. The classification task was found to be difficult, as indicated by the 68% accuracy rate that was achieved. This suggests that the current models have room for improvement. There might still be relationships that the current models were not able to capture. It could also be that there is considerable noise in the data set because the outcome measures were self-reported by help seekers, which might not be equally reliable for every help seeker. Furthermore, the indicated influential messages might be messages that are not the main cause but rather a result of a different action. For example, we saw multiple situations where the counsellor expressed gratitude for a compliment. This was most likely the result of the help seeker being grateful for something; however, it does not necessarily mention what the help seeker was grateful for. Second, a limitation of this study is the challenge posed by modeling a large amount of text. Current methods have limitations in capturing dependencies over long ranges or in exceeding maximum memory thresholds, which was the case with the chat conversations used in our data set. Therefore, hierarchical models were used, which had the limitation that dependencies between words from different messages were not captured. Third, a limitation of using chat messages as output to determine categories of influential messages is the need for human judgement. This introduces subjectivity and the potential for bias, as different judges may interpret the same messages differently or possibly miss a connection between the different messages.

Future Work

Considering the findings presented in this study, we identified 3 potential directions for future work that could further enhance the classification and identification of influential messages in suicide helpline chat conversations. First, while larger models have the potential to improve performance, explainability needs to be considered as well. The use of larger models can sometimes lead to decreased interpretability, and it may be challenging to identify the most influential features that contribute to the classification of a message. Therefore, future research could explore the use of models such as Longformer, which are designed to handle long sequences of text through windowed attention and global attention for the class. This global attention can possibly be leveraged for explainability. Second, with additional computational resources, another potential area of research is to forgo the use of sentence embedders and input messages directly into a transformer model. This approach could potentially improve the performance of the model by better capturing individual sentences rather than relying on message embeddings that are not trained for the specific task. Third, in addition to model improvements, future research could explore additional processing techniques of influential messages, such as clustering. Clustering could be used to group similar messages together, allowing for an analysis of influential messages. This could be useful for easier identification of patterns in influential messages.

Practical Implications

Engaging with help seekers expressing suicidal thoughts while recognizing they can be better helped elsewhere is important. However, counselors should be mindful of empathetically guiding them toward the appropriate channels. It is important to keep validating their emotions and ensuring they feel supported rather than dismissed. Standardized responses from macros can be beneficial in the right circumstances. If used without having good rapport with the help seekers, they can appear distant. Being transparent with the help seeker about the use of macros is important, as well as ensuring good enough rapport between the help seeker and counselor has been established with personal responses before using standardized responses. Collaborative problem solving and building rapport are proven ways to foster better conversations. Positive reinforcement might be another method that counsellors can employ. Including positive reinforcement more regularly in their responses might be beneficial for helpline conversations.

Conclusion

This study compares the performance of different models for classifying suicide helpline chat conversations and found that a weighted average model using message embeddings performed the best. This study is unique compared to other studies that aim to gain insight into the quality and effectiveness of suicide prevention helplines. Many studies use questionnaires to evaluate implemented counseling approaches. In this study, we identified influential messages that contributed to better or worse scores on a suicide risk questionnaire through a machine learning approach. This initial application showed that we could extract explanations from the model and identified distinct situations for improvement and deterioration of help seekers’ emotional states.

Acknowledgments

We would like to thank the senior psychologists Mirjam van Driel and Maryke Geerdink for their interpretation of the impactful messages.

Authors' Contributions

SS had full access to all the data in this study and takes responsibility for the integrity, acquisition, and interpretability of the data and accuracy of the data analysis. SS drafted the manuscript and performed statistical analysis. RG and RvdM contributed to the administrative, technical, or material support. SB and SM supervised the study. All authors contributed to the study concept and design and critical revision of the manuscript for important intellectual content. The authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Gould MS, Pisani A, Gallo C, Ertefaie A, Harrington D, Kelberman C, et al. Crisis text-line interventions: Evaluation of texters' perceptions of effectiveness. Suicide Life Threat Behav 2022 Jun;52(3):583-595 [FREE Full text] [doi: 10.1111/sltb.12873] [Medline: 35599358]
2. Pisani AR, Gould MS, Gallo C, Ertefaie A, Kelberman C, Harrington D, et al. Individuals who text crisis text line: Key characteristics and opportunities for suicide prevention. Suicide Life Threat Behav 2022 Jun;52(3):567-582 [FREE Full text] [doi: 10.1111/sltb.12872] [Medline: 35615898]

3. Gould MS, Lake AM, Munfakh JL, Galfalvy H, Kleinman M, Williams C, et al. Helping callers to the national suicide prevention lifeline who are at imminent risk of suicide: evaluation of caller risk profiles and interventions implemented. Suicide Life Threat Behav 2016 Apr 04;46(2):172-190. [doi: 10.1111/sltb.12182] [Medline: 26242234]
4. Mokkenstorm JK, Eikelenboom M, Huisman A, Wiebenga J, Gilissen R, Kerkhof AJFM, et al. Evaluation of the 113 online suicide prevention crisis chat service: outcomes, helper behaviors and comparison to telephone hotlines. Suicide Life Threat Behav 2016 Aug 19;47(3):282-296. [doi: 10.1111/sltb.12286]
5. Mishara BL, Chagnon F, Daigle M, Balan B, Raymond S, Marcoux I, et al. Which helper behaviors and intervention styles are related to better short-term outcomes in telephone crisis intervention? Results from a Silent Monitoring Study of Calls to the U.S. 1-800-SUICIDE Network. Suicide Life Threat Behav 2007 Jun;37(3):308-321. [doi: 10.1521/suli.2007.37.3.308] [Medline: 17579543]
6. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. ArXiv Preprint posted online on August 2, 2023. [doi: 10.48550/ARXIV.1706.03762]
7. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. ArXiv Preprint posted online on May 24, 2019. [doi: 10.48550/ARXIV.1810.04805]
8. Janssen W, van Raak J, van der Lucht Y, van Ballegooijen W, Mérelle S. Can outcomes of a chat-based suicide prevention helpline be improved by training counselors in motivational interviewing? a non-randomized controlled trial. Front Digit Health 2022;4:871841 [FREE Full text] [doi: 10.3389/fdgth.2022.871841] [Medline: 35801118]
9. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2016:1480-1489. [doi: 10.18653/v1/n16-1174]
10. Pappagari R, Zelasko P, Villalba J, Carmiel Y, Dehak N. Hierarchical transformers for long document classification. 2019 Dec Presented at: IEEE Automatic Speech Recognition and Understanding Workshop; December 14-18; Singapore. [doi: 10.1109/asru46091.2019.9003958]
11. Lu J, Henchion M, Bacher I, Namee BM. A sentence-level hierarchical BERT model for document classification with limited labelled data. ArXiv Preprint posted online on June 12, 2021. [doi: 10.48550/arXiv.2106.06738]
12. Gao S, Qiu JX, Alawad M, Hinkle JD, Schaefferkoetter N, Yoon H, et al. Classifying cancer pathology reports with hierarchical self-attention networks. Artif Intell Med 2019 Nov;101:101726 [FREE Full text] [doi: 10.1016/j.artmed.2019.101726] [Medline: 31813492]
13. Ilias L, Askounis D. Explainable identification of dementia from transcripts using transformer networks. IEEE J Biomed Health Inform 2022 Aug;26(8):4153-4164. [doi: 10.1109/jbhi.2022.3172479]
14. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A robustly optimized BERT pretraining approach. ArXiv Preprint posted online on July 26, 2019 [FREE Full text] [doi: 10.48550/ARXIV.1907.11692]
15. Serrano S, Smith NA. Is attention interpretable? ArXiv Preprint posted online on June 9, 2019 [FREE Full text] [doi: 10.48550/ARXIV.1906.03731]
16. Jain S, Wallace B. Attention is not explanation. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) 2019 Jun:3543-3556. [doi: 10.18653/v1/N19-1357]
17. Wiegreffe S, Pinter Y. Attention is not not explanation. ArXiv Preprint posted online on September 5, 2019 [FREE Full text] [doi: 10.48550/arXiv.1908.04626]
18. Wallace E, Tuyls J, Wang J, Subramanian S, Gardner M, Singh S. AllenNLP interpret: A framework for explaining predictions of NLP models. ArXiv Preprint posted online on September 19, 2019. [doi: 10.18653/v1/d19-3002]
19. Bastings J, Filippova K. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? Proceedings of the Third Blackbox NLP Workshop on Analyzing and Interpreting Neural Networks for NLP 2020:149-155. [doi: 10.18653/v1/2020.blackboxnlp-1.14]
20. Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?": Explaining the predictions of any classifier. 2016 Presented at: The 22nd ACM SIGKDD international conference on knowledge discovery and data mining; August 13-17; San Francisco, CA p. 1135-1144. [doi: 10.1145/2939672.2939778]
21. Beltagy I, Peters M, Cohan A. Longformer: The long-document transformer. ArXiv Preprint posted online on December 2, 2020. [doi: 10.48550/ARXIV.2004.05150]
22. Kleiman EM, Turner BJ, Fedor S, Beale EE, Huffman JC, Nock MK. Examination of real-time fluctuations in suicidal ideation and its risk factors: Results from two ecological momentary assessment studies. J Abnorm Psychol 2017 Aug;126(6):726-738. [doi: 10.1037/abn0000273] [Medline: 28481571]
23. Pachkowski MC, May AM, Tsai M, Klonsky ED. A brief measure of unbearable psychache. Suicide Life Threat Behav 2019 Dec;49(6):1721-1734. [doi: 10.1111/sltb.12556] [Medline: 31074006]
24. Nuij C, van Ballegooijen W, Ruwaard J, de Beurs D, Mokkenstorm J, van Duijn E, et al. Smartphone-based safety planning and self-monitoring for suicidal patients: Rationale and study protocol of the CASPAR (Continuous Assessment for Suicide Prevention And Research) study. Internet Interv 2018 Sep;13:16-23. [doi: 10.1016/j.invent.2018.04.005] [Medline: 30206514]

25. Forkmann T, Teismann T, Stenzel J, Glaesmer H, de Beurs D. Defeat and entrapment: more than meets the eye? Applying network analysis to estimate dimensions of highly correlated constructs. BMC Med Res Methodol 2018 Jan 25;18(1):16 [FREE Full text] [doi: 10.1186/s12874-018-0470-5] [Medline: 29370770]
26. Littlewood DL, Kyle SD, Carter L, Peters S, Pratt D, Gooding P. Short sleep duration and poor sleep quality predict next-day suicidal ideation: an ecological momentary assessment study. Psychol Med 2018 Apr 26;49(3):403-411. [doi: 10.1017/s0033291718001009]
27. Hallensleben N, Glaesmer H, Forkmann T, Rath D, Strauss M, Kersting A, et al. Predicting suicidal ideation by interpersonal variables, hopelessness and depression in real-time. An ecological momentary assessment study in psychiatric inpatients with depression. Eur Psychiatry 2019 Feb;56:43-50. [doi: 10.1016/j.eurpsy.2018.11.003] [Medline: 30530103]
28. Spangenberg L, Glaesmer H, Hallensleben N, Rath D, Forkmann T. (In)stability of capability for suicide in psychiatric inpatients: longitudinal assessment using ecological momentary assessments. Suicide Life Threat Behav 2019 Dec;49(6):1560-1572. [doi: 10.1111/sltb.12547] [Medline: 30834576]
29. Delobelle P, Winters T, Berendt B. RobBERT: a Dutch RoBERTa-based language model. ArXiv Preprint posted online on September 16, 2020. [doi: 10.48550/ARXIV.2001.06286]
30. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997 Nov 15;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735] [Medline: 9377276]
31. Uther W, Mladeni D, Ciaramita M, Berendt B, Kocz A, Grobelnik M, et al. Text visualization. In: Encyclopedia of Machine Learning. USA: Springer; 2011:986-987.
32. Cristianini N, Ricci E. Support vector machines. In: Encyclopedia of Algorithms. USA: Springer; 2008:928-932.
33. Côté LP, Mishara BL. Effect of helping suicidal people using text messaging: An evaluation of effects and best practices of the Canadian suicide prevention service's text helpline. Suicide Life Threat Behav 2022 Dec;52(6):1140-1148. [doi: 10.1111/sltb.12908] [Medline: 35946317]
34. Gilat I, Rosenau S. Volunteers perspective of effective interactions with helpline callers: qualitative study. British Journal of Guidance & Counselling Informa UK Limited 2011 Aug:325-337. [doi: 10.1080/03069885.2011.567327]

Abbreviations

BERT: bidirectional encoder representations from transformers
LIME: local interpretable model-agnostic explanations
LSTM: long short-term memory
RoBERTa: robustly optimized bidirectional encoder representations from transformers pretraining approach
TF-IDF: term frequency–inverse document frequency

Edited by J Torous; submitted 14.02.24; peer-reviewed by L Lilias, T Zhang, CS Soma, IT Hwang; comments to author 30.05.24; revised version received 25.07.24; accepted 11.08.24; published 26.09.24.

Please cite as:
Salmi S, Mérelle S, Gilissen R, van der Mei R, Bhulai S
The Most Effective Interventions for Classification Model Development to Predict Chat Outcomes Based on the Conversation Content in Online Suicide Prevention Chats: Machine Learning Approach
JMIR Ment Health 2024;11:e57362
URL: https://mental.jmir.org/2024/1/e57362
doi:10.2196/57362
PMID:

Generation of Backward-Looking Complex Reflections for a Motivational Interviewing–Based Smoking Cessation Chatbot Using GPT-4: Algorithm Development and Validation

Ash Tanuj Kumar[1], BSc, MASc; Cindy Wang[1]; Alec Dong[1]; Jonathan Rose[1,2], PhD

1

2

Corresponding Author:
Jonathan Rose, PhD

Abstract

Background: Motivational interviewing (MI) is a therapeutic technique that has been successful in helping smokers reduce smoking but has limited accessibility due to the high cost and low availability of clinicians. To address this, the MIBot project has sought to develop a chatbot that emulates an MI session with a client with the specific goal of moving an ambivalent smoker toward the direction of quitting. One key element of an MI conversation is reflective listening, where a therapist expresses their understanding of what the client has said by uttering a *reflection* that encourages the client to continue their thought process. *Complex* reflections link the client's responses to relevant ideas and facts to enhance this contemplation. Backward-looking complex reflections (BLCRs) link the client's most recent response to a relevant selection of the client's previous statements. Our current chatbot can generate complex reflections—but not BLCRs—using large language models (LLMs) such as GPT-2, which allows the generation of unique, human-like messages customized to client responses. Recent advancements in these models, such as the introduction of GPT-4, provide a novel way to generate complex text by feeding the models instructions and conversational history directly, making this a promising approach to generate BLCRs.

Objective: This study aims to develop a method to generate BLCRs for an MI-based smoking cessation chatbot and to measure the method's effectiveness.

Methods: LLMs such as GPT-4 can be stimulated to produce specific types of responses to their inputs by "asking" them with an English-based description of the desired output. These descriptions are called *prompts*, and the goal of writing a description that causes an LLM to generate the required output is termed *prompt engineering*. We evolved an instruction to prompt GPT-4 to generate a BLCR, given the portions of the transcript of the conversation up to the point where the reflection was needed. The approach was tested on 50 previously collected MIBot transcripts of conversations with smokers and was used to generate a total of 150 reflections. The quality of the reflections was rated on a 4-point scale by 3 independent raters to determine whether they met specific criteria for acceptability.

Results: Of the 150 generated reflections, 132 (88%) met the level of acceptability. The remaining 18 (12%) had one or more flaws that made them inappropriate as BLCRs. The 3 raters had pairwise agreement on 80% to 88% of these scores.

Conclusions: The method presented to generate BLCRs is good enough to be used as one source of reflections in an MI-style conversation but would need an automatic checker to eliminate the unacceptable ones. This work illustrates the power of the new LLMs to generate therapeutic client-specific responses under the command of a language-based specification.

(JMIR Ment Health 2024;11:e53778) doi:10.2196/53778

KEYWORDS

motivational interviewing; smoking cessation; therapy; automated therapy; natural language processing; large language models; GPT-4; chatbot; dialogue agent; reflections; reflection generation; smoking; cessation; ChatGPT; smokers; smoker; effectiveness; messages

Introduction

Background

Smoking cessation therapists have long used the motivational interviewing (MI) talk therapy to guide clients toward positive behavioral change [1]. MI engages clients in a structured conversation that encourages them to contemplate their behavior more deeply and motivates them to change it. MI has been shown to be successful in helping clients reduce or quit their smoking habits [2], but the availability of MI-trained clinicians is limited to hospitals and medical centers, and MI therapy is usually only initiated after a smoking-related health issue occurs [3]. These restrictions make it difficult for smokers to access therapy outside of medical centers and occur too late to have a preventative effect.

Our research seeks to automate the therapist side of an MI conversation which, if successful, could broaden access to care at a population level. We have been developing a chatbot, called MIBot [4], whose purpose is to move ambivalent smokers toward the direction of quitting. MIBot is being developed by an interdisciplinary research collaboration among expert MI-trained clinicians, social scientists, and computer engineers. The initial version of the MIBot chatbot guides the client through a fairly simple MI conversation by combining scripted interactions with context-specific responses generated by natural language models, based on elements of the MI approach.

The focus of the initial version of the MIBot chatbot is on one core skill of MI: reflective listening [1], in which the chatbot provides reflections on what the client has most recently said. In general, reflections are meant to express the therapist's current understanding of the client's most recent response and invite the client to continue further contemplation of their behavior. Reflections can be simple or complex [1]. A simple reflection rephrases a client's response, sending the message that the response was understood and inviting the client to continue. A complex reflection attempts to infer relevant information about the client from the client's utterance by linking the client's response to relevant facts or ideas. A good quality complex reflection may further infer something about the emotional state of the client through their utterance.

In a complex reflection, when these relevant facts come from a client's earlier responses in the conversation, we call this a backward-looking complex reflection (BLCR). Preferably, a BLCR does not simply summarize all the past conversational information in order but is composed of the information that is sensible for the context. Textbox 1 shows an example of a conversation in which the final statement by the therapist is a BLCR.

Textbox 1. Example motivational interviewing conversation in which the last utterance by the therapist is a backward-looking complex reflection.

Therapist: What is one thing you like about smoking?

Client: It makes me have less stress and keeps me connected to my friends.

Therapist: What is one thing you dislike about smoking?

Client: It leaves bad breath.

Therapist: What is one thing about your smoking addiction that you would like to change?

Client: I would like to reduce smoking.

Therapist: [backward-looking complex reflection] It seems like you want to reduce your smoking, which might help your concern about bad breath

The initial MIBot chatbot [4] only generates reflections using the client's most recent utterance and does not make use of prior utterances. The ability to generate BLCRs can expand the chatbot's options for generating context-appropriate complex reflections.

The goal of this work is to develop and evaluate a method to automatically generate BLCRs given a prior conversation. It has become possible to do this kind of generation through recent dramatically powerful advancements in natural language processing [5], and more specifically the most recent large language models (LLMs) from GPT-3.5 and later [6-8].

LLMs are language models which take text as input and generate textual output. GPT-4, an LLM introduced in March 2023, has significantly improved capability to generate text to satisfy particular requirements compared to previous LLMs [6-9]. One way to use GPT-4 is to write a *prompt*, which is a language-based instruction that literally tells the model the processing that is desired [9]. This processing is potentially anything that can be described in language, which is a truly remarkable, new capability that will have many applications. We describe a method for developing the prompts needed to "tell" the model to create BLCRs.

This paper is organized as follows: the *Prior Work* section introduces MI, GPT-4, and the relevant parts of the MIBot project that we build on. The *Methods* section describes the prompt developed to generate a BLCR, the specific structure of the input to GPT-4, the rating scale developed to assess when a BLCR is acceptable, the experimental procedure to test the acceptability of BLCRs generated by the prompt, and the data used to test this procedure. The *Results* section provides the evaluation, and the *Discussion* section interprets the results of the experiment and lists limitations. The *Conclusions* section suggests avenues for further work.

Prior Work

Motivational Interviewing

MI is a therapeutic technique in which a therapist engages in a conversation to guide and motivate clients who are ambivalent about their behaviors to move toward changing them [1]. These guided conversations use 4 MI core skills: asking open-ended questions, providing reflections, affirmations, and summarization. In an MI conversation, the therapist will typically begin with an open-ended question, listen to the client's response, and reply with 1 of the other 3 core skill types, depending on the circumstances and the direction the therapist wishes to guide the conversation.

While all 4 core skill types are integral to a successful MI, we focus on the role of reflections and the related reflective listening. Reflective listening requires the therapist to listen to what the client has most recently said and formulate a response—called a reflection—that displays the therapist's understanding while also guiding the conversation. The content of a reflection depends on the current context of the conversation. Reflections can be divided into 2 types: simple reflections and complex reflections. Simple reflections restate the client's response, typically using different words, so that the therapist and client can establish that they are on the same page. Complex reflections allow the therapist to link what the client has most recently said to other facts or information about the client's life and emotional state, usually providing some kind of inference. Complex reflections are used to guide the conversation toward new topics.

MI has been shown to be a successful therapy for moving clients toward reducing their smoking habits [2], and reflections in particular have been correlated with high perceived support for patient autonomy in MI sessions [10].

LLMs and GPT-4

LLMs are digital models of natural language that are able to generate text from an input by autoregressively predicting the next word in a given sequence [6]. These models learn how to predict semantically and syntactically reasonable words by being trained to "fill in the blanks" on large amounts of diverse human-written text, which encompass questions and answers, web-based conversations, informative articles, and other kinds of digitized text. The wide range of data that LLMs are trained on have made them effective systems for generating solutions to various problems in the domain of natural language processing, such as answering questions, summarizing long text, and conversational dialogue generation [7].

The GPT (Generative Pretrained Transformer) family of LLMs has proven to be state of the art in a number of general-purpose tasks [8]. ChatGPT and the related GPT-4 [9] model can generate human-like text and answer questions correctly to the point that it has successfully passed m any professional and academic examinations [9].

Due to the size of the model and the large amounts of human-produced textual data it is trained on, 1 emergent characteristic of the GPT-4 is its ability to answer questions and interpret human-readable text to follow instructions. This has led researchers to try and directly "ask" GPT-4 to generate some kind of desired text given some input. The study of ways to ask GPT-4 to generate desired text is a newly emerging field called prompt engineering. Thus, this asking process is called *prompting* a GPT-4 model, and these "asks" are typically called *prompts* [9,11].

A prompt usually consists of a *request* of the model to generate or process some desired text, usually followed by requirements that the generated text must satisfy or instructions that tell the model how to generate this text [11]. The request can optionally be followed by an *input*, with the goal that the model will use the input to process and generate the requested text (Table 1).

Table . Prompt example and generated result. The request is the first sentence, and the input is the italicized second line. The request and input message can be tested live on the OpenAI playground by copy and pasting the entire prompt text [9].

	Prompt example and GPT-4–generated result
Prompt (request and input)	Given the Keywords below, write a paragraph that incorporates them into a story about a princess on the moon. Keywords: spaceship, cat, star, Pluto, Brazil
GPT-4	*Once upon a time, there lived a princess, unlike any other, on the moon. She had a spaceship that allowed her to travel across galaxies, making frequent visits to far-off planets. It was on one such trip that she found a stray cat wandering amidst the frozen nitrogen landscapes of Pluto, looking as lost and lonely as a star flickering bleakly in an infinite sea of black. She couldn't bear its sight and decided to take this feline friend along on her adventures. To cheer the cat up, they took spontaneous trips to Earth, her favorite being Brazil. Their spaceship would hover quietly over the vibrant Amazon forest as the princess and her cat would gaze down at the colorful montage of life. The brilliant night sky with millions of stars reflected in the cat's eyes reminded the princess that, just like the starlight, even the loneliest souls could brighten up the darkest corners of the universe.*

GPT-3– and GPT-4–based prompting has been shown to be highly effective in generating text to solve various natural language processing tasks [9,11,12] and has already found applications in a diverse set of technical fields. However, a prompted GPT model does not always produce factually correct answers [9,11,12]. In addition, a prompted GPT model is not deterministic, and a single prompt may produce different texts each time that a prompt is used to generate a completion [11,13]. Recent research on prompt engineering has produced new methods to structure prompts for generating satisfactory texts [13].

The ability to prompt is not restricted by the architecture of GPT-3 or GPT-4. Prompting is possible with any LLM of similar structure, and the difference in output depends on how much knowledge and prediction capability has been retained by an LLM. Thus, while our work specifically used GPT-4, this paper's method can be used with any LLM, including future improvements on GPT-4, and we will indicate this by referring to LLMs broadly in our methods and discussions.

Existing MI Smoking Cessation Chatbots and the MIBot Project

The research and development of MI-based chatbots across several therapeutic domains remains an open problem, with numerous approaches incorporating different natural language processing techniques, and nothing yet deployed in a commercial or therapeutic context for mass adoption. For MI focusing on smoking cessation, several research teams have independently developed chatbots that have been tested and evaluated on experimental study participants. Our particular work has focused on an early step in smoking cessation, which is moving ambivalent smokers toward the decision to quit smoking.

Almusharraf et al [14] designed an MI chatbot, which used predefined answers in a scripted conversation and measured its effectiveness on clients' confidence to quit smoking with an 11-point scale. After testing this method on 97 participants, they found that the average confidence among clients to quit smoking increased by 0.8 (P<.001 via paired 1-tailed t test) 1 week after the conversation. The scripted nature of these MI conversations, with answers not unique to clients' responses, was suggested as a future point of improvement to investigate further.

Independently, He et al [15] sought to investigate whether chatbots using MI techniques had any differing effects from neutral chatbots. They designed 2 chatbots—an MI-based chatbot and a neutral, affirming chatbot—and found that while there were no significant differences in clients' reception of the 2 chatbots, both chatbots increased the clients' motivations to quit smoking. The conclusions of He et al [15] combined with the results of Almusharraf et al [14] indicate that nonscripted responses from chatbots may be better received.

The text produced by generative models are an alternative to scripted responses, and Shen et al [16] displayed how generative models could generate reflections dependent on context. Using a GPT-2–based architecture, they created unique, context-dependent generative responses by incorporating a combination of client and therapist utterances from an existing dialogue history, and drawing from a database of previous transcripts to help select between context-relevant responses based on semantic similarity. These generated reflections were compared to a seq2seq model baseline, an older model of conditional text generation that is not LLM based, and human evaluation using a 5-point Likert scale for absolute effectiveness. The generated reflections produced by this system were considered improvements over the baseline model using standard metrics such as the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score and, in terms of absolute effectiveness, were on-par or above ground truth reference reflections. These results indicate that custom reflections from generative models may be effective for MI-based smoking cessation chatbots to increase users' confidence and motivation in quitting smoking.

To explore this possibility, Brown et al [4] have been iteratively developing MIBot, an MI-based smoking cessation chatbot that uses GPT-2 to generate custom reflections. They tested 3 versions of the chatbot—labeled v5.0, v5.1, and v5.2—on independent groups of recruited smokers to measure the effect of GPT-2–based generative reflections on moving smokers towards changing their smoking habits. They also used a version of the chatbot that did not generate reflections—v4.7—for comparison. MIBot v5.0, v5.1, and v5.2 asked 5 core questions, shown in Textbox 2 in sequence, expected a participant response after each question, and used a pretrained GPT-2 model to generate a custom reflection. MIBot v5.2 added extra secondary questions after questions 1 and 2 to allow participants to follow-up with their initial responses to a core question, and a specific version of question 4 if the answer to question 3 was to reduce smoking. MIBot v4.7 also asked these questions, but responded with "thank you" to each response rather than generating a reflection.

Textbox 2. The 5 motivational interviewing conversational questions in the MIBot v5.2 conversation used in this paper.

1. What is one thing you like about smoking?
2. What is one thing you dislike about smoking?
3. What is one thing about your smoking addiction that you would like to change?
4. What will your life look like once you make this change?
5. What is one step you need to make this change?

The effect of MIBot versions on readiness to quit was measured using a numerical scale called the Readiness Ruler [17]. Here, each participant was asked to rate their confidence, importance, and readiness to quit smoking from 0 to 10, with 10 indicating the highest value. Participants were asked to fill out the Readiness Ruler 3 times: just before, immediately after, and 1 week after the conversation with MIBot. Participants were also asked to score the perceived empathy of MIBot through the CARES (Consultation and Relational Empathy Survey) metric, a validated tool used to measure the perceived empathy of a health care interaction by asking a participant 10 statements that are each rated using a 6-point Likert scale [18].

Brown et al [4] found that there were statistically significant increases in participant confidence to quit smoking across all four chatbots 1 week after the conversation, with no statistically significant differences between them. This finding agreed with He et al's [15] results, and Brown et al [4] posited that asking questions may be enough to evoke an impact on confidence to quit. Version v5.2 did display statistically significant increases in importance and readiness to quit smoking when the other versions did not. In addition, v5.2 did exhibit a statistically significant increase in perceived empathy compared to v4.7 (P=.004) on the CARE scale. Both results were in contrast to He et al's [15] findings that there were no statistically significant differences between neutral and MI-style chatbot conversations, and Brown et al [4] postulated that this may be due to the effect of v5.2's LLM-based generative reflections.

MIBot v5.0, v5.1, and v5.2 generate GPT-2–based reflections that only use a participant's latest response. This precludes the generation of complex reflections that can refer to earlier responses in a conversation, which are the essential element of the BLCRs that are the focus of this paper. This work builds upon Brown et al's [4] work by creating and evaluating a method to generate BLCRs using GPT-4.

Methods

Overview

In this section, we describe the structure of the method used to generate BLCRs, the set of data we test our BLCR generation method on and how the resulting BLCRs are assessed.

Ethical Considerations

Ethical standards and approval directly follow those of Brown et al [4] as per the use of the data in the experiments described in that paper. The research used to acquire that data was approved by the University of Toronto Research Ethics Board under protocol number 35567, amended June 29, 2022, and all participants provided consent before participating in the Brown et al [4] study.

BLCR Generation Structure

In a chatbot conversation with a client, the client's latest and previous responses, along with the questions that were asked to evoke those responses, are packaged into a text called the *client message input*. A set of instructions, called the *BLCR prompt*, tells an LLM how to generate a BLCR from the client message input. These 2 texts are used together to generate a BLCR.

Client Message Input

The client message input (Textbox 3) consists of (1) conversation—the sequence of therapist questions and client responses up to the client's response right before the therapist's latest question—and (2) latest question-response—the therapist's latest question and the client's latest response.

Textbox 3. A sample client message input.

Conversation:

Therapist: What is one thing you like about smoking?

Client: It makes me to be more relaxed and releases my tension levels

Therapist: What is one thing you dislike about smoking?

Client: It would be the number of cigarettes I smoke a day plus the affordability of cigarettes theses days

Therapist: What is one thing about your smoking habit that you would like to change?

Client: The number or quantity I smoke a week

Therapist: What will your life look like when you make this change?

Client: If I can reduce by smoking 2 cigarettes a day and I would have some extra cash to do other things

Latest Question-Response:

Therapist: What are the steps you need to make this change?

Client: I need to probably set a smoking schedule that I need to stick too and also find a hobby to keep me distracted from my cravings

Backward-looking complex reflection:

The client message input is unique to each client response, and so changes on every client response. An LLM processes this input to generate a BLCR by first processing the instructions given in the BLCR prompt.

Prompt Design

The BLCR prompt, shown in Textbox 4, consists of (1) a request to generate a BLCR meeting the standards of MI, using terms presented in the client message input (see *Client Message Input* section); (2) a description of a complex reflection, taken from Miller and Rollnick [1]; (3) constraints and criteria to ensure the generated text meets the criteria of a complex reflection; (4) constraints and criteria to ensure the generated text meets the criteria of a BLCR; and (5) repetition of the request to generate a BLCR, given the above constraints and criteria.

The BLCR prompt is the same regardless of the client input message used. The BLCR prompt draws upon an LLM's implicit domain knowledge of MI [4,11], combined with a specific definition of a complex reflection, and constraints and criteria on what the output must follow to be an acceptable BLCR. For each client message input, an LLM can use the BLCR prompt's guidelines to generate a BLCR.

Textbox 4. The full backward-looking complex reflection prompt.

Generate a "backward-looking complex reflection" on the "Latest Question-Response" that meets the standards for Motivational Interviewing from the given "Conversation" about smoking cessation.

Refer to the following operational definition of a complex reflection in the context of Motivational Interviewing (MI):

Reflective listening statements are made by the clinician in response to client statements. A reflection may introduce new meaning or material, but it essentially captures and returns to clients something about what they have just said. Reflections are further categorized as simple or complex reflections.

Complex reflections typically add substantial meaning or emphasis to what the client has said. These reflections serve the purpose of conveying a deeper or more complex picture of what the client has said. Sometimes the clinician may choose to emphasize a particular part of what the client has said to make a point or take the conversation in a different direction. Clinicians may add subtle or very obvious content to the client's words, or they may combine statements from the client to form complex summaries.

A complex reflection has these hard constraints:

1. A complex reflection must be a statement and not a question.
2. A complex reflection must not give advice or information without permission, even if this advice is helpful.
3. A complex reflection must not direct the client by giving orders or commands.
4. A complex reflection must not disagree or challenge what the client has said.
5. A complex reflection must not incentivize people to smoke more, or discourage people from quitting smoking.
6. A complex reflection must not be factually wrong about smoking.
7. A complex reflection must be grammatically correct.

Here are some additional hard constraints for backward-looking complex reflections:

1. A backward-looking complex reflection must directly reference the Client statement and the Therapist question it is responding to in the Latest Question-Response.
2. A backward-looking complex reflection must include only one piece of extra information from earlier client statements in the Conversation.
3. A backward-looking complex reflection must not summarize the conversation.
4. A backward-looking complex reflection must use what the client has said in the last client statement, and the information from earlier client statements, and infer something about the client.

Given all the context above, generate a backward-looking complex reflection on the "Latest Question-Response" from the given "Conversation" that meets the Motivational Interviewing criteria of a complex reflection and satisfies all above hard constraints.

The BLCR prompt was created through an iterative process. Starting with an initial description was set of rules describing a BLCR and the requirements to generate a BLCR. This initial prompt was used to generate reflections on preexisting conversational data from prior conversations. These reflections were evaluated using the scale described in the *Evaluation of Quality of a BLCR* section. The prompt was subsequently revised to improve the responses, and the method attempted again on another set of independent conversational data. The revisions consisted of additional constraints and guidance, written in English, to address the shortcomings of the generated reflections. This iterative process continued until a prompt of sufficiently high evaluation score of the generated reflections was was achieved. The following sections describe both the data and the scale used.

Data

To test the BLCR prompt and client message inputs on real conversational data, 50 conversations were randomly selected from the MIBot version5.1 experiment data [4]. Each conversation consisted of the 5 MIBot core questions shown in (Textbox 2), along with their respective participant responses. As described in Brown et al [4], the participants were 50 anonymous volunteers from the Prolific platform who self-selected based on being current smokers. All 50 participants wrote their responses in text via the MIBot text-based chat interface. Multimedia Appendix 1 provides a sample conversation. Using the BLCR prompt and client message input, BLCRs would be generated for responses to Q3, Q4, and Q5 for each conversation, giving a total of 150 candidate BLCRs to assess.

Evaluation of Quality of a BLCR

A rating scale was developed to numerically evaluate the quality of a BLCR. This scale allows one to determine whether a BLCR is *acceptable*, that is, it meets the definition of a BLCR described in the *Prior Work* section.

The BLCR rating scale (Textbox 5) is an ordinal scale where higher number ratings successively include and build upon lower number ratings. If a BLCR achieves a rating of 3, this means it meets the criteria of 1 (referencing a client's latest response), 2 (referencing previous information in the conversation), and 3 (makes an inference about the client using present and past information). Satisfying these 3 requirements meets the definition of a BLCR as defined in the *Prior Work* section; therefore, we call any BLCRs rated 3 or greater acceptable BLCRs. A further rating of 4 is included to meet the preference for a “good” BLCR, which does not summarize the previous contents of the conversation, an optional condition that was deemed useful for indicating an unambiguous BLCR that exceeds the minimum acceptability requirements.

Textbox 5. The backward-looking complex reflection rating scale.

1: does the output reference the client's latest response somewhere?

- the output contains 1 or more references to the client's latest response

2: 1 + does the output reference some extra information from earlier in the conversation?

- the output contains 1 or more references to 1 or more previous client responses

3: 2 + does the output make an inference about the client using information in criteria 1 and 2?

- the output generates 1 or more novel assumptions about the client using information in 1 and 2

4: 3 + is the output not summarizing the sequence of the conversation word for word?

- the output does not repeat the information in each client response in sequence

Criteria to accept as a backward-looking complex reflection (score a 1 [True]): it is rated 3 or greater on the above rating scale.

A Python script was written to parse 50 conversations and build a formatted client message input for every Q3, Q4, and Q5 conversational sequence, creating 150 total inputs. These were fed to an LLM alongside the BLCR prompt, and the LLM generated 150 candidate BLCRs.

Three human raters were deployed to use the criteria of the BLCR Rating Scale to independently score all 150 generated BLCRs as acceptable or unacceptable. Using a binary score, an acceptable BLCR was scored 1 (true) if it received a rating of 3 or greater on the BLCR Rating Scale, while an unacceptable BLCR was scored 0 (false). The binary scoring was used to determine the *acceptability*: the percentage of accepted BLCRs among all generated BLCRs. The interrater reliability between the binary scores of the 3 raters was assessed using percent agreement and the calculation of Cohen κ. This metric was chosen specifically to measure interrater reliability with an ordinal scale, and was chosen instead of a similar metric such as Fleiss κ due to the latter's unsuitability in a case where all raters rate all items, which is the case for this BLCR assessment experiment [19].

Results

Overview

This section reports the fraction of the BLCRs generated using the evaluation method described in the *Methods* section that were deemed acceptable by each of the 3 human raters. The first section reports the percentage of accepted BLCRs between the 3 raters and between the 3 questions, along with a breakdown of the frequency of ranking scores per question and rater. The second section reports the interrater reliability between 3 pairs of the 3 raters (rater 1 and rater 2, rater 1 and rater 3, and rater 2 and rater 3) using percent agreement, with a brief discussion on the κ results.

BLCR Acceptability Statistics

Table 2 displays the percentage of BLCRs meeting the BLCR rating criteria as acceptable (BLCR rating of 3 or greater) broken down by the rater and the question. Table 3 displays the frequency of rating ranks broken down by question and by rater.

Table . Percentage of backward-looking complex reflections deemed acceptable by question and rater.

	Q3 (n=50)	Q4 (n=50)	Q5 (n=50)	Total (N=150)
Rater 1 (%)	92	90	96	93
Rater 2 (%)	73	90	88	84
Rater 3 (%)	90	88	86	88
Average acceptance (%)	85 (10)	89 (1)	90 (5)	88 (5)

Table . Frequency of rating by question and rater.

Question and rater	Rating, n				
	0	1	2	3	4
Q3					
Rater 1	4	0	0	3	44
Rater 2	1	0	13	2	35
Rater 3	2	0	3	6	40
Q4					
Rater 1	1	2	2	0	46
Rater 2	1	0	4	0	46
Rater 3	1	0	5	0	45
Q5					
Rater 1	1	0	1	0	49
Rater 2	1	0	5	2	43
Rater 3	3	2	2	0	43
Total (all questions combined)					
Rater 1	6	2	3	3	139
Rater 2	3	0	22	4	124
Rater 3	6	2	10	6	129

Table 2 breaks down the percent of acceptable BLCRs by rater and question, and the total column indicates the percent of BLCRs scored acceptable across all 150 responses by a single rater. The percentages in parentheses indicate the SD of the acceptability percentage.

The combination of high acceptability (Table 2) and high frequency of "4" ratings (Table 3) indicates that the majority of BLCRs generated by this method were considered "good" among all 3 raters. This is an indication that the LLM GPT-4 is highly capable of generating a BLCR. Multimedia Appendices 2 and 3 graph the frequencies of rating by question and rater, with both indicating a large skew toward "4" ratings.

Interrater Reliability

To assess the agreement of the results provided in Tables 1-3, Table 4 displays the percent agreement and Cohen κ for each rater pair. All 3 raters agreed on results at least 80% of the time.

Table . Percent agreement and Cohen κ for rater pairs.

	Rater 1, rater 2	Rater 1, rater 3	Rater 2, rater 3
Agreement (%)	84	88	80
Cohen κ	0.26	0.36	0.16

Discussion

Principal Findings

Altogether, the combination of high "4" frequency and a rating agreement of 80% and above indicates that this BLCR generation method can be expected to produce "good" BLCRs in the large majority of cases. In comparison, the κ values (Table 3) indicated weak to fair agreement between all 3 pairs of raters, based on standard interpretation criteria of κ. The discrepancy between high percentage agreement and weak to fair κ may be due to the majority of BLCRs being rated "4" by all 3 raters. The lack of contrastive negative examples (very few generated BLCRs that were rated 0, 1, or 2) skews the calculation of κ toward treating the labeling of widespread agreement as random chance. Therefore, percent agreement is thought to be a more realistic assessment of effectiveness in this context.

Multimedia Appendix 1 contains an example of a real conversation from Brown et al [4], with Brown et al's [4] reflections (labeled MIBot [data]) and BLCRs generated by this paper's method (labeled MIBot [BLCR]) below those reflections. Overall, the BLCRs generated successfully iterate on Brown et al's [4] provided reflections by better incorporating direct reflections on responses and linkages to previous responses to make inferences. A high-quality MI reflection would further infer about the emotional state of the client, and while the generated BLCRs are able to make rudimentary inferences about the mental state of the client ("it seems that…"), more work may be necessary to turn these inferences into those of emotional states. The high percentage of accepted BLCRs shows promise in prompt-based methods being an effective technique for MIBot to generate complex reflections that incorporate information from the past.

Limitations

The prompt-based BLCR generation method is restricted to MI conversations for smoking cessation and has only been tested in the context of 5-question MIBot conversations. Beyond this scope, this work may not generalize to other MI smoking cessation therapeutic contexts without changes to the prompt. However, the structure of the prompt itself is not specific to the data or the situation. The prompt can in theory be modified to remove references to smoking cessation and replace these with references to other domains, potentially offering a degree of domain generalizability across different subjects of MI therapy beyond smoking cessation. GPT-4 was the LLM model used in this work, but this method is applicable to any LLM model in theory. Newer LLM models, including future GPT models, may provide more robust results.

Conclusions

This paper presented a method to use an LLM-based prompt to generate BLCRs for a version of MIBot's MI smoking cessation conversation. It provided a definition of a BLCR, a prompt used to generate BLCRs, and a BLCR rating scale to assess whether a BLCR is acceptable. We found that 88% (n=150) of the generated BLCRs were deemed acceptable. This paper extends the work of Brown et al [4] by providing a method to generate complex reflections that incorporate information from earlier in the conversation, and uses GPT-4's strong text-generation capability rather than GPT-2.

Future work may build upon the definitions and methods introduced by this paper in three ways. First, the definition of a BLCR and the BLCR rating scale may be further refined to provide an accurate conceptual model of what the BLCR is trying to capture in a MI conversation. Second, the BLCR's prompt method can be adjusted to different MI therapeutic contexts beyond smoking cessation or refined to be more accurate for the smoking cessation context. Finally, the BLCR prompt method can be incorporated into MIBot, and its generated BLCRs can be assessed qualitatively and quantitatively in live experimental conversations.

Acknowledgments

This research was funded by a Natural Sciences and Engineering Research Council of Canada Discovery grant (RGPIN-2019-04395) and the Edward S Rogers Sr Department of Electrical and Computer Engineering at the University of Toronto.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Example of a MIBot v5.1 conversation. Generated reflections are italicized. Reflections marked "mibot (data)" are from the original dataset. Reflections marked "mibot (blcr)" are generated from this paper's prompting method. Both are provided for comparison.

[DOCX File, 6 KB - mental_v11i1e53778_app1.docx]

Multimedia Appendix 2

Frequency of backward-looking complex reflection rating score by question.

[PNG File, 20 KB - mental_v11i1e53778_app2.png]

Multimedia Appendix 3

Frequency of backward-looking complex reflection rating score by rater.

[PNG File, 19 KB - mental_v11i1e53778_app3.png]

References

1. Miller WR, Rollnick S. Motivational Interviewing: Helping People Change: Guilford Press; 2012.
2. Burke BL, Arkowitz H, Menchola M. The efficacy of motivational interviewing: a meta-analysis of controlled clinical trials. J Consult Clin Psychol 2003 Oct;71(5):843-861. [doi: 10.1037/0022-006X.71.5.843] [Medline: 14516234]
3. Kramer Schmidt L, Andersen K, Søgaard Nielsen A. Differences in the delivery of motivational interviewing across three countries. J Ethn Subst Abuse 2022 Jul;21(3):823-844. [doi: 10.1080/15332640.2020.1824838] [Medline: 33032500]
4. Brown A, Kumar AT, Melamed O, et al. A motivational-interviewing chatbot with generative reflections for increasing readiness to quit among smokers. JMIR Ment Health 2023 Oct 17;10:e49132. [doi: 10.2196/49132] [Medline: 37847539]
5. Jurafsky D, Martin J. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition: Prentice Hall; 2008.
6. Zhao WX, Zhou K, Li J, et al. A survey of large language models. arXiv. Preprint posted online on Nov 24, 2023. [doi: 10.48550/arXiv.2303.18223]
7. Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models. arXiv. Preprint posted online on Dec 29, 2023. [doi: 10.48550/arXiv.2307.03109]

8. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. Papers With Code. 2018. URL: https://paperswithcode.com/paper/improving-language-understanding-by [accessed 2024-06-03]
9. OpenAI, Achiam J, Adler S, et al. GPT-4 technical report. arXiv. Preprint posted online on Mar 4, 2024. [doi: 10.48550/arXiv.2303.08774]
10. Pollak KI, Alexander SC, Tulsky JA, et al. Physician empathy and listening: associations with patient satisfaction and autonomy. J Am Board Fam Med 2011;24(6):665-672. [doi: 10.3122/jabfm.2011.06.110025] [Medline: 22086809]
11. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. arXiv. Preprint posted online on Jul 28, 2021. [doi: 10.48550/arXiv.2107.13586]
12. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. arXiv. Preprint posted online on Jul 22, 2020. [doi: 10.48550/arXiv.2005.14165]
13. White J, Fu Q, Hays S, et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv. Preprint posted online on Feb 21, 2023. [doi: 10.48550/arXiv.2302.11382]
14. Almusharraf F, Rose J, Selby P. Engaging unmotivated smokers to move toward quitting: design of motivational interviewing-based chatbot through iterative interactions. J Med Internet Res 2020 Nov 3;22(11):e20251. [doi: 10.2196/20251] [Medline: 33141095]
15. He L, Basar E, Wiers RW, Antheunis ML, Krahmer E. Can chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance. BMC Public Health 2022 Apr 12;22(1):726. [doi: 10.1186/s12889-022-13115-x] [Medline: 35413887]
16. Shen S, Welch C, Mihalcea R, Pérez-Rosas V. Counseling-style reflection generation using generative pretrained transformers with augmented context. Presented at: 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue; Jul 1-3, 2020 URL: https://aclanthology.org/2020.sigdial-1 [accessed 2023-08-06] [doi: 10.18653/v1/2020.sigdial-1.2]
17. Readiness Ruler. Case Western Reserve University. 2010. URL: https://case.edu/socialwork/centerforebp/resources/readiness-ruler [accessed 2023-08-05]
18. Mercer SW, Maxwell M, Heaney D, Watt GC. The Consultation and Relational Empathy (CARE) measure: development and preliminary validation and reliability of an empathy-based consultation process measure. Fam Pract 2004 Dec;21(6):699-705. [doi: 10.1093/fampra/cmh621] [Medline: 15528286]
19. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282. [Medline: 23092060]

Abbreviations

BLCR: backward-looking complex reflection
CARES: Consultation and Relational Empathy Survey
GPT: Generative Pretrained Transformer
LLM: large language model
MI: motivational interviewing
ROUGE: Recall-Oriented Understudy for Gisting Evaluation

Edited by J Torous; submitted 18.10.23; peer-reviewed by J Andrade, J Heffner, S Siddals; revised version received 29.03.24; accepted 15.04.24; published 26.09.24.

Please cite as:
Kumar AT, Wang C, Dong A, Rose J
Generation of Backward-Looking Complex Reflections for a Motivational Interviewing–Based Smoking Cessation Chatbot Using GPT-4: Algorithm Development and Validation
JMIR Ment Health 2024;11:e53778
URL: https://mental.jmir.org/2024/1/e53778
doi:10.2196/53778

Original Paper

Assessing the Short-Term Efficacy of Digital Cognitive Behavioral Therapy for Insomnia With Different Types of Coaching: Randomized Controlled Comparative Trial

Wai Sze Chan[1], PhD; Wing Yee Cheng[1], BSocSc; Samson Hoi Chun Lok[1], MSocSc; Amanda Kah Mun Cheah[1], MSocSc; Anna Kai Win Lee[1], MSc; Albe Sin Ying Ng[1], PhD; Tobias Kowatsch[2,3,4], PhD

[1]Department of Psychology, The University of Hong Kong, Hong Kong, China (Hong Kong)

[2]Institute for Implementation Science in Health Care, University of Zurich, Zurich, Switzerland

[3]School of Medicine, University of St.Gallen, St. Gallen, Switzerland

[4]Centre for Digital Health Interventions, Department of Management, Technology, and Economics, Eidgenössische Technische Hochschule Zürich, Zurich, Switzerland

Corresponding Author:
Wai Sze Chan, PhD
Department of Psychology
The University of Hong Kong
Room 627, the Jockey Club Tower
Pokfulam, Hong Kong
Hong Kong
China (Hong Kong)
Phone: 852 39172295
Email: chanwais@hku.hk

Abstract

Background: Digital cognitive behavioral therapy for insomnia (dCBTi) is an effective intervention for treating insomnia. The findings regarding its efficacy compared to face-to-face cognitive behavioral therapy for insomnia are inconclusive but suggest that dCBTi might be inferior. The lack of human support and low treatment adherence are believed to be barriers to dCBTi achieving its optimal efficacy. However, there has yet to be a direct comparative trial of dCBTi with different types of coaching support.

Objective: This study examines whether adding chatbot-based and human coaching would improve the treatment efficacy of, and adherence to, dCBTi.

Methods: Overall, 129 participants (n=98, 76% women; age: mean 34.09, SD 12.05 y) whose scores on the Insomnia Severity Index [ISI] were greater than 9 were recruited. A randomized controlled comparative trial with 5 arms was conducted: dCBTi with chatbot-based coaching and therapist support (dCBTi-therapist), dCBTi with chatbot-based coaching and research assistant support, dCBTi with chatbot-based coaching only, dCBTi without any coaching, and digital sleep hygiene and self-monitoring control. Participants were blinded to the condition assignment and study hypotheses, and the outcomes were self-assessed using questionnaires administered on the web. The outcomes included measures of insomnia (the ISI and the Sleep Condition Indicator), mood disturbances, fatigue, daytime sleepiness, quality of life, dysfunctional beliefs about sleep, and sleep-related safety behaviors administered at baseline, after treatment, and at 4-week follow-up. Treatment adherence was measured by the completion of video sessions and sleep diaries. An intention-to-treat analysis was conducted.

Results: Significant condition-by-time interaction effects showed that dCBTi recipients, regardless of having any coaching, had greater improvements in insomnia measured by the Sleep Condition Indicator (P=.003; d=0.45) but not the ISI (P=.86; d=–0.28), depressive symptoms (P<.001; d=–0.62), anxiety (P=.01; d=–0.40), fatigue (P=.02; d=–0.35), dysfunctional beliefs about sleep (P<.001; d=–0.53), and safety behaviors related to sleep (P=.001; d=–0.50) than those who received digital sleep hygiene and self-monitoring control. The addition of chatbot-based coaching and human support did not improve treatment efficacy. However, adding human support promoted greater reductions in fatigue (P=.03; d=–0.33) and sleep-related safety behaviors (P=.05; d=–0.30) than dCBTi with chatbot-based coaching only at 4-week follow-up. dCBTi-therapist had the highest video and diary completion rates compared to other conditions (video: 16/25, 60% in dCBTi-therapist vs <3/21, <25% in dCBTi without any coaching), indicating greater treatment adherence.

Conclusions: Our findings support the efficacy of dCBTi in treating insomnia, reducing thoughts and behaviors that perpetuate insomnia, reducing mood disturbances and fatigue, and improving quality of life. Adding chatbot-based coaching and human support did not significantly improve the efficacy of dCBTi after treatment. However, adding human support had incremental benefits on reducing fatigue and behaviors that could perpetuate insomnia, and hence may improve long-term efficacy.

Trial Registration: ClinicalTrials.gov NCT05136638; https://www.clinicaltrials.gov/study/NCT05136638

(JMIR Ment Health 2024;11:e51716) doi:10.2196/51716

KEYWORDS

insomnia; cognitive behavioral therapy; digital intervention; mobile health; mHealth; chatbot-based coaching; human support; mobile phone

Introduction

Background

Insomnia disorder is the most common sleep-wake disorder. It is characterized by difficulty initiating or maintaining sleep despite having adequate opportunities to sleep or having nonrestorative sleep, not explained by other sleep disorders [1,2]. The prevalence of insomnia disorder ranges from 2.3% to 25.5% globally [3] and was heightened during the COVID-19 pandemic, affecting up to one-third of the general population [4-6]. Insomnia is associated with high societal and economic costs resulting from health care utilization, work absenteeism, and lost productivity [7,8]. In the United States, the estimated annual cost of insomnia reaches up to US $100 billion [9]. In addition, insomnia contributes to substantial losses in annual gross domestic product, amounting to US $19.6 billion in Canada, US $41.4 billion in the United Kingdom, and US $19.2 billion in Australia [10]. The cost of untreated insomnia outweighs the cost of treating insomnia, with cognitive behavioral therapy for insomnia (CBTi) achieving greater cost-effectiveness than pharmacological treatments [11]. Technology-aided delivery of CBTi may have even greater cost-effectiveness, given its high scalability and reduced demands for human resources compared to face-to-face CBTi [12].

CBTi is an evidence-based, first-line treatment for insomnia disorder recommended by health organizations around the world [13,14]. It targets the cognitive and behavioral mechanisms perpetuating sleep difficulties, namely compromised sleep drive, disturbed circadian rhythm, and hyperarousal, especially hyperarousal associated with the sleeping environment or sleep per se [15-17]. Integrating multiple treatment techniques, CBTi aims to preserve sleep drive, stabilize circadian rhythm, and reduce hyperarousal; it is typically delivered in 6 to 8 hourly sessions by a trained mental health professional [18]. CBTi effectively improves sleep quality and reduces insomnia symptoms across populations, including populations with medical and psychiatric comorbidities. On average, CBTi leads to 20- to 30-minute reductions in both sleep onset latency and wake after sleep onset and approximately 10% increases in sleep efficiency at the end of treatment [19] with less consistent effects on total sleep time (TST) [20]. CBTi also improves mood, fatigue, and quality of life—indirectly through sleep improvements or directly through changes in behavior and cognition [21,22]. Despite its strong evidence base, CBTi is not consistently delivered to most people due to various implementation obstacles, notably the lack of mental health workers trained in delivering CBTi and limitations in time and resources required for delivering and receiving CBTi [23].

Digital CBTi (dCBTi) is a promising alternative mode of delivery of CBTi, given its high scalability and low demands for human resources. Meta-analyses have found comparable efficacy estimates of dCBTi compared to face-to-face CBTi [24,25]. However, a direct comparative trial found dCBTi less efficacious than face-to-face CBTi [26]. Notably, fully automated dCBTi is the most cost-effective treatment of insomnia, followed by group CBTi and individual CBTi [27]. Of the different types of dCBTi, dCBTi with therapist support is found to have the highest efficacy compared to other types of dCBTi without therapist support [28]. Indeed, CBTi recipients attribute treatment success to the working alliance with the therapist, and they perceive therapist-assisted support, such as the provision of personalized feedback, motivational enhancement, and accountability, to be critical for increasing engagement with dCBTi [23]. Nonetheless, the need for therapist support hinders the scalability and accessibility of dCBTi, which are the core benefits of dCBTi over face-to-face CBTi.

dCBTi with nontherapist guidance is a lower-cost alternative. Although it may be limited in delivering expert advice and addressing challenging motivational or emotional barriers, nontherapist support can provide personalized feedback, motivational enhancement, and accountability. Nontherapist support has been found to improve treatment outcomes of self-help CBTi [29]. Technological advancements have enabled the development of virtual conversational agents designed to mimic patient-therapist interactions and provide personalized content and feedback, also known as chatbot-based coaching or e-coaching [30]. Nontherapist and chatbot-based support require no therapist and preserve the core benefit of dCBTi over face-to-face CBTi. Meta-analyses have found mixed results regarding the effect sizes of dCBTi with different types of coaching support [24,25], but a direct comparative trial has not yet been conducted.

Low treatment adherence and high attrition rates are the major challenges in implementing dCBTi and even face-to-face CBTi [31,32]. On average, half of the dCBTi recipients do not adhere to treatment [33] compared to 14% to 44% in face-to-face CBTi [34]. While dCBTi offers greater potential for scalability, if engagement and treatment adherence are low, its impact on population health will remain minimal. Adding human or chatbot-based guidance to dCBTi is one of the most discussed solutions to improving engagement [31,33]; however, empirical support for its effects on treatment adherence is still being determined.

In sum, dCBTi is undoubtedly a promising intervention for insomnia that could have a major impact at the population level. Nevertheless, more research on how to optimize its benefits is needed. In particular, adding human or chatbot-based coaching has been frequently regarded as a useful strategy to improve treatment efficacy. However, a direct comparative trial of dCBTi with different types of coaching compared to dCBTi without any coaching (dCBTi-unguided) has yet to be conducted. Furthermore, the effects of different types of coaching on treatment adherence have not been evaluated.

Objectives

This study is the first empirical evaluation of the efficacy of dCBTi with different types of coaching. Moreover, we aim to evaluate whether different coaching types improve treatment adherence along with treatment outcomes. Specifically, in a 5-arm randomized controlled comparative trial, we aim to evaluate whether (1) a fully automated mobile phone–based dCBTi has superior efficacy to an active digital sleep hygiene and self-monitoring control (dSH); (2) adding coaching, regardless of type, would enhance treatment adherence and efficacy compared to dCBTi-unguided; (3) dCBTi with human support would enhance treatment adherence and efficacy compared to dCBTi with chatbot-based coaching only (dCBTi-chatbot); and (4) support from a therapist is superior to support from a nontherapist. We hypothesized that (1) dCBTi, regardless of the presence of coaching, would be efficacious for improving insomnia symptoms compared to dSH; (2) dCBTi with coaching would promote greater improvements in insomnia symptoms and greater treatment adherence than dCBTi-unguided; (3) dCBTi with human coaching would promote greater improvements in insomnia symptoms and treatment adherence than dCBTi-chatbot; and (4) dCBTi with chatbot-based coaching and therapist support (dCBTi-therapist) would promote greater improvements in insomnia symptoms and greater treatment adherence than dCBTi with chatbot-based coaching and research assistant support (dCBTi-assistant).

Methods

Study Design

The study was a 5-arm, parallel, participant-blinded, randomized controlled comparative trial. The 5 conditions included dCBTi-therapist, dCBTi-assistant, dCBTi-chatbot, dCBTi-unguided, and dSH. The study protocol was preregistered on ClinicalTrials.gov (NCT05136638).

Ethical Considerations

This study was approved by the University of Hong Kong Human Research Ethics Committee before data collection (EA210458). Electronic informed consent was obtained from each participant before study participation. Each participant was informed that participation was entirely voluntary, and they could withdraw from the study at any point without negative consequences. All data were kept confidential in a password-protected drive. Only the research team had access to the data. All personal identifying information was removed from the research data and will be kept separately from the research data for 3 years after the publication of the main study findings to ensure that there are no problems regarding consent, fabrication, or falsification. Anonymous data will be kept indefinitely. Each participant was compensated HK $60 (approximately US $8) for completing research measures at each time point.

Randomization and Blinding

Simple randomization with equal chances of assigning a participant to 1 of the 5 conditions was conducted using the randomization function implemented in Sleep Sensei, a mobile app developed with the MobileCoach platform [35,36] specifically for this study. Participants were informed that they were assigned to one of the intervention conditions. However, they were not informed about the conditions or the condition to which they were assigned. They were not informed about the study hypotheses either. As all participants were given access to Sleep Sensei, we considered them blinded to the condition assignment and study hypotheses. The therapists and assistants who provided coaching support were not blinded to the assignment. The assessments of treatment outcomes were all self-administered using Qualtrics (Qualtrics International Inc).

Participants

The inclusion criteria were as follows: participants who (1) have an Insomnia Severity Index (ISI) score of ≥10, indicating clinically significant insomnia [37]; (2) are aged 18 to 65 years; (3) have access to a smartphone and a local telephone number; and (4) are able to understand written Chinese and spoken Cantonese, the languages used in Sleep Sensei. The exclusion criteria were as follows: (1) self-reported sleep apnea or high risk of sleep apnea identified using the Berlin Questionnaire [38]; (2) self-reported acute, untreated mental or medical illnesses that would interfere with participation; (3) suicidal ideation suggested by a score of ≥1 on the Patient Health Questionnaire-9 (PHQ-9) and confirmed by a follow-up interview by a clinical psychology trainee; (4) unstable medication use that can affect sleep; (5) currently receiving psychotherapy for insomnia; and (6) other conditions that prevent adherence to CBTi recommendations, such as shift work. Participants who did not complete baseline research assessments were also excluded. Eligible participants showed sufficient digital literacy to be able to complete the web-based screening survey and use the mobile app.

An a priori power analysis was conducted to determine the sample size required for detecting significant group-by-time interaction effects if there were clinically meaningful differences in the primary outcome (a 4-point difference in the ISI total score [39]) in the patterns specified in our 4 hypotheses. A simulation-based power analysis was performed using the R package *mixedpower* [40]. We simulated a database using the means and SDs of ISI scores from a local sample for another insomnia trial as the baseline values [41] because these data were most likely the closest estimates of baseline ISI values in our study sample. For hypothesis 1, we simulated a database with after-treatment ISI values to be 4 points lower than the baseline ISI values in the experimental condition. For hypotheses

2 to 4, we simulated a database with after-treatment ISI values to be 4 points lower than the baseline's and 4 points lower than the comparison group's. On the basis of these simulations, a sample size of 120 would be required for detecting significant results with statistical power >95%, >80%, >90%, and >80% for hypotheses 1, 2, 3, and 4, respectively.

Procedures

Participants were recruited using mass emails sent to students, staff, and affiliates of the university, as well as social media advertisements with the institutional affiliation displayed. Inclusion and exclusion criteria were evaluated based on potential participants' responses to the screening survey, followed by telephone interviews as needed by authors SHCL and AKMC, both of whom were clinical psychology trainees under the supervision of lead author WSC, a licensed clinical psychologist. Eligible participants then downloaded Sleep Sensei for free. All conditions were delivered via Sleep Sensei. They were given access to modules and functions according to the condition to which they were assigned (Table 1). They had access to Sleep Sensei from baseline to follow-up. Assessments of outcomes were administered at baseline, after treatment, and at 4-week after-treatment (follow-up) using Qualtrics. Figure 1 presents the CONSORT (Consolidated Standards of Reporting Trials) flow diagram.

Table 1. Intervention components in each treatment condition.

Treatment components	dCBTi-therapist[a]	dCBTi-assistant[b]	dCBTi-chatbot[c]	dCBTi-unguided[d]	dSH[e]
Video lessons	✓	✓	✓	✓	✓[f]
Resource library	✓	✓	✓	✓	✓[f]
Daily diary entry and visualization	✓	✓	✓	✓	✓
Automatic customized sleep schedule suggestions	✓	✓	✓	✓	
Weekly goal-setting and action-planning entries	✓	✓	✓	✓	
Chatbot-based coaching	✓	✓	✓		
Assistant support		✓			
Therapist support	✓				

[a]dCBTi-therapist: digital cognitive behavioral therapy for insomnia with chatbot-based coaching and therapist support.

[b]dCBTi-assistant: digital cognitive behavioral therapy for insomnia with chatbot-based coaching and research assistant support.

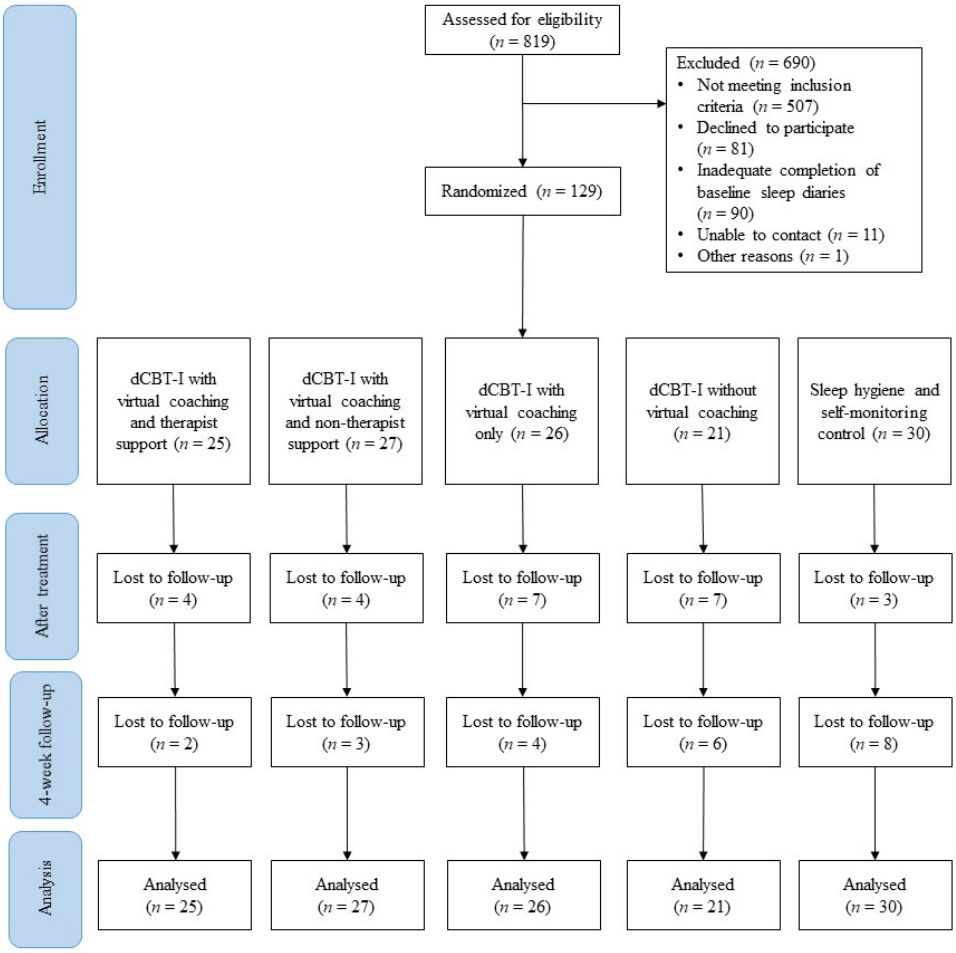
[c]dCBTi-chatbot: digital cognitive behavioral therapy for insomnia with chatbot-based coaching only.

[d]dCBTi-unguided: digital cognitive behavioral therapy for insomnia without any coaching.

[e]dSH: digital sleep hygiene and self-monitoring control.

[f]Sleep hygiene only.

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flowchart of participants. dCBTi: digital cognitive behavioral therapy for insomnia.



Interventions

Sleep Sensei

Sleep Sensei was developed specifically for this study and hence was customized to deliver the 5 conditions with differing combinations of content and functionalities. Sleep Sensei used the talk-and-tools user-interface paradigm, which comprised a *talk* system enabling text-based chat interactions between the user and a conversational agent (ie, a chatbot-based coach) and a *tools* system allowing the user to observe and manipulate the objects in the interface [30]. Before launching the app, Sleep Sensei underwent 3 rounds of usability testing with 6 volunteers.

In Sleep Sensei, the core CBTi treatment components were implemented using the tools system, which consisted of (1) 6 video lessons providing psychoeducation and the rationale for each treatment recommendation, delivered chronologically in the order described in the next subsection; (2) a resource library storing and presenting video lesson content in written format as well as additional resources such as relaxation recordings; (3) daily diary entries and visualization of diary data; (4) automatic, individually tailored weekly sleep schedule suggestions; and (5) weekly goal-setting and action-planning entries. Multimedia Appendix 1 presents screenshots of the Sleep Sensei interface.

CBTi Core Treatment Components

CBTi is a multicomponent intervention that combines behavioral techniques and cognitive therapy [14]. The core treatment components of CBTi include sleep restriction, stimulus control, sleep hygiene, psychoeducation about sleep, relaxation, and cognitive therapy. Sleep hygiene is included in CBTi but is not considered CBTi on its own. It is often used as an active control condition in clinical trials of CBTi [42].

Sleep restriction is a technique to improve one's sleep efficiency by restoring a high sleep drive via limiting time in bed (TIB), specifically to match one's current sleep needs. It is achieved by tailoring TIB to match one's average TST. Once adequate sleep efficiency is achieved, TIB can gradually extend until optimal sleep duration is reached. In this study, the rationale and procedures for sleep restriction are presented in video lesson 1. If the participant's previous-week sleep diary data show that their sleep efficiency is ≥85%, and they indicated that they were not sleep deprived by answering a *yes* or *no* question, they would be asked to maintain their TIB with consistent wake time and bedtime. If their sleep efficiency is ≥85%, and they indicated that they were sleep deprived, the app would recommend TIB with a 20-minute increase. If their sleep efficiency is <85%, the app would recommend a range of TIBs, from the minimum TST (equivalent to the previous week's average TST based on diary data but at least 5 hours) to 20 minutes less than the previous week's TIB. The participant will be asked to specify a consistent wake time suitable for them and choose a bedtime that would result in a TIB within the suggested range. This implementation of the sleep restriction procedures enables the participant to choose to restrict TIB more aggressively or gradually by 20 minutes each week, allowing for greater flexibility and potentially greater adherence.

Sleep hygiene refers to daily habits that influence sleep drive, circadian rhythm, or arousal associated with the sleeping environment. The daily habits include keeping the sleeping environment dark, quiet, and cool; having a consistent sleep schedule, especially a consistent wake time; maintaining adequate light exposure and activity levels; reducing stimulating activities before bedtime such as eating too much, intense exercise, alcohol or nicotine consumption, and screen time; and setting up a relaxation routine before bedtime. The rationale and procedures for sleep hygiene are presented in video lesson 2.

Stimulus control refers to reconditioning the sleeping environment to be a place only for sleeping and to reduce conditioned arousal. The procedures instruct the participant to enter the sleeping environment only when they feel ready to fall asleep, leave the sleeping environment if they cannot fall asleep or stay asleep for approximately >20 minutes by estimation, and go back to bed only when one feels sleepy. The distinction between feeling fatigued and sleepy is also discussed. The rationale and procedures for stimulus control are presented in video lesson 3.

Relaxation includes practices of progressive muscle relaxation and visual imagery exercises that can be used to reduce hyperarousal. The rationale and procedures for relaxation are presented in video lesson 4. Participants are guided to practice relaxation during the video lesson and told that recordings of relaxation guidance could be accessed in the resource library. In addition, another clinical technique targeting excessive worry is also introduced in this lesson—setting up worry time (referring to setting aside a 30-minute period each day [not close to bedtime] for worrying) and limiting worrying only to this time to restrict the impact of worrying on mental health and sleep.

Cognitive therapy refers to a set of techniques to identify and reframe thoughts and beliefs that may promote and sustain sleep difficulties, such as thoughts that lead to heightened worries and frustration about the consequences of not having good sleep or thoughts that reduce motivation and treatment adherence. The rationale and procedures for cognitive therapy are presented in video lessons 5 and 6.

Chatbot-Based Coaching

The chatbot-based coaching element was delivered using the *talk* system via a series of logic-based preprogrammed conversational turns created by WSC, SHCL, and AKMC with the goal to simulate therapeutic interactions. These text-based conversations covered the following areas: (1) after-video summary and quiz (the chatbot-based coach guided the participants to reflect on what they learned from the video lessons and facilitated them to apply the learned strategies to their own situations), (2) weekly goal setting and action planning (the participants were guided to set up goals and action plans to implement the treatment recommendations during the week), (3) positive feedback and reflection (encouraging messages were sent to participants when they completed daily sleep diaries and achieved their weekly goals), and (4) problem-solving (when the participants did not complete daily diaries or achieve weekly goals, the chatbot-based coach guided them to think about different solutions to remove the barriers to implement the treatment recommendations).

Research Assistant Support

Research assistant support was provided to participants in dCBTi-assistant at the end of sessions 1, 3, and 6 by 2 undergraduate research interns who had no prior experience in counseling or CBTi. They were trained by WSC to provide supportive contact, including expressing appreciation for the participants' time and effort in using the app, encouraging them to continue using the app and complete daily diaries, and asking whether they had encountered any technical issues that needed support. If the participant asked sleep- or CBTi-related questions, the assistants would direct the participant to review the CBTi materials on the app. The contact time was restricted to between 20 and 30 minutes. The percentages of telephone calls completed were 74% (20/27), 33% (9/27), and 30% (8/27) for sessions 1, 3, and 6, respectively.

Therapist Support

Therapist support was provided to participants in dCBTi-therapist at the end of sessions 1, 3, and 6 by 2 postgraduate clinical psychology trainees (SHCL and AKMC) who had received at least 1 year of clinical psychology training and training in a CBTi protocol. The support included reinforcing the understanding of the intervention materials, providing sleep- and CBTi-related expert advice, reviewing treatment progress, identifying and resolving barriers to implementing CBTi treatment strategies, and addressing any motivational issues. Similar to the dCBTi-assistant condition, the contact time was restricted to between 20 and 30 minutes. The percentages of telephone calls completed were 84% (21/25), 76% (19/25), and 68% (17/25) for sessions 1, 3, and 6, respectively.

Measures

Overview

The assessments were all electronic questionnaires administered using Qualtrics. These questionnaires had been tested by research assistants before being distributed to the participants. The questionnaires were distributed through email, and participants were instructed to complete them within 1 week. The order of the questions was the same for each participant. All questions were mandatory.

Primary Outcome (ISI)

The ISI is a widely used self-report questionnaire assessing insomnia symptoms in the previous 2 weeks [43]. It consists of 7 items asking the participants to rate the severity of their insomnia symptoms on a 5-point Likert scale ranging from 0 to 4, with higher scores indicating greater insomnia severity. The composite score ranges from 0 to 28, with a score of ≥10 indicating clinical insomnia. The validated Chinese version was used in this study [44]. The ISI showed acceptable to excellent internal consistencies across the 3 time points, with Cronbach α values in the range of 0.77 to 0.90.

Secondary Outcomes

Sleep Condition Indicator

The Sleep Condition Indicator (SCI) is a newer measure of insomnia symptoms developed based on the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (*DSM-5*), diagnostic criteria [1]; research diagnostic criteria [45]; and recommended quantitative parameters [46]. It consists of 8 items assessing the severity of sleep difficulties and daytime impairment during the past month. Four items assess insomnia symptoms on 5-point scales with quantitative anchors (ie, frequency or duration); for instance, item 1 asks, “How long does it take for you to fall asleep?” and the participants respond on a scale ranging from 0 to 4 (0=*0-15 min*, 1=*16-30 min*, 2=*31-45 min*, 3=*46-60 min*, and 4=*>61 min*). The other 4 items assess insomnia symptoms on 5-point scales with qualitative anchors; for instance, item 4 asks, “How do you rate your sleep quality?” and the participants respond on a scale ranging from 0 (*very good*) to 4 (*very bad*). The SCI has been validated against the ISI, and it has shown good reliability and convergent validity. The validated Chinese version of the SCI was used in this study [47]. The total score ranges from 0 to 32; higher scores indicate lower levels of insomnia, and a score of ≤16 indicates insomnia disorder. In this study, the SCI had acceptable to good internal consistency, with Cronbach α values in the range of 0.70 to 0.88.

PHQ-9 for Depressive Symptoms

The PHQ-9 was used to assess depressive symptoms [48]. The PHQ-9 consists of 9 items asking participants to rate the frequency of their depressive symptoms in the past 2 weeks on a 4-point Likert scale ranging from 0 (*not at all*) to 3 (*nearly every day*). The total score ranges from 0 to 27; higher scores indicate higher levels of depressive symptoms, with a PHQ-9 cutoff score of ≥10 indicating clinically significant depressive symptoms. The validated Chinese version was used in this study [49]. The PHQ-9 showed good internal consistencies across the 3 time points, with Cronbach α values in the range of 0.81 to 0.88.

Generalized Anxiety Disorder-7

The Generalized Anxiety Disorder-7 (GAD-7) was used to measure anxiety symptoms [50]. The GAD-7 consists of 7 items asking participants to rate the frequency of their anxiety symptoms in the past 2 weeks on a 4-point Likert scale ranging from 0 (*not at all*) to 3 (*nearly every day*). The composite score ranges from 0 to 21; higher scores indicate higher levels of anxiety symptoms, with a GAD-7 cutoff score of ≥8 indicating clinically significant anxiety symptoms. The validated Chinese version was used in this study [51]. The GAD-7 showed excellent internal consistencies across the 3 time points, with Cronbach α values in the range of 0.92 to 0.95.

Fatigue Assessment Scale

The Fatigue Assessment Scale (FAS) was used to assess fatigue [52]. The FAS consists of 10 items asking participants to rate the frequency of their fatigue symptoms on a 5-point Likert scale ranging from 1 (*never*) to 5 (*always*). The total score ranges from 10 to 50, with higher scores indicating higher levels of fatigue. The validated Chinese version was used in this study [53]. In this study, the FAS had good internal consistencies across the 3 time points, with Cronbach α values in the range of 0.88 to 0.90.

Epworth Sleepiness Scale

The validated Chinese version of the Epworth Sleepiness Scale (ESS) was used to assess daytime sleepiness [54]. The ESS consists of 8 items asking participants to rate their chances of dozing in different situations on a 4-point Likert scale, ranging from 0 (*never*) to 3 (*high chance*). The composite score ranges from 0 to 24, and higher scores indicate greater daytime sleepiness, with an ESS score of ≥11 indicating excessive or clinically significant daytime sleepiness. The ESS showed good internal consistencies across the 3 time points, with Cronbach α values in the range of 0.82 to 0.83.

Satisfaction With Life Scale

The Satisfaction With Life Scale (SWLS) was used to measure general psychological well-being [55]. The SWLS asks participants to rate their agreement with 5 statements of life satisfaction on a 7-point Likert scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). The total score ranges from 5 to 35; higher scores indicate greater psychological well-being. The validated Chinese version was used in this study [56]. The SWLS showed good internal consistencies across the 3 time points, with Cronbach α values in the range of 0.90 to 0.92.

Mechanistic Outcomes

Dysfunctional Beliefs and Attitudes About Sleep-16

The Dysfunctional Beliefs and Attitudes About Sleep-16 (DBAS-16) measures one’s endorsement of dysfunctional thoughts and beliefs that could elevate anxiety and frustration about sleep difficulties, leading to the perpetuation of insomnia [57]; for example, 1 of the items is “I need 8 hours of sleep to feel refreshed and function well during the day.” The DBAS-16 asks participants to rate how much they believe the 16 statements about sleep on an 11-point Likert scale ranging from 0 (*strongly disagree*) to 10 (*strongly agree*). The total score ranges from 0 to 160; higher scores indicate stronger dysfunctional beliefs about sleep. The validated Chinese version was used in this study [58]. The DBAS-16 showed good to excellent internal consistencies across the 3 time points, with Cronbach α values in the range of 0.85 to 0.95.

Sleep-Related Behaviors Questionnaire

The Sleep-Related Behaviors Questionnaire (SRBQ) measures one’s engagement in sleep-related safety behaviors, which are behaviors that aim to alleviate the distress and consequences of insomnia but inadvertently perpetuate insomnia [59] (eg, preoccupation with sleep, such as “I spend time considering ways to improve sleep”; and reduced daytime engagement to preserve energy, such as “I take on fewer social commitments”). The validated Chinese version of the SRBQ, which consists of 20 statements, was used [41]. Participants rated these statements on a 5-point Likert scale ranging from 0 (*almost never*) to 4 (*almost always*). The composite score ranges from 0 to 80; higher scores indicate greater engagement in sleep-related safety behaviors. The SRBQ had good to excellent internal consistencies across the 3 time points, with Cronbach α values in the range of 0.85 to 0.93.

Treatment Adherence

In this study, video completion and sleep diary completion for each dCBTi session were used as the indicators of treatment adherence, which are common global indicators of treatment adherence that have been used in other CBTi and dCBTi trials [60].

Statistical Analysis

Analyses were performed with R (version 4.0.2; R Foundation for Statistical Computing). All tests for significance were 2-tailed, and $P<.05$ was considered statistically significant. Intention-to-treat analyses were conducted using linear mixed models with the restricted maximum likelihood method for handling missing data. The restricted maximum likelihood method incorporates the observed data and model covariance structure to estimate the variance parameters in the model with missing data [61]. The models included the treatment groups (dCBTi-therapist, dCBTi-assistant, dCBTi-chatbot, dCBTi-unguided, and dSH), time (baseline, after treatment, and follow-up), and the group-by-time interaction as fixed effects. Participant IDs were included as the random effect in the model. Planned contrasts were specified in the models to test our 4 hypotheses: hypothesis 1, all dCBTi conditions compared to dSH; hypothesis 2, all guided dCBTi conditions compared to unguided dCBTi; hypothesis 3, dCBTi with human support compared to dCBTi-chatbot; and hypothesis 4, dCBTi-therapist compared to dCBTi-assistant.

Cohen *d* was calculated from the mean differences between the conditions after treatment to indicate the effect size of each significant effect. Fisher exact tests were conducted to evaluate whether the percentages of participants who completed the video and sleep diary for each session were different across the conditions. In addition, we conducted chi-square tests on remission rates to evaluate the differences in remission rates across the conditions after treatment. Remission was defined as having an ISI score of <10 or an SCI score of >21. We also conducted an analysis on the percentages of participants achieving a clinically meaningful reduction in depressive symptoms (a 5-point reduction in the PHQ-9 score), anxiety symptoms (a 4-point reduction in the GAD-7 score), fatigue (a 4-point reduction in the FAS score), and daytime sleepiness (a 2-point reduction in the ESS score).

Results

Descriptives

Of the 819 individuals who completed the screening survey, 690 (84.2%) were not eligible, declined to participate, or did not complete the baseline measures; thus the final sample consisted of 129 (15.8%) participants (age: mean 34.09, SD 12.05 y). Most of the participants were female (98/129, 76%), had never married (87/129, 67.4%), had completed tertiary education (104/129, 80.6%), and were employed full time (77/129, 59.7%). There were no significant differences in age, marital status, education level, employment status, or monthly household income across treatment conditions (Table 2). Table 3 presents the mean values and SDs of each outcome at 3 time points. No significant differences were observed in all outcomes at baseline. Significant differences in the ISI, SCI, PHQ-9, GAD-7, DBAS-16, and SRBQ scores across the conditions were observed after treatment, favoring the treatment conditions over the control.

Table 2. Demographic characteristics of participants at baseline.

Characteristics	dCBTi-therapist[a] (n=25)	dCBTi-assistant[b] (n=27)	dCBTi-chatbot[c] (n=26)	dCBTi-unguided[d] (n=21)	dSH[e] (n=30)	Full sample (n=129)	*F* test (*df*)	Chi-square (*df*)	*P* value
Age (y), mean (SD)	34.28 (12.18)	34.22 (13.62)	35.58 (12.23)	30.76 (10.45)	34.83 (11.68)	34.09 (12.05)	0.52 (4,124)	—[f]	.72
Sex, n (%)							—	10.2 (4)	.04
Female	18 (72)	21 (77.8)	21 (80.8)	11 (52.4)	27 (90)	98 (76)			
Male	7 (28)	6 (22.2)	5 (19.2)	10 (47.6)	3 (10)	31 (24)			
Marital status, n (%)							—	17.1 (12)	.15
Never married	16 (64)	20 (74.1)	14 (53.8)	18 (85.7)	19 (63.3)	87 (67.4)			
Cohabiting	2 (8)	0 (0)	3 (11.5)	0 (0)	0 (0)	5 (3.9)			
Married	5 (20)	5 (18.5)	9 (34.6)	2 (9.5)	10 (33.3)	31 (24)			
Divorced or separated	2 (8)	1 (7.4)	0 (0)	1 (4.8)	1 (3.3)	6 (4.7)			
Highest educational level, n (%)							—	12.1 (8)	.15
Secondary	4 (16)	2 (7.4)	4 (15.4)	0 (0)	2 (6.7)	12 (9.3)			
Tertiary (nondegree)	3 (12)	3 (11.1)	1 (3.8)	0 (0)	6 (20)	13 (10.1)			
Tertiary (degree)	18 (72)	22 (81.5)	21 (80.8)	21 (100)	22 (73.3)	104 (80.6)			
Employment status, n (%)							—	13.9 (20)	.84
Full time	15 (60)	15 (55.6)	19 (73.1)	13 (61.9)	15 (50)	77 (59.7)			
Part time	1 (4)	3 (11.1)	0 (0)	1 (4.8)	4 (13.3)	9 (7)			
Unemployed	1 (4)	2 (7.4)	1 (3.8)	1 (4.8)	2 (6.7)	7 (5.4)			
Retired	1 (4)	1 (3.7)	2 (7.7)	0 (0)	1 (3.3)	5 (3.9)			
Homemaker	1 (4)	0 (0)	0 (0)	0 (0)	2 (6.7)	3 (2.3)			
Student	6 (24)	6 (22.2)	4 (15.4)	6 (28.6)	6 (20)	28 (21.7)			
Monthly household income, HK $ (US $), n (%)							—	12.3 (20)	.71
<15,000 (1950)	7 (28)	9 (33.3)	4 (15.4)	4 (19)	9 (30)	33 (25.6)			
15,000 (1950)-24,999 (3249.87)	5 (20)	4 (14.8)	4 (15.4)	6 (28.6)	7 (23.3)	26 (20.2)			
25,000 (3250)-39,999 (5199.87)	6 (24)	6 (22.2)	11 (42.3)	5 (23.8)	5 (16.7)	33 (25.6)			
40,000 (5200)-59,999 (7799.87)	3 (12)	4 (14.8)	3 (11.5)	2 (9.5)	1 (3.3)	13 (10.1)			
>60,000 (7800)	4 (16)	4 (14.8)	2 (7.7)	4 (19)	7 (23.3)	21 (16.28)			

[a]dCBTi-therapist: digital cognitive behavioral therapy for insomnia with chatbot-based coaching and therapist support.

[b]dCBTi-assistant: digital cognitive behavioral therapy for insomnia with chatbot-based coaching and research assistant support.

[c]dCBTi-chatbot: digital cognitive behavioral therapy for insomnia with chatbot-based coaching only.

[d]dCBTi-unguided: digital cognitive behavioral therapy for insomnia without any coaching.

[e]dSH: digital sleep hygiene and self-monitoring control.

[f]Not applicable.

Table 3. Measures of outcomes by time points.

Variables and time points	dCBTi-therapist[a] (n=25), mean (SD)	dCBTi-assistant[b] (n=27), mean (SD)	dCBTi-chatbot[c] (n=26), mean (SD)	dCBTi-unguided[d] (n=21), mean (SD)	dSH[e] (n=30), mean (SD)	Full sample (n=129), mean (SD)	*F* test (*df*)	*P* value
ISI[f]								
Baseline	14.96 (3.65)	13.70 (3.86)	14.50 (4.96)	15.76 (5.17)	16.30 (4.39)	15.05 (4.44)	1.47 (4,124)	.22
After treatment	9.10 (4.23)[g]	9.00 (4.59)[g]	9.42 (4.44)	9.64 (5.00)	13.22 (5.07)[g]	10.28 (4.92)	3.66 (4,99)	.01
Follow-up	9.26 (4.85)	7.55 (4.86)	10.00 (5.32)	6.88 (4.67)	12.06 (6.20)	9.36 (5.43)	2.28 (4,75)	.07
SCI[h]								
Baseline	11.88 (4.22)	13.48 (5.28)	12.81 (5.21)	12.81 (5.01)	11.70 (4.62)	12.51 (4.85)	0.62 (4,124)	.65
After treatment	19.86 (4.90)[i]	18.30 (5.80)	19.16 (6.48)[i]	19.43 (7.55)	13.96 (6.03)[i]	17.80 (6.41)	3.81 (4,99)	.01
Follow-up	21.00 (6.38)	20.40 (7.00)	19.00 (6.11)	22.75 (4.71)	15.67 (8.13)	19.45 (7.00)	2.23 (4,75)	.07
PHQ-9[j]								
Baseline	11.60 (5.32)	9.52 (4.59)	10.16 (4.61)	12.86 (6.33)	10.38 (5.36)	10.80 (5.28)	1.50 (4,122)	.21
After treatment	7.10 (5.35)	6.35 (5.02)[k]	7.16 (3.86)	8.07 (5.41)	10.92 (6.36)[k]	8.04 (5.51)	2.77 (4,98)	.03
Follow-up	6.74 (5.43)	6.40 (5.50)	7.47 (4.66)	7.86 (4.74)	9.44 (6.24)	7.51 (5.44)	0.87 (4,74)	.48
GAD-7[l]								
Baseline	9.52 (5.55)	7.44 (5.06)	8.92 (4.56)	11.95 (5.67)	9.69 (5.86)	9.40 (5.46)	2.16 (4,122)	.08
After treatment	6.10 (3.99)	5.35 (5.01)[m]	6.05 (4.56)	8.79 (6.28)	9.54 (6.04)[m]	7.16 (5.40)	2.82 (4,98)	.03
Follow-up	7.42 (6.09)	6.05 (4.95)	6.87 (5.25)	9.43 (7.50)	7.33 (5.98)	7.13 (5.69)	0.48 (4,74)	.75
FAS[n]								
Baseline	32.24 (7.23)	28.74 (7.07)	28.04 (7.20)	30.81 (7.40)	31.00 (7.61)	30.15 (7.36)	1.42 (4,122)	.23
After treatment	25.81 (6.27)	25.61 (5.61)	26.74 (8.09)	27.64 (8.58)	30.54 (8.52)	27.38 (7.56)	1.77 (4,98)	.14
Follow-up	25.47 (8.20)	25.75 (6.89)	27.87 (7.37)	27.57 (7.02)	28.56 (6.96)	26.89 (7.27)	0.61 (4,74)	.66
ESS[o]								
Baseline	9.32 (4.63)	10.59 (4.76)	10.24 (4.38)	11.86 (4.61)	8.55 (5.10)	10.12 (4.77)	1.75 (4,122)	.14
After treatment	8.57 (4.51)	8.57 (4.53)	10.47 (4.23)	9.36 (5.68)	8.23 (5.22)	8.94 (4.80)	0.71 (4,98)	.59
Follow-up	7.16 (5.21)	8.30 (4.43)	10.67 (3.70)	11.29 (5.94)	7.33 (4.43)	8.52 (4.78)	2.13 (4,74)	.08
SWLS[p]								
Baseline	14.64 (6.25)	17.26 (6.17)	18.00 (6.43)	13.48 (5.20)	17.34 (5.25)	16.28 (6.04)	2.63 (4,122)	.04
After treatment	18.29 (7.24)	19.74 (6.05)	18.00 (6.72)	15.86 (6.31)	17.54 (5.46)	18.04 (6.32)	0.88 (4,98)	.48
Follow-up	19.16 (8.39)	20.00 (5.67)	19.47 (6.15)	19.43 (6.58)	18.17 (5.76)	19.23 (6.47)	0.19 (4,74)	.94
DBAS-16[q]								
Baseline	6.22 (1.23)	5.94 (1.71)	6.09 (1.20)	6.04 (1.18)	6.20 (1.68)	6.10 (1.42)	0.17 (4,124)	.95
After treatment	2.51 (2.27)[r]	3.33 (2.14)[r]	3.50 (2.57)[r]	2.73 (2.54)[r]	5.55 (2.74)[r]	3.63 (2.68)	6.67 (4,124)	<.001
Follow-up	2.55 (2.53)	2.57 (2.43)	2.99 (2.81)	1.87 (2.77)	3.30 (3.02)	2.71 (2.73)	0.96 (4,124)	.43
SRBQ[s]								

Variables and time points	dCBTi-therapist[a] (n=25), mean (SD)	dCBTi-assistant[b] (n=27), mean (SD)	dCBTi-chatbot[c] (n=26), mean (SD)	dCBTi-unguided[d] (n=21), mean (SD)	dSH[e] (n=30), mean (SD)	Full sample (n=129), mean (SD)	*F* test (*df*)	*P* value
Baseline	37.68 (9.21)	32.85 (11.50)	36.20 (9.46)	34.43 (10.40)	36.47 (10.30)	35.55 (10.21)	0.89 (4,123)	.47
After treatment	28.05 (11.11)	27.13 (10.15)[t]	29.11 (11.36)	30.57 (14.66)	38.73 (15.62)[t]	31.08 (13.32)	3.30 (4,98)	.01
Follow-up	24.11 (14.36)	25.00 (11.97)	30.53 (11.72)	29.38 (11.71)	34.44 (13.69)	28.39 (13.22)	1.97 (4,75)	.11

[a]dCBTi-therapist: digital cognitive behavioral therapy for insomnia with chatbot-based coaching and therapist support.

[b]dCBTi-assistant: digital cognitive behavioral therapy for insomnia with chatbot-based coaching and research assistant support.

[c]dCBTi-chatbot: digital cognitive behavioral therapy for insomnia with chatbot-based coaching only.

[d]dCBTi-unguided: digital cognitive behavioral therapy for insomnia without any coaching.

[e]dSH: digital sleep hygiene and self-monitoring control.

[f]ISI: Insomnia Severity Index.

[g]Significant differences between treatment groups in post hoc multiple comparisons (dCBTi-therapist vs dSH, *P*=.03; dCBTi-assistant vs dSH, *P*=.02).

[h]SCI: Sleep Condition Indicator.

[i]Significant differences between treatment group in post hoc multiple comparisons (dCBTi-therapist vs dSH, *P*=.01; dCBTi-chatbot vs dSH, *P*=.04).

[j]PHQ-9: Patient Health Questionnaire-9.

[k]Significant differences between treatment group in post hoc multiple comparisons (dCBTi-assistant vs dSH, *P*=.03).

[l]GAD-7: Generalized Anxiety Disorder-7.

[m]Significant differences between treatment group in post hoc multiple comparisons (dCBTi-assistant vs dSH, *P*=.05).

[n]FAS: Fatigue Assessment Scale.

[o]ESS: Epworth Sleepiness Scale.

[p]SWLS: Satisfaction With Life Scale.

[q]DBAS-16: Dysfunctional Beliefs and Attitudes About Sleep-16.

[r]Significant differences between treatment group in post hoc multiple comparisons (dCBTi-therapist vs dSH, *P*<.001; dCBTi-assistant vs dSH, *P*=.008; dCBTi-chatbot vs dSH, *P*=.02).

[s]SRBQ: Sleep-Related Behaviors Questionnaire.

[t]Significant differences between treatment group in post hoc multiple comparisons (dCBTi-assistant vs dSH, *P*=.02).

Treatment Efficacy

Hypothesis 1: dCBTi, Regardless of the Presence of Coaching, Would Be Efficacious for Improving Insomnia Symptoms Compared to dSH

As shown in Table 4, the condition-by-time interaction effects on the SCI, PHQ-9, GAD-7, FAS, DBAS-16, and SRBQ scores were significant, indicating that participants who received dCBTi had greater improvements in insomnia symptoms measured by the SCI and greater reductions in fatigue, depressive symptoms, anxiety symptoms, dysfunctional thoughts about sleep, and safety behaviors related to sleep than participants who received dSH (Figure 2). At follow-up, significant condition-by-time interaction effects were observed on the SCI, FAS, PHQ-9, SWLS, and SRBQ scores (Table 4), indicating greater improvements in these outcomes experienced by dCBTi recipients compared to dSH recipients (Figure 2). In addition, as shown in Table 5, the remission rate based on the ISI scores was 58% (45/77) in the dCBTi conditions, which was significantly greater than that in the dSH condition (6/27, 22%). The remission rate based on the SCI score was 36% (28/77) in the dCBTi conditions, which was significantly greater than that in the dSH condition (2/27, 7%). No significant differences were observed for the rates of achieving clinically meaningful differences in the PHQ-9, GAD-7, FAS, and ESS scores.

Table 4. Linear mixed models results.

Outcome measures and assessment time points	Interaction effects for dCBTi conditions[a] vs dSH[b]			Interaction effects for guided dCBTi conditions[c] vs dCBTi-unguided			Interaction effects for dCBTi with human support[d] vs dCBTi-chatbot			Interaction effects for dCBTi-therapist vs dCBTi-assistant		
	Estimate	*P* value	Cohen *d* (95% CI)	Estimate	*P* value	Cohen *d* (95% CI)	Estimate	*P* value	Cohen *d* (95% CI)	Estimate	*P* value	Cohen *d* (95% CI)
ISI[e]												
After treatment	–2.01	.06	–0.28 (–0.42 to –0.14)	0.24	.86	0.02 (–0.08 to 0.13)	–0.30	.81	–0.03 (–0.15 to 0.08)	–1.16	.41	–0.12 (–0.22 to –0.02)
Follow-up	–1.89	.12	–0.22 (–0.34 to –0.11)	2.57	.12	0.22 (0.13 to 0.30)	–1.54	.26	–0.16 (–0.27 to –0.06)	0.26	.86	0.03 (–0.07 to 0.12)
SCI[f]												
After treatment	*3.83[g]*	*.003*	*0.45 (0.33 to 0.56)*	–0.20	.90	–0.02 (–0.11 to 0.07)	–0.08	.96	–0.01 (–0.10 to 0.09)	2.94	.08	0.26 (0.17 to 0.34)
Follow-up	*3.51*	*.02*	0.35 (0.24 to 0.45)	–1.03	.60	–0.07 (–0.140 to –0.002)	2.25	.17	0.20 (0.11 to 0.29)	2.54	.15	0.21 (0.13 to 0.29)
PHQ-9[h]												
After treatment	*–4.12*	*<.001*	*–0.62 (–0.77 to–0.47)*	0.75	.55	0.09 (–0.03 to 0.21)	–0.40	.74	–0.05 (–0.18 to 0.08)	–1.47	.26	–0.17 (–0.29 to –0.06)
Follow-up	*–4.00*	*<.001*	*–0.52 (–0.65 to–0.39)*	1.65	.31	0.15 (0.06 to 0.24)	–1.44	.26	–0.17 (–0.29 to –0.05)	–2.16	.11	–0.24 (–0.35 to –0.13)
GAD-7[i]												
After treatment	*–2.56*	*.01*	*–0.40 (–0.56 to–0.25)*	–0.10	.93	–0.01 (–0.13 to 0.11)	0.33	.77	0.04 (–0.09 to 0.17)	–1.57	.21	–0.19 (–0.31 to –0.07)
Follow-up	–1.70	.12	–0.23 (–0.37 to –0.09)	1.14	.46	0.11 (0.01 to 0.20)	–0.58	.63	–0.07 (–0.19 to 0.05)	–1.06	.42	–0.12 (–0.24 to –0.01)
FAS[j]												
After treatment	*–2.87*	*.02*	*–0.35 (–0.47 to –0.23)*	–1.05	.50	–0.10 (–0.200 to –0.004)	–2.15	.14	–0.22 (–0.32 to –0.12)	–2.93	.07	–0.28 (–0.37 to –0.18)
Follow-up	*–4.01*	*.01*	*–0.42 (–0.53 to –0.33)*	1.43	.47	0.11 (0.03 to 0.18)	*–3.47*	*.03*	*–0.33 (–0.43 to –0.24)*	*–4.56*	*.01*	*–0.41 (–0.50 to –0.32)*
ESS[k]												

Outcome measures and assessment time points	Interaction effects for dCBTi conditions[a] vs dSH[b]			Interaction effects for guided dCBTi conditions[c] vs dCBTi-unguided			Interaction effects for dCBTi with human support[d] vs dCBTi-chatbot			Interaction effects for dCBTi-therapist vs dCBTi-assistant		
	Estimate	*P* value	Cohen *d* (95% CI)	Estimate	*P* value	Cohen *d* (95% CI)	Estimate	*P* value	Cohen *d* (95% CI)	Estimate	*P* value	Cohen *d* (95% CI)
After treatment	–0.81	.35	–0.14 (–0.31 to 0.03)	0.20	.85	0.03 (–0.11 to 0.16)	–1.47	.16	–0.21 (–0.36 to –0.07)	0.83	.47	0.11 (–0.02 to 0.24)
Follow-up	–0.88	.38	–0.13 (–0.28 to 0.02)	0.73	.61	0.08 (–0.03 to 0.18)	–1.48	.18	–0.20 (–0.33 to –0.06)	–0.15	.90	–0.02 (–0.15 to 0.11)
SWLS[l]												
After treatment	2.14	.07	0.27 (0.14 to 0.39)	0.40	.79	0.04 (–0.06 to 0.14)	1.94	.17	0.20 (0.10 to 0.31)	1.30	.40	0.13 (0.03 to 0.22)
Follow-up	*3.02*	*.03*	*0.33 (0.22 to 0.43)*	–1.88	.33	–0.14 (–0.22 to –0.07)	2.55	.10	0.25 (0.15 to 0.35)	1.75	.28	0.16 (0.07 to 0.25)
DBAS-16[m]												
After treatment	*–2.41*	*<.001*	*–0.53 (–0.75 to –0.31)*	0.34	.62	0.06 (–0.12 to 0.25)	–0.58	.39	–0.11 (–0.30 to 0.08)	–1.09	.16	–0.18 (–0.34 to –0.02)
Follow-up	–0.68	.24	–0.15 (–0.37 to 0.07)	0.79	.25	0.15 (–0.04 to 0.33)	–0.42	.53	–0.08 (–0.27 to 0.11)	–0.30	.70	–0.05 (–0.21 to 0.11)
SRBQ[n]												
After treatment	*–7.74*	*.001*	*–0.50 (–0.57 to –0.43)*	–3.64	.21	–0.18 (–0.23 to –0.13)	–0.90	.74	–0.05 (–0.10 to 0.01)	–2.30	.45	–0.11 (–0.16 to –0.06)
Follow-up	*–8.71*	*.001*	*–0.49 (–0.54 to –0.43)*	–2.17	.54	–0.09 (–0.13 to –0.05)	*–5.93*	*.05*	*–0.30 (–0.35 to –0.25)*	–5.70	.07	–0.27 (–0.32 to –0.22)

[a]Digital cognitive behavioral therapy for insomnia (dCBTi) with chatbot-based coaching and therapist support (dCBTi-therapist), dCBTi with chatbot-based coaching and research assistant support (dCBTi-assistant), dCBTi with chatbot-based coaching only (dCBTi-chatbot), and dCBTi without any coaching (dCBTi-unguided).

[b]dSH: digital sleep hygiene and self-monitoring control.

[c]dCBTi-therapist, dCBTi-assistant, and dCBTi-chatbot.

[d]dCBTi-therapist and dCBTi-assistant.

[e]ISI: Insomnia Severity Index.

[f]SCI: Sleep Condition Indicator.

[g]Italicization refers to significant results. See the respective columns of *P* value.

[h]PHQ-9: Patient Health Questionnaire-9.

[i]GAD-7: Generalized Anxiety Disorder-7.

[j]FAS: Fatigue Assessment Scale.

[k]ESS: Epworth Sleepiness Scale.

[l]SWLS: Satisfaction With Life Scale.

[m]DBAS-16: Dysfunctional Beliefs and Attitudes About Sleep-16.

[n]SRBQ: Sleep-Related Behaviors Questionnaire.

Figure 2. Changes in outcomes from baseline to follow-up. Error bars indicate the SEs, "a" indicates significant group-by-time effects of digital cognitive behavioral therapy for insomnia (dCBTi) versus digital sleep hygiene and self-monitoring control (dSH), "b" indicates significant group-by-time effects of dCBTi with human (therapist or research assistant) support versus dCBTi with chatbot-based coaching only (dCBTi-chatbot), and "c" indicates significant group-by-time effects of dCBTi with chatbot-based coaching and therapist support (dCBTi-therapist) versus dCBTi with chatbot-based coaching and research assistant support (dCBTi-assistant). DBAS-16: Dysfunctional Beliefs and Attitudes About Sleep-16; ESS: Epworth Sleepiness Scale; FAS: Fatigue Assessment Scale; GAD-7: Generalized Anxiety Disorder-7; ISI: Insomnia Symptom Index; PHQ-9: Patient Health Questionnaire-9; SCI: Sleep Condition Indicator; SRBQ: Sleep-Related Behaviors Questionnaire; SWLS: Satisfaction With Life Scale.

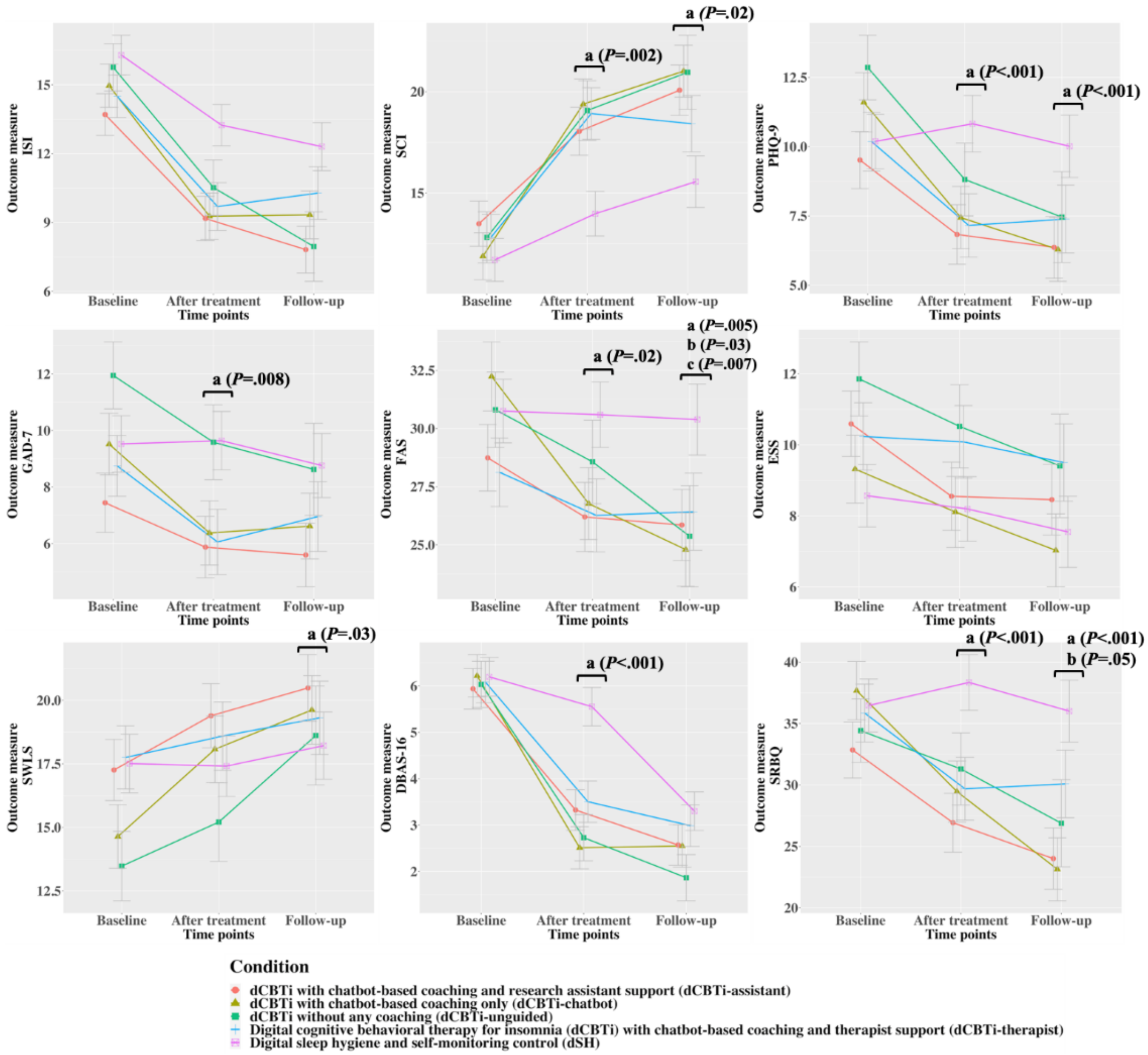


Table 5. Comparison of the rates of remission or clinically meaningful changes across conditions after treatment.

Outcome	Comparison between dCBTi conditions[a] and dSH[b]				Comparison between guided dCBTi conditions[c] and dCBTi-unguided				Comparison between dCBTi with human support[d] and dCBTi-chatbot				Comparison between dCBTi-therapist and dCBTi-assistant			
	dCBTi conditions (n=77), n (%)	dSH (n=27), n (%)	Chi-square (*df*)	*P* value	Guided dCBTi conditions (n=63), n (%)	dCBTi-unguided (n=14), n (%)	Chi-square (*df*)	*P* value	dCBTi with human support (n=44), n (%)	dCBTi-chatbot (n=19), n (%)	Chi-square (*df*)	*P* value	dCBTi-therapist (n=21), n (%)	dCBTi-assistant (n=23), n (%)	Chi-square (*df*)	*P* value
ISI[e]	*45 (58.44)*[f]	*6 (22.22)*	*9.1 (1)*	*.003*	37 (58.73)	8 (57.14)	<0.1 (1)	.99	26 (59.09)	11 (57.89)	<0.1 (1)	.99	13 (61.90)	13 (56.52)	<0.1 (1)	.96
SCI[g]	*28 (36.36)*	*2 (7.41)*	*6.8 (1)*	*.01*	22 (42.86)	6 (34.92)	0.1 (1)	.80	15 (34.09)	7 (36.84)	<0.01 (1)	.99	8 (38.10)	7 (30.43)	0.1 (1)	.83
PHQ-9[h]	28 (36.84)	4 (16)	2.9 (1)	.09	21 (33.87)	7 (50)	0.7 (1)	.41	17 (38.64)	4 (22.22)	0.9 (1)	.35	10 (47.62)	7 (30.43)	0.7 (1)	.39
GAD-7[i]	25 (32.89)	5 (20)	0.9 (1)	.33	20 (32.36)	5 (35.71)	<0.1 (1)	.99	13 (29.55)	7 (38.89)	0.2 (1)	.68	*10 (47.62)*	*3 (13.04)*	*4.8 (1)*	*.03*
FAS[j]	35 (46.05)	8 (32)	1.0 (1)	.32	30 (48.39)	5 (35.71)	0.3 (1)	.57	22 (50)	8 (44.44)	<0.1 (1)	.91	13 (61.90)	9 (39.13)	1.5 (1)	.23
ESS[k]	33 (43.42)	6 (24)	2.2 (1)	.14	28 (45.16)	5 (35.71)	0.1 (1)	.73	22 (50)	6 (33.33)	0.8 (1)	.36	10 (47.62)	12 (52.17)	<0.1 (1)	.99

[a]Digital cognitive behavioral therapy for insomnia (dCBTi) with chatbot-based coaching and therapist support (dCBTi-therapist), dCBTi with chatbot-based coaching and research assistant support (dCBTi-assistant), dCBTi with chatbot-based coaching only (dCBTi-chatbot), and dCBTi without any coaching (dCBTi-unguided).

[b]dSH: digital sleep hygiene and self-monitoring control.

[c]dCBTi-therapist, dCBTi-assistant, and dCBTi-chatbot.

[d]dCBTi-therapist and dCBTi-assistant.

[e]ISI: Insomnia Severity Index (criterion of remission: ISI score <10).

[f]Italicization refers to significant results. See the respective columns of *P* value.

[g]SCI: Sleep Condition Indicator (criterion of remission: SCI score >21).

[h]PHQ-9: Patient Health Questionnaire-9 (criterion of reaching clinically meaningful difference: 5-point change).

[i]GAD-7: Generalized Anxiety Disorder-7 (criterion of reaching clinically meaningful difference: 4-point change).

[j]FAS: Fatigue Assessment Scale (criterion of reaching clinically meaningful difference: 4-point change).

[k]ESS: Epworth Sleepiness Scale (criterion of reaching clinically meaningful difference: 2-point change).

Hypothesis 2: dCBTi With Coaching Would Promote Greater Improvements in Insomnia Symptoms and Greater Treatment Adherence Than dCBTi-Unguided

No significant interaction effects were found on all outcomes after treatment and at follow-up when comparing guided dCBTi and dCBTi-unguided (Table 4), suggesting that adding chatbot-based coaching and human support did not improve treatment efficacy. Similarly, the rates of remission of insomnia and the rates of achieving clinically meaningful changes in the secondary outcomes did not differ significantly between guided dCBTi and dCBTi-unguided (Table 5).

Hypothesis 3: dCBTi With Human Coaching Would Promote Greater Improvements in Insomnia Symptoms and Treatment Adherence Than dCBTi-Chatbot

Significant condition-by-time interaction effects were observed on the FAS and SRBQ scores at follow-up (Table 4), indicating that participants who received dCBTi-therapist or dCBTi-assistant experienced greater reductions in fatigue and sleep-related safety behaviors than those who received dCBTi-chatbot (Figure 2). The rates of remission of insomnia and the rates of achieving clinically meaningful changes in the secondary outcomes did not differ significantly between dCBTi with human support and dCBTi-chatbot (Table 5).

Hypothesis 4: dCBTi-Therapist Would Promote Greater Improvements in Insomnia Symptoms and Greater Treatment Adherence Than dCBTi-Assistant

A significant condition-by-time interaction effect was observed on the FAS scores at follow-up (Table 4), indicating that participants who received dCBTi-therapist experienced greater reductions in fatigue than those who received dCBTi-assistant (Figure 2). In addition, the rate of achieving clinically meaningful changes in the GAD-7 scores was significantly greater in dCBTi-therapist than in dCBTi-assistant (Table 5).

Treatment Adherence

Table 6 presents the results of treatment adherence across the conditions. As expected, participants in dCBTi-therapist and dCBTi-assistant completed significantly more video sessions than participants in dCBTi-unguided. They also completed more weeks of sleep diaries than participants in dSH. Significant differences in treatment adherence were observed especially in later sessions, with more participants in dCBTi-therapist completing sleep diaries during sessions 4 to 6 compared to those in dCBTi-unguided and dSH.

Table 6. Video completion and sleep diary completion across conditions.

Variables	dCBTi-therapist[a] (n=25)	dCBTi-assistant[b] (n=27)	dCBTi-chatbot[c] (n=26)	dCBTi-unguided[d] (n=21)	dSH[e] (n=30)	*F* test (*df*)	*P* value
Video sessions completed							
Total, mean (SD)	4.24 (1.83)[f]	3.37 (2.50)[f]	2.69 (2.24)	1.48 (2.02)[f]	—[g]	6.58 (3,95)	<.001
Session 1, n (%)	20 (80)[f]	18 (67)	15 (56)	10 (48)[f]	—	—	.03
Session 2, n (%)	19 (76)	18 (67)	14 (54)	9 (43)	—	—	.09
Session 3, n (%)	21 (84)[f]	17 (63)	12 (46)[f]	6 (29)[f]	—	—	<.001
Session 4, n (%)	18 (72)	14 (52)	12 (46)	5 (24)	—	—	.10
Session 5, n (%)	16 (64)[f]	14 (52)	10 (38)	5 (24)[f]	—	—	.01
Session 6, n (%)	16 (64)[f]	13 (48)[f]	8 (32)[f]	3 (14)[f]	—	—	<.001
Weeks of diaries completed							
Total, mean (SD)	5.32 (1.55)[h]	4.74 (1.87)	3.96 (2.57)	3.81 (2.18)	3.60 (2.33)[h]	2.95 (4,124)	.02
Week 1, n (%)	24 (96)	24 (89)	21 (81)	19 (90)	27 (90)	—	.57
Week 2, n (%)	21 (84)	25 (93)	19 (73)	17 (81)	22 (73)	—	.31
Week 3, n (%)	22 (88)	21 (78)	17 (65)	14 (67)	18 (60)	—	.15
Week 4, n (%)	22 (88)[h]	21 (78)	16 (62)	12 (571)	15 (50)[h]	—	.02
Week 5, n (%)	22 (88)[h]	18 (67)	16 (62)	10 (48)[h]	14 (47)[h]	—	.01
Week 6, n (%)	22 (88)[h]	19 (70)	14 (54)	8 (38)[h]	12 (39)[h]	—	.001

[a]dCBTi-therapist: digital cognitive behavioral therapy for insomnia with chatbot-based coaching and therapist support.

[b]dCBTi-assistant: digital cognitive behavioral therapy for insomnia with chatbot-based coaching and research assistant support.

[c]dCBTi-chatbot: digital cognitive behavioral therapy for insomnia with chatbot-based coaching only.

[d]dCBTi-unguided: digital cognitive behavioral therapy for insomnia without any coaching.

[e]dSH: digital sleep hygiene and self-monitoring control.

[f]Significant differences between treatment groups in the same row in post hoc multiple comparisons with adjustments for multiple tests (dCBTi-therapist vs dCBTi-unguided, *P*<.001; dCBTi-assistant vs dCBTi-unguided, *P*=.02).

[g]Not applicable.

[h]Significant differences between treatment groups (dCBTi-therapist vs dSH, *P*=.03).

Discussion

Principal Findings

This work presents the first randomized controlled comparative trial that evaluates the effects of chatbot-based coaching and human support on the treatment efficacy of, and adherence to, dCBTi. We found that participants who received dCBTi had greater improvements in insomnia symptoms (measured using the SCI), mood disturbances, fatigue, and life satisfaction as well as greater reductions in dysfunctional beliefs and safety behaviors related to insomnia than those who received dSH, with medium effect sizes comparable to those in previous studies of dCBTi [62-65]. Most of the improvements in the dCBTi conditions were sustained at 4-week follow-up. Surprisingly, adding chatbot-based coaching and human support did not significantly improve treatment effects on insomnia. Nonetheless, adding human support, especially therapist support, promoted greater improvements in fatigue as well as greater reduction in safety behaviors related to sleep. Adding chatbot-based coaching and human support also improved some indicators of treatment adherence.

Does Fully Automated dCBTi-Unguided Work?

Supporting hypothesis 1, dCBTi delivered by a fully automated mobile app is efficacious for improving insomnia, mood disturbances, fatigue, and quality of life in adults with insomnia, with effect sizes comparable to those of other tested versions of dCBTi [62-65]. Recipients of dCBTi, regardless of having coaching, achieved an average increase of 12% (SD 13.43%) in sleep efficiency after treatment, from 75.4% to 87.4%; noting that ≥85% sleep efficiency is considered remission of insomnia [45]. The remission rate reached 58% (45/77) in dCBTi conditions compared to 22% (6/27) in dSH. This study was one of the few randomized controlled trials of dCBTi conducted in non-Western populations, and this was the only dCBTi mobile app implemented in Cantonese with published efficacy. This study also extended previous findings by showing that dCBTi was also efficacious for reducing dysfunctional beliefs about sleep and maladaptive behaviors related to sleep—the mechanisms theorized to bring about the treatment effects in CBTi. This finding provided even stronger support for dCBTi by showing that it worked in a way that was consistent with the theory.

Unexpectedly, a greater reduction in insomnia symptoms in the dCBTi group was only reflected by the SCI scores but not the ISI scores. The SCI differs from the ISI in that its ratings on sleep difficulties are based on the recommended quantitative criteria from the *DSM-5* as opposed to qualitative impressions of insomnia symptom severity. The inconsistent results reflected by the scores on the two scales might suggest that quantitative anchors are more sensitive in detecting changes in insomnia symptoms. Nevertheless, the absence of a treatment effect of dCBTi on the ISI scores in this study differed from the findings of previous studies [62,65,66]. These previous studies used more stringent participant inclusion criteria; for instance, in addition to scoring ≥10 on the ISI, participants had to meet the duration (≥3 mo) and frequency (≥3 d/wk) diagnostic criteria of insomnia disorder. Participants in these studies might have had more chronic and severe insomnia to begin with and hence experienced greater improvements. Indeed, the ISI scores were 17 [65] and 19 [62] in prior studies and 15 in our sample. Our sample might have also included individuals with acute insomnia. As acute and subclinical insomnia could predict chronic insomnia and depressive episodes [67,68], the evidence for the efficacy of dCBTi for this group of participants with a potentially wider range of symptom duration and severity adds confidence for the impact of dCBTi at the population level where people with differing symptom duration and severity could benefit from dCBTi.

Does Adding Chatbot-Based Coaching and Human Support Improve dCBTi?

Partially consistent with hypothesis 2, adding coaching support improved treatment adherence to dCBTi but not efficacy. Our findings suggested that both human-assisted guidance and chatbot-based coaching were useful strategies to enhance engagement in the middle and late stages of dCBTi. The treatment adherence rates in the guided dCBTi conditions were double those of dCBTi-unguided. Among all types of guided dCBTi, dCBTi-therapist had the highest adherence rate, followed by dCBTi-assistant and dCBTi-chatbot. However, increased adherence rates were not associated with greater efficacy. Consistent with previous meta-analyses [24,69,70] showing that dCBTi was not inferior to face-to-face CBTi for alleviating insomnia, this study did not find any meaningful differences (a change of >4 points in the ISI total score) between guided CBTi conditions and dCBTi-unguided. This study was the first direct comparison of guided CBTi and dCBTi-unguided with different types of guidance and provided primary evidence indicating that adding either therapist or research assistant support does not promote meaningfully greater treatment efficacy.

It is possible that the high degrees of personalization offered by dCBTi might have minimized the benefit of coaching support on treatment efficacy; for instance, in the dCBTi-unguided condition, participants still received a tailored sleep schedule suggestion based on their diary-reported sleep data in the previous week. They were also prompted to set up individualized weekly goals and action plans. With mobile technology, even dCBTi-unguided could deliver tailored treatment recommendations, which is one of the promising benefits of dCBTi. Furthermore, our sample might have included participants with acute insomnia, and the insomnia symptoms experienced by these individuals might not necessitate coaching support. In addition, our sample is overrepresented by highly educated young adults (104/129, 80.6%) with few psychiatric comorbidities. Coaching support may not be most needed for this population; adding therapist support may be beneficial specifically for patients with psychiatric comorbidities [71]. Nevertheless, the differences in adherence between guided dCBTi and dCBTi-unguided reflected the utility of coaching support for enhancing engagement and potentially reducing early dropouts and motivational barriers.

Similarly, partially consistent with hypothesis 3, adding human support did not promote greater improvements in insomnia and most outcomes compared to dCBTi-chatbot. However, greater improvements in fatigue and greater reductions in safety behaviors related to sleep were observed in dCBTi with human support compared to dCBTi-chatbot. These incremental benefits might promote greater or more sustained improvements in sleep and well-being in the long term because reduced fatigue and safety behaviors related to sleep could potentially enhance the maintenance of positive changes resulting from dCBTi, such as maintaining adequate daytime activities and inhibiting anxiety and frustration about sleep. Indeed, as shown in a previous study, adding human support to a self-help CBTi did not lead to greater improvements in insomnia symptoms after treatment and at 4-week follow-up; however, the incremental improvements appeared later at the 3-month follow-up [29].

Is Support From a Therapist Better Than That From a Research Assistant?

Inconsistent with hypothesis 4, support from a therapist did not promote superior treatment efficacy for most outcomes or superior treatment adherence compared to support from a research assistant. While this study was the first to directly compare human therapist support and research assistant support in dCBTi, our results converged with a prior study on digital intervention for depression to suggest that treatment efficacy was comparable between therapist-guided and nontherapist-guided digital interventions [72]. However, it should be noted that the dCBTi-assistant telephone call completion rate was much lower than the dCBTi-therapist telephone call rate, suggesting that therapist support was much more welcomed by the participants in comparison. The lack of differences between these two conditions could be explained by the lack of statistical power for detecting smaller effects. There were no prior data on the expected effect size for the difference between dCBTi-therapist and dCBTi-assistant. This study could have missed smaller effects between these two conditions. Indeed, as shown in Table 4, differences between dCBTi-therapist and dCBTi-assistant amounting to small effect sizes were observed in the SCI, PHQ-9, GAD-7, FAS, and SRBQ scores. Future studies with larger sample sizes statistically powered to detect small effects are needed to further elucidate whether therapist support promotes incremental treatment efficacy.

Limitations

Our findings need to be interpreted in light of the following limitations. First, our sample consisted of mostly highly educated young adults (104/129, 80.6%); therefore, the findings may not be generalizable to other populations. While dCBTi was found to be effective across demographic groups [67], all studies were conducted in samples of working-age adults [68]. It remains unclear whether older adults also respond as well to dCBTi. Second, although this study was adequately powered for detecting meaningful differences in the ISI scores, it could not detect smaller effects that might have existed in the comparison between dCBTi conditions with different coaching types. Nonetheless, we argue that such differences would have limited practical and clinical implications. Third, the follow-up assessment was conducted 4 weeks after treatment, thereby limiting any conclusions that could be drawn about the long-term efficacy of the dCBTi intervention. In particular, although we did not find significant meaningful differences in the primary outcomes between dCBTi conditions with different types of coaching at short-term follow-up, dCBTi with coaching, especially dCBTi-therapist, performed better than dCBTi without coaching on the mechanism of action, that is, sleep-related safety behaviors. This greater improvement in the mechanism of action may promote incremental benefits on the primary outcome that appear at a longer follow-up. Future studies with longer follow-up are necessary to fully evaluate the potential benefits of adding coaching to dCBTi. Finally, we did not collect data on participants' use of strategies learned in dCBTi. Although adding therapist support improved video session and sleep diary completions, it remains unclear whether the addition of such support increased participants' use of the learned strategies in their daily lives. More detailed assessments of adherence would provide greater insights into the relationship between treatment adherence and efficacy or the lack thereof.

Conclusions

Our findings supported the efficacy of a fully automated dCBTi intervention, Sleep Sensei, compared to an active control for treating insomnia, reducing mood disturbances and fatigue, and improving quality of life. Adding chatbot-based coaching and human support did not significantly improve the efficacy of Sleep Sensei for treating insomnia, but doing so may improve long-term efficacy, given their effects on increasing treatment adherence and additional benefits on reducing fatigue and behaviors that could perpetuate insomnia. In sum, Sleep Sensei can be used as a stand-alone intervention for treating insomnia and is the only Cantonese mobile app for CBTi published with demonstrated efficacy.

Acknowledgments

The authors thank Christy Li Lok Yan for assistance in the proofreading and language editing of the manuscript. This study was funded by a research development grant awarded to WSC.

Authors' Contributions

WSC contributed to conceptualization, methodology, formal analysis, resources, writing the original draft, reviewing and editing the manuscript, supervision, and funding acquisition. WYC contributed to software, formal analysis, reviewing and editing the manuscript, and visualization. SHCL and AKMC contributed to investigation and reviewing and editing the manuscript. AKWL contributed to data curation, project administration, and reviewing and editing the manuscript. ASYN contributed to reviewing and editing the manuscript and visualization. TK contributed to resources and reviewing and editing the manuscript.

Conflicts of Interest

TK is affiliated with the Centre for Digital Health Interventions, a joint initiative of the Institute for Implementation Science in Health Care at the University of Zurich; the Department of Management, Technology, and Economics at ETH Zurich; the Future Health Technologies Program at the Singapore-ETH Centre; and the School of Medicine and Institute of Technology Management at the University of St Gallen. The Centre for Digital Health Interventions is funded in part by CSS, a Swiss health insurer; Mavie Next (owned by the UNIQA Group), an Austrian care provider; and MTIP, a Swiss investor company. TK is also a cofounder of Pathmate Technologies, a university spin-off company that creates and delivers digital clinical pathways. However, CSS, Pathmate Technologies, Mavie Next, and MTIP were not involved in this research. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
Screenshots of the Sleep Sensei interface.
[PDF File (Adobe PDF File), 1639 KB - mental_v11i1e51716_app1.pdf]

Multimedia Appendix 2
Consort Ehealth checklist.
[PDF File (Adobe PDF File), 13886 KB - mental_v11i1e51716_app2.pdf]

References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders Fifth Edition (DSM-5). Washington, DC: American Psychiatric Association Publishing; 2013.
2. The AASM International Classification of Sleep Disorders – Third Edition, Text Revision (ICSD-3-TR). American Academy of Sleep Medicine. URL: https://aasm.org/clinical-resources/international-classification-sleep-disorders/ [accessed 2024-06-17]
3. Aernout E, Benradia I, Hazo JB, Sy A, Askevis-Leherpeux F, Sebbane D, et al. International study of the prevalence and factors associated with insomnia in the general population. Sleep Med 2021 Jun;82:186-192. [doi: 10.1016/j.sleep.2021.03.028] [Medline: 33957414]
4. Lam CS, Yu BY, Cheung DS, Cheung T, Lam SC, Chung KF, et al. Sleep and mood disturbances during the COVID-19 outbreak in an urban Chinese population in Hong Kong: a longitudinal study of the second and third waves of the outbreak. Int J Environ Res Public Health 2021 Aug 10;18(16):8444 [FREE Full text] [doi: 10.3390/ijerph18168444] [Medline: 34444192]
5. Morin CM, Bjorvatn B, Chung F, Holzinger B, Partinen M, Penzel T, et al. Insomnia, anxiety, and depression during the COVID-19 pandemic: an international collaborative study. Sleep Med 2021 Nov;87:38-45 [FREE Full text] [doi: 10.1016/j.sleep.2021.07.035] [Medline: 34508986]
6. Meaklim H, Junge MF, Varma P, Finck WA, Jackson ML. Pre-existing and post-pandemic insomnia symptoms are associated with high levels of stress, anxiety, and depression globally during the COVID-19 pandemic. J Clin Sleep Med 2021 Oct 01;17(10):2085-2097 [FREE Full text] [doi: 10.5664/jcsm.9354] [Medline: 33985647]
7. Streatfeild J, Smith J, Mansfield D, Pezzullo L, Hillman D. The social and economic cost of sleep disorders. Sleep 2021 Nov 12;44(11):zsab132. [doi: 10.1093/sleep/zsab132] [Medline: 34015136]
8. Daley M, Morin CM, LeBlanc M, Grégoire JP, Savard J. The economic burden of insomnia: direct and indirect costs for individuals with insomnia syndrome, insomnia symptoms, and good sleepers. Sleep 2009 Jan;32(1):55-64 [FREE Full text] [Medline: 19189779]
9. Wickwire EM, Shaya FT, Scharf SM. Health economics of insomnia treatments: the return on investment for a good night's sleep. Sleep Med Rev 2016 Dec;30:72-82. [doi: 10.1016/j.smrv.2015.11.004] [Medline: 26874067]
10. Hafner M, Romanelli RJ, Yerushalmi E, Troxel WM. The societal and economic burden of insomnia in adults: an international study. RAND Europe. 2023. URL: https://www.rand.org/content/dam/rand/pubs/research_reports/RRA2100/RRA2166-1/RAND_RRA2166-1.pdf [accessed 2023-07-31]
11. Natsky AN, Vakulin A, Chai-Coetzer CL, Lack L, McEvoy RD, Lovato N, et al. Economic evaluation of cognitive behavioural therapy for insomnia (CBT-I) for improving health outcomes in adult populations: a systematic review. Sleep Med Rev 2020 Dec;54:101351. [doi: 10.1016/j.smrv.2020.101351] [Medline: 32739824]
12. Buntrock C, Lehr D, Smit F, Horvath H, Berking M, Spiegelhalder K, et al. Guided internet-based cognitive behavioral therapy for insomnia: health-economic evaluation from the societal and public health care perspective alongside a randomized controlled trial. J Med Internet Res 2021 May 24;23(5):e25609 [FREE Full text] [doi: 10.2196/25609] [Medline: 34028361]
13. Edinger JD, Arnedt JT, Bertisch SM, Carney CE, Harrington JJ, Lichstein KL, et al. Behavioral and psychological treatments for chronic insomnia disorder in adults: an American academy of sleep medicine clinical practice guideline. J Clin Sleep Med 2021 Feb;17(2):255-262 [FREE Full text] [doi: 10.5664/jcsm.8986] [Medline: 33164742]
14. Riemann D, Baglioni C, Bassetti C, Bjorvatn B, Dolenc Groselj L, Ellis JG, et al. European guideline for the diagnosis and treatment of insomnia. J Sleep Res 2017 Dec;26(6):675-700. [doi: 10.1111/jsr.12594] [Medline: 28875581]
15. Spielman AJ, Caruso LS, Glovinsky PB. A behavioral perspective on insomnia treatment. Psychiatr Clin North Am 1987 Dec;10(4):541-553. [Medline: 3332317]
16. Harvey AG. A cognitive model of insomnia. Behav Res Ther 2002 Aug;40(8):869-893. [doi: 10.1016/s0005-7967(01)00061-4] [Medline: 12186352]
17. Riemann D, Spiegelhalder K, Feige B, Voderholzer U, Berger M, Perlis M, et al. The hyperarousal model of insomnia: a review of the concept and its evidence. Sleep Med Rev 2010 Feb;14(1):19-31. [doi: 10.1016/j.smrv.2009.04.002] [Medline: 19481481]
18. Morin CM. Cognitive-behavioral approaches to the treatment of insomnia. J Clin Psychiatry 2004;65 Suppl 16:33-40. [Medline: 15575803]

19. Trauer JM, Qian MY, Doyle JS, Rajaratnam SM, Cunnington D. Cognitive behavioral therapy for chronic insomnia: a systematic review and meta-analysis. Ann Intern Med 2015 Aug 04;163(3):191-204. [doi: 10.7326/M14-2841] [Medline: 26054060]
20. Chan WS, McCrae CS, Ng AS. Is cognitive behavioral therapy for insomnia effective for improving sleep duration in individuals with insomnia? A meta-analysis of randomized controlled trials. Ann Behav Med 2023 May 23;57(6):428-441. [doi: 10.1093/abm/kaac061] [Medline: 36461882]
21. Alimoradi Z, Jafari E, Broström A, Ohayon MM, Lin CY, Griffiths MD, et al. Effects of cognitive behavioral therapy for insomnia (CBT-I) on quality of life: a systematic review and meta-analysis. Sleep Med Rev 2022 Aug;64:101646 [FREE Full text] [doi: 10.1016/j.smrv.2022.101646] [Medline: 35653951]
22. Lin YH, Lin SH, Chen IM. The direct effect of cognitive behavioral therapy for insomnia on depression prevention and the mediation effect via insomnia remission. JAMA Psychiatry 2022 May 01;79(5):514-515. [doi: 10.1001/jamapsychiatry.2022.0149] [Medline: 35293968]
23. Koffel E, Amundson E, Polusny G, Wisdom JP. "You're missing out on something great": patient and provider perspectives on increasing the use of cognitive behavioral therapy for insomnia. Behav Sleep Med 2020;18(3):358-371 [FREE Full text] [doi: 10.1080/15402002.2019.1591958] [Medline: 30907144]
24. Soh HL, Ho RC, Ho CS, Tam WW. Efficacy of digital cognitive behavioural therapy for insomnia: a meta-analysis of randomised controlled trials. Sleep Med 2020 Nov;75:315-325. [doi: 10.1016/j.sleep.2020.08.020] [Medline: 32950013]
25. Zachariae R, Lyby MS, Ritterband LM, O'Toole MS. Efficacy of internet-delivered cognitive-behavioral therapy for insomnia - a systematic review and meta-analysis of randomized controlled trials. Sleep Med Rev 2016 Dec;30:1-10. [doi: 10.1016/j.smrv.2015.10.004] [Medline: 26615572]
26. Kallestad H, Saksvik S, Vedaa Ø, Langsrud K, Morken G, Lydersen S, et al. Digital cognitive-behavioural therapy for insomnia compared with digital patient education about insomnia in individuals referred to secondary mental health services in Norway: protocol for a multicentre randomised controlled trial. BMJ Open 2021 Jun 28;11(6):e050661 [FREE Full text] [doi: 10.1136/bmjopen-2021-050661] [Medline: 34183350]
27. Darden M, Espie CA, Carl JR, Henry AL, Kanady JC, Krystal AD, et al. Cost-effectiveness of digital cognitive behavioral therapy (Sleepio) for insomnia: a Markov simulation model in the United States. Sleep 2021 Apr 09;44(4):zsaa223. [doi: 10.1093/sleep/zsaa223] [Medline: 33151330]
28. Hasan F, Tu YK, Yang CM, James Gordon C, Wu D, Lee HC, et al. Comparative efficacy of digital cognitive behavioral therapy for insomnia: a systematic review and network meta-analysis. Sleep Med Rev 2022 Feb;61:101567. [doi: 10.1016/j.smrv.2021.101567] [Medline: 34902820]
29. Ho FY, Chung KF, Yeung WF, Ng TH, Cheng SK. Weekly brief phone support in self-help cognitive behavioral therapy for insomnia disorder: relevance to adherence and efficacy. Behav Res Ther 2014 Dec;63:147-156. [doi: 10.1016/j.brat.2014.10.002] [Medline: 25461790]
30. Beun RJ, Fitrianie S, Griffioen-Both F, Spruit S, Horsch C, Lancee J, et al. Talk and tools: the best of both worlds in mobile user interfaces for E-coaching. Pers Ubiquitous Comput 2017 May 19;21:661-674. [doi: 10.1007/s00779-017-1021-5]
31. Luik AI, van der Zweerde T, van Straten A, Lancee J. Digital delivery of cognitive behavioral therapy for insomnia. Curr Psychiatry Rep 2019 Jun 04;21(7):50 [FREE Full text] [doi: 10.1007/s11920-019-1041-0] [Medline: 31161406]
32. van Straten A, Lancee J. Digital cognitive behavioural therapy for insomnia: the answer to a major public health issue? Lancet Digit Health 2020 Aug;2(8):e381-e382 [FREE Full text] [doi: 10.1016/S2589-7500(20)30167-9] [Medline: 33328041]
33. Horsch C, Lancee J, Beun RJ, Neerincx MA, Brinkman WP. Adherence to technology-mediated insomnia treatment: a meta-analysis, interviews, and focus groups. J Med Internet Res 2015 Sep 04;17(9):e214 [FREE Full text] [doi: 10.2196/jmir.4115] [Medline: 26341671]
34. Agnew S, Vallières A, Hamilton A, McCrory S, Nikolic M, Kyle SD, et al. Adherence to cognitive behavior therapy for insomnia: an updated systematic review. Sleep Med Clin 2021 Mar;16(1):155-202 [FREE Full text] [doi: 10.1016/j.jsmc.2020.11.002] [Medline: 33485527]
35. Filler A, Kowatsch T, Haug S, Wahle F, Staake T, Fleisch E. MobileCoach: a novel open source platform for the design of evidence-based, scalable and low-cost behavioral health interventions: overview and preliminary evaluation in the public health context. In: Proceedings of the 2015 Wireless Telecommunications Symposium. 2015 Presented at: WTS 2015; April 15-17, 2015; New York, NY. [doi: 10.1109/wts.2015.7117255]
36. Kowatsch T, Volland D, Shih I, Rüegger D, Künzler F, Barata F, et al. Design and evaluation of a mobile chat app for the open source behavioral health intervention platform MobileCoach. In: Proceedings of the 12th International Conference Designing the Digital Transformation. 2017 Presented at: DESRIST 2017; May 30-June 1, 2017; Karlsruhe, Germany. [doi: 10.1007/978-3-319-59144-5_36]
37. Morin CM, Belleville G, Bélanger L, Ivers H. The Insomnia Severity Index: psychometric indicators to detect insomnia cases and evaluate treatment response. Sleep 2011 May 01;34(5):601-608 [FREE Full text] [doi: 10.1093/sleep/34.5.601] [Medline: 21532953]
38. Netzer NC, Stoohs RA, Netzer CM, Clark K, Strohl KP. Using the Berlin Questionnaire to identify patients at risk for the sleep apnea syndrome. Ann Intern Med 1999 Oct 05;131(7):485-491. [doi: 10.7326/0003-4819-131-7-199910050-00002] [Medline: 10507956]

39. Yang M, Morin CM, Schaefer K, Wallenstein GV. Interpreting score differences in the Insomnia Severity Index: using health-related outcomes to define the minimally important difference. Curr Med Res Opin 2009 Oct;25(10):2487-2494. [doi: 10.1185/03007990903167415] [Medline: 19689221]
40. Kumle L, Võ ML, Draschkow D. Estimating power in (generalized) linear mixed models: an open introduction and tutorial in R. Behav Res Methods 2021 Dec;53(6):2528-2543 [FREE Full text] [doi: 10.3758/s13428-021-01546-0] [Medline: 33954914]
41. Chung HK, Louie K, Chan WS. Development and evaluation of a Chinese short-form of the sleep-related behaviors questionnaire in Hong Kong Chinese adults using item response theory. J Health Psychol 2024 Mar;29(4):255-265. [doi: 10.1177/13591053231195518] [Medline: 37688382]
42. Kaldo V, Jernelöv S, Blom K, Ljótsson B, Brodin M, Jörgensen M, et al. Guided internet cognitive behavioral therapy for insomnia compared to a control treatment - a randomized trial. Behav Res Ther 2015 Aug;71:90-100 [FREE Full text] [doi: 10.1016/j.brat.2015.06.001] [Medline: 26091917]
43. Buysse DJ, Reynolds CF3, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res 1989 May;28(2):193-213. [doi: 10.1016/0165-1781(89)90047-4] [Medline: 2748771]
44. Tsai PS, Wang SY, Wang MY, Su CT, Yang TT, Huang CJ, et al. Psychometric evaluation of the Chinese version of the Pittsburgh Sleep Quality Index (CPSQI) in primary insomnia and control subjects. Qual Life Res 2005 Oct;14(8):1943-1952. [doi: 10.1007/s11136-005-4346-x] [Medline: 16155782]
45. Edinger JD, Bonnet MH, Bootzin RR, Doghramji K, Dorsey CM, Espie CA, et al. Derivation of research diagnostic criteria for insomnia: report of an American Academy of Sleep Medicine Work Group. Sleep 2004 Dec 15;27(8):1567-1596. [doi: 10.1093/sleep/27.8.1567] [Medline: 15683149]
46. Lichstein KL, Durrence HH, Taylor DJ, Bush AJ, Riedel BW. Quantitative criteria for insomnia. Behav Res Ther 2003 Apr;41(4):427-445. [doi: 10.1016/s0005-7967(02)00023-2] [Medline: 12643966]
47. Wong ML, Lau KN, Espie CA, Luik AI, Kyle SD, Lau EY. Psychometric properties of the Sleep Condition Indicator and Insomnia Severity Index in the evaluation of insomnia disorder. Sleep Med 2017 May;33:76-81. [doi: 10.1016/j.sleep.2016.05.019] [Medline: 28449911]
48. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
49. Yu X, Tam WW, Wong PT, Lam TH, Stewart SM. The Patient Health Questionnaire-9 for measuring depressive symptoms among the general population in Hong Kong. Compr Psychiatry 2012 Jan;53(1):95-102. [doi: 10.1016/j.comppsych.2010.11.002] [Medline: 21193179]
50. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
51. Tong X, An D, McGonigal A, Park SP, Zhou D. Validation of the Generalized Anxiety Disorder-7 (GAD-7) among Chinese people with epilepsy. Epilepsy Res 2016 Feb;120:31-36. [doi: 10.1016/j.eplepsyres.2015.11.019] [Medline: 26709880]
52. Michielsen HJ, De Vries J, Van Heck GL, Van de Vijver FJ, Sijtsma K. Examination of the dimensionality of fatigue. Eur J Psychol Assess 2004 Jan;20(1):39. [doi: 10.1027/1015-5759.20.1.39]
53. Ho LY, Lai CK, Ng SS. Measuring fatigue following stroke: the Chinese version of the Fatigue Assessment Scale. Disabil Rehabil 2021 Nov;43(22):3234-3241. [doi: 10.1080/09638288.2020.1730455] [Medline: 32142618]
54. Chen NH, Johns MW, Li HY, Chu CC, Liang SC, Shu YH, et al. Validation of a Chinese version of the Epworth sleepiness scale. Qual Life Res 2002 Dec;11(8):817-821. [doi: 10.1023/a:1020818417949] [Medline: 12482165]
55. Pavot W, Diener E. Review of the satisfaction with life scale. Psychol Assess 1993;5(2):164-172. [doi: 10.1037/1040-3590.5.2.164]
56. Wang KT, Yuen M, Slaney RB. Perfectionism, depression, loneliness, and life satisfaction: a study of high school students in Hong Kong. Counsel Psychol 2008 Jul 31;37(2):249-274. [doi: 10.1177/0011000008315975]
57. Morin CM, Vallières A, Ivers H. Dysfunctional beliefs and attitudes about sleep (DBAS): validation of a brief version (DBAS-16). Sleep 2007 Nov;30(11):1547-1554 [FREE Full text] [doi: 10.1093/sleep/30.11.1547] [Medline: 18041487]
58. Chung KF, Ho FY, Yeung WF. Psychometric comparison of the full and abbreviated versions of the dysfunctional beliefs and attitudes about sleep scale. J Clin Sleep Med 2016 Jun 15;12(6):821-828 [FREE Full text] [doi: 10.5664/jcsm.5878] [Medline: 26857054]
59. Ree MJ, Harvey AG. Investigating safety behaviours in insomnia: the development of the sleep-related behaviours questionnaire (SRBQ). Behav change 2012 Feb 22;21(1):26-36. [doi: 10.1375/bech.21.1.26.35971]
60. Mellor A, Kavaliotis E, Mascaro L, Drummond SP. Approaches to the assessment of adherence to CBT-I, predictors of adherence, and the association of adherence to outcomes: a systematic review. Sleep Med Rev 2022 Jun;63:101620. [doi: 10.1016/j.smrv.2022.101620] [Medline: 35398650]
61. Corbeil RR, Searle SR. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. Technometrics 1976 Feb;18(1):31-38. [doi: 10.2307/1267913]

62. Rajabi Majd N, Broström A, Ulander M, Lin CY, Griffiths MD, Imani V, et al. Efficacy of a theory-based cognitive behavioral technique app-based intervention for patients with insomnia: randomized controlled trial. J Med Internet Res 2020 Apr 01;22(4):e15841 [FREE Full text] [doi: 10.2196/15841] [Medline: 32234700]
63. Cheng P, Luik AI, Fellman-Couture C, Peterson E, Joseph CL, Tallent G, et al. Efficacy of digital CBT for insomnia to reduce depression across demographic groups: a randomized trial. Psychol Med 2019 Feb;49(3):491-500 [FREE Full text] [doi: 10.1017/S0033291718001113] [Medline: 29792241]
64. Batterham PJ, Christensen H, Mackinnon AJ, Gosling JA, Thorndike FP, Ritterband LM, et al. Trajectories of change and long-term outcomes in a randomised controlled trial of internet-based insomnia treatment to prevent depression. BJPsych Open 2017 Sep 25;3(5):228-235 [FREE Full text] [doi: 10.1192/bjpo.bp.117.005231] [Medline: 28959453]
65. Kuhn E, Miller KE, Puran D, Wielgosz J, YorkWilliams SL, Owen JE, et al. A pilot randomized controlled trial of the insomnia coach mobile app to assess its feasibility, acceptability, and potential efficacy. Behav Ther 2022 May;53(3):440-457. [doi: 10.1016/j.beth.2021.11.003] [Medline: 35473648]
66. Horsch CH, Lancee J, Griffioen-Both F, Spruit S, Fitrianie S, Neerincx MA, et al. Mobile phone-delivered cognitive behavioral therapy for insomnia: a randomized waitlist controlled trial. J Med Internet Res 2017 Apr 11;19(4):e70 [FREE Full text] [doi: 10.2196/jmir.6524] [Medline: 28400355]
67. Ellis JG, Cushing T, Germain A. Treating acute insomnia: a randomized controlled trial of a "single-shot" of cognitive behavioral therapy for insomnia. Sleep 2015 Jun 01;38(6):971-978 [FREE Full text] [doi: 10.5665/sleep.4752] [Medline: 25515106]
68. Ellis JG, Gehrman P, Espie CA, Riemann D, Perlis ML. Acute insomnia: current conceptualizations and future directions. Sleep Med Rev 2012 Feb;16(1):5-14. [doi: 10.1016/j.smrv.2011.02.002] [Medline: 21596596]
69. van Straten A, Cuijpers P. Self-help therapy for insomnia: a meta-analysis. Sleep Med Rev 2009 Feb;13(1):61-71. [doi: 10.1016/j.smrv.2008.04.006] [Medline: 18952469]
70. Deng W, M J J van der Kleij R, Shen H, Wei J, Brakema EA, Guldemond N, et al. eHealth-based psychosocial interventions for adults with insomnia: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2023 Mar 14;25:e39250 [FREE Full text] [doi: 10.2196/39250] [Medline: 36917145]
71. van der Zweerde T, Lancee J, Ida Luik A, van Straten A. Internet-delivered cognitive behavioral therapy for insomnia: tailoring cognitive behavioral therapy for insomnia for patients with chronic insomnia. Sleep Med Clin 2020 Jun;15(2):117-131. [doi: 10.1016/j.jsmc.2020.02.001] [Medline: 32386688]
72. Titov N, Andrews G, Davies M, McIntyre K, Robinson E, Solley K. Internet treatment for depression: a randomized controlled trial comparing clinician vs. technician assistance. PLoS One 2010 Jun 08;5(6):e10939 [FREE Full text] [doi: 10.1371/journal.pone.0010939] [Medline: 20544030]

Abbreviations

CBTi: cognitive behavioral therapy for insomnia
CONSORT: Consolidated Standards of Reporting Trials
DBAS-16: Dysfunctional Beliefs and Attitudes About Sleep-16
dCBTi: digital cognitive behavioral therapy for insomnia
dCBTi-assistant: digital cognitive behavioral therapy for insomnia with chatbot-based coaching and research assistant support
dCBTi-chatbot: digital cognitive behavioral therapy for insomnia with chatbot-based coaching only
dCBTi-therapist: digital cognitive behavioral therapy for insomnia with chatbot-based coaching and therapist support
dCBTi-unguided: digital cognitive behavioral therapy for insomnia without any coaching
dSH: digital sleep hygiene and self-monitoring control
DSM-5: Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
ESS: Epworth Sleepiness Scale
FAS: Fatigue Assessment Scale
GAD-7: Generalized Anxiety Disorder-7
ISI: Insomnia Severity Index
PHQ-9: Patient Health Questionnaire-9
SCI: Sleep Condition Indicator
SRBQ: Sleep-Related Behaviors Questionnaire
SWLS: Satisfaction With Life Scale
TIB: time in bed
TST: total sleep time

Edited by J Torous; submitted 09.08.23; peer-reviewed by K Spiegelhalder, CB Miller, A Baharav, K Coteur; comments to author 01.11.23; revised version received 07.03.24; accepted 25.05.24; published 07.08.24.

Please cite as:
Chan WS, Cheng WY, Lok SHC, Cheah AKM, Lee AKW, Ng ASY, Kowatsch T
Assessing the Short-Term Efficacy of Digital Cognitive Behavioral Therapy for Insomnia With Different Types of Coaching: Randomized Controlled Comparative Trial
JMIR Ment Health 2024;11:e51716
URL: https://mental.jmir.org/2024/1/e51716
doi:10.2196/51716
PMID:

Social Media Use in Adolescents: Bans, Benefits, and Emotion Regulation Behaviors

Kelsey L McAlister, MS; Clare C Beatty, MA; Jacqueline E Smith-Caswell, MS; Jacqlyn L Yourell, PhD; Jennifer L Huberty, PhD

Fit Minded, Inc, Phoenix, AZ, United States

Corresponding Author:
Kelsey L McAlister, MS

Abstract

Social media is an integral part of adolescents' daily lives, but the significant time they invest in social media has raised concerns about the effect on their mental health. Bans and severe restrictions on social media use are quickly emerging as an attempt to regulate social media use; however, evidence supporting their effectiveness is limited. Adolescents experience several benefits from social media, including increased social connection, reduced loneliness, and a safe space for marginalized groups (eg, LGBTQ+) to interact. Rather than enforcing bans and severe restrictions, emotion regulation should be leveraged to help adolescents navigate the digital social environment. This viewpoint paper proposes a nuanced approach toward regulating adolescent social media use by (1) discontinuing the use of ineffective bans, (2) recognizing the benefits social media use can have, and (3) fostering emotion regulation skills in adolescents to encourage the development of self-regulation.

(JMIR Ment Health 2024;11:e64626) doi:10.2196/64626

KEYWORDS

adolescent social media; social media bans; emotion regulation; youth; adolescent; media use; social platform; social network; self-regulation; behavioral health; mental health; digital health; technology; digital literacy

Introduction

Social media, defined as technologically based platforms using public, continually accessible, algorithmic-based social interactions, is an integral part of adolescents' daily lives [1]. Ninety percent of adolescents (aged 13‐19 years) in the United States report that they have a social media platform [2], with more than half using these platforms for more than 4 hours daily [3]. The significant time adolescents invest in social media has raised concerns about the impact on their mental health and well-being, prompting various local and national initiatives advocating for bans and severe restrictions on social media usage. Yet, research suggests that time on social media is not predictive of mental health outcomes and bans often fail to achieve their intended purpose, with little to no improvement in mental health or academic performance [4,5]. Evidence across various types of bans for adolescents (eg, smartphone bans, banning transgender youth from sports participation, and book bans) indicates that such restrictions carry several negative consequences for their mental health [6-9]. In this viewpoint paper, we advocate that instead of enforcing bans and severe restrictions, we focus our efforts on fostering emotion regulation in adolescents, so they develop skills specifically tailored for self-regulating their social media interactions. This redirection is crucial given the low efficacy of bans, the numerous benefits of social media use, and the necessity of developing emotion regulation skills for healthy development.

Limitations of Bans and Severe Restrictions

Social media bans and severe restrictions are a rigid, ineffective response to evolving issues that warrant continuous evaluation. Social media bans inhibit adolescent psychosocial needs by keeping adolescents from a source of meaningful connection without offering a valuable alternative. Strict social media restrictions and bans for adolescents have resulted in several negative consequences, such as instilling feelings of isolation, fostering rebellion against authority, and contributing to underdeveloped digital literacy skills [10]. Additionally, legislators have been attempting to impose time restrictions on social media use for minors [11,12] but have faced challenges in passing such regulations due to constitutional concerns. Even if implemented, these time restrictions do not address the quality of adolescents' social media interactions and other contextual factors, which are critical for evaluating their impact on mental health. The current evidence on the effectiveness of bans in improving adolescent mental health remains weak and inconclusive [13-15]. A recent scoping review found that only 6 of the 22 studies examined the effects of mobile phone bans on mental health and well-being in schools [5]. Of these, 2 included anecdotal support for banning mobile phones while 4 found no evidence to support bans for adolescent mental health and well-being [5]. Further, studies on other adolescent outcomes (eg, academic performance) were mixed, likely due to variation in the type of ban (eg, partial, full, and social media only) and the aspect of social media (eg, time spent, highly visual apps, and instant messaging–based apps) investigated, which hinders definitive conclusions [16]. Future work examining specific types of restrictions and bans is needed to elucidate their efficacy. Taken together, the limited evidence supporting bans underscores that bans are not the most effective solution for improving adolescent well-being.

In spite of the limited empirical support, social media bans and severe restrictions continue to emerge in an attempt to ameliorate adolescent mental health concerns. Recent examples of such initiatives include a Florida law banning social media use for those aged 14 years and younger, and requiring parental consent for 12-15 years old [11]. Similarly, the Los Angeles Unified School District, the second largest in the United States (with more than 429,000 students), has banned cellphone and social media use during school hours for K-12 students [17]. This trend extends beyond the United States, as France proposed a law to ban social media use for those aged 15 years and younger [18]. Adding to these measures, US Surgeon General Dr Vivek H Murthy called on Congress to implement warning labels on social media platforms, similar to those found on cigarette packages. Dr Murthy argued for swift action, citing social media as "an important contributor" to the mental health crisis among young people. This characterization oversimplifies social media use and neglects the diverse experiences adolescents can have on social media, including positive ones.

Benefits of Social Media

Social media is often regarded as solely negative for adolescents, overlooking its many benefits. For example, a thematic meta-synthesis of *qualitative* studies demonstrated that social media had a nuanced impact on adolescent well-being, revealing both positive effects (eg, promoting learning) and negative effects (eg, impacting mood through exposure to upsetting content) [19]. Social media fulfills several psychological and social needs for adolescents, such as self-expression, social validation, and peer interaction [20-22]. A survey conducted by EdWeek Research Center among 1054 high school adolescents found that one-third of adolescents report feeling less alone because of social media and 72% report that social media has either no impact or a positive impact on their mental health [23]. Adolescents value social media for its unique benefits, such as overcoming shyness, building relationships, staying connected over distances, and facilitating group interactions and shared experiences that in-person communication cannot always provide [19]. Evidence also shows that social media can provide a safe space for specific adolescent populations, such as those in the LGBTQ+ community [24]. For these marginalized groups, dedicated web-based communities serve as crucial support systems [25]. Additionally, social media can connect adolescents with varied interests, such as fandom communities where adolescents can engage with others who have shared interests in books, films, or others [26]. In a recent study of fandom communities, sexual and gender minority youth reported that web-based communities contribute to their identity development by providing a safe and anonymous space, as well as offering validation and normalization [26]. While social media may foster various types of connections, most existing naturalistic research to date has focused on negative mental health outcomes. Consequently, there is a dearth of research examining its potential positive effects.

Social media platforms can be leveraged to deliver interventions providing mental health support for adolescents. Various interventions show promise, from brief, targeted approaches [27] to ongoing peer support networks [28]. For example, single-session interventions integrated into popular social media platforms demonstrate potential in addressing acute mental health needs, such as self-harm [29]. Exposure to personal recovery narratives on social platforms can reduce suicidal ideation [30]. Furthermore, the power of peer support on these platforms appears to be mutually beneficial, as providing support to others online can lead to improvements in one's own mental health and coping strategies [31,32]. These findings emphasize the benefits of social media and underscore the need for a more nuanced approach to adolescent use. Greater emphasis should be placed on fostering positive experiences online and mitigating negative ones. Adolescents need assistance in developing the skills necessary to navigate the complex landscape of social media, including the ability to understand and manage their emotions.

Emotion Regulation in Adolescence

Adolescence is marked by significant psychosocial shifts, including the development of identity and independence away from parents and caregivers [33]. These shifts are accompanied by various emotional challenges (eg, increased social and academic pressures). Understanding and effectively managing emotions is critical for navigating social media experiences in a healthy manner. This skill set, known as emotion regulation, is broadly defined as one's ability to intentionally or unintentionally influence or manage one's emotions [34]. The process model of emotion regulation, developed by Gross, provides a framework for understanding how individuals regulate their emotions [35,36]. This model identifies 5 stages in the emotion regulation process: situation selection, situation modification, attentional deployment, cognitive change, and response modulation [35]. Each stage represents a point at which individuals can intervene to influence their emotional experiences.

Situation selection involves choosing to approach or avoid situations that are likely to elicit emotional responses [35,36]. Situation selection can precede the onset of an emotion, meaning people can elect into or out of situations that elicit emotional responses. This is unique to the other stages of emotion regulation which occur after the emotion has already been elicited. By identifying and avoiding situations (or social media content) that may bring about negative emotional experiences, one can prevent those negative emotions from arising. For example, an adolescent who feels anxious might avoid engaging with highly visual and stimulating apps (eg, TikTok) to prevent exacerbating their anxiety, thereby proactively managing their emotional state (ie, situation selection). Alternatively, the adolescent might select social media platforms that induce positive emotional experiences (eg, a mindfulness-based app or a communication platform). Depending upon the developmental stage of the adolescent, identification and situation selection could either be self-directed or could be taught and supported by parents or caregivers.

Several emotion regulation strategies can be deployed once an emotion has been triggered. *Situation modification* and *attentional deployment* are 2 strategies that leverage behavior once an emotion has been elicited, either by redirecting attention or manipulating a situation to better suit one's needs [35,36]. To do so, adolescents must develop the ability to recognize their current emotional experiences. *Situation modification* refers to altering aspects of the situation to minimize negative emotional experiences or foster positive ones. In the context of social media use, this could include recognizing negative emotional shifts when they occur while using social media (as generative content means less direct control over what is viewed) and navigating away from it (eg, leaving the app or website). For example, TikTok, a popular social media app, contains categories of content for users to engage with. Adolescents could modify their emotional experience while staying on the app by searching strategically for content that elicits the desired emotion. *Attentional deployment* includes directing one's attention toward or away from emotionally distressing scenarios and could include focusing on positive aspects, such as comments that foster connectedness or encouraging aspects of a platform.

Cognitive change refers to reappraising one's attitudes or beliefs toward a situation, which then alters the emotional value and importance [35,36]. In the context of social media, *cognitive change* could include caregiver and adolescent discussions about the role and value of social media in one's life. The examination of those beliefs can then foster the ability to challenge or change previously unexamined assumptions about social media use. This could be an opportunity to reframe or correct held beliefs on social media, especially during adolescence where reinforcement from peers is developing and particularly sensitive. For example, if an adolescent feels pressure to post about their achievements to gain validation from their peers, they may feel distressed if they have low engagement on their posts. Caregivers can help by discussing how self-worth is not tied to online validation and encouraging activities that build self-esteem.

Lastly, *response modulation* includes adjusting one's behavioral response to an emotion, either by inhibiting or amplifying it [35,36]. Given the interactive nature of social media, this modulation is particularly relevant for content creation and engagement. In this context, response modulation includes an initial emotion to content (internal) that is then either inhibited or amplified through a behavioral response (external). For example, an adolescent who feels excited after receiving a compliment on their post might amplify this positive emotion by engaging more with the content (ie, liking and commenting on other posts), sharing their excitement with friends, or posting more frequently to maintain positive interaction. Conversely, if the adolescent receives a negative comment that makes them feel upset, they might inhibit their emotional response by choosing not to respond to the comment or by taking a break from social media to calm down. By modulating their responses, users can shape their social media experience to be more positive and constructive.

The application of Gross and colleagues' process model to emotion regulation presents an opportunity for adaptive social media use behaviors more so than bans. These suggestions frame one aspect of the social media use experience. An additional aspect to consider is the influence of user-driven content on algorithm-based applications. A study found that, although adolescents might understand that their actions influence the algorithm, thus determining the content they see, they often do not adjust their social media behaviors or clearly express their understanding of how algorithms work [37]. To effectively support emotion regulation in this context, it is crucial to help adolescents bridge the gap between their awareness of algorithmic influences and practical strategies for managing their social media use.

Another consideration is the role that caregivers play in helping adolescents develop and use effective emotion regulation strategies. Caregivers can aid in managing social media use and emotion regulation even while adolescents develop autonomy and become less reliant on the caregiver's guidance throughout the adolescent years. For example, research shows that parents use social networking sites to maintain communication with their adolescents [38], highlighting how social media serves as a key avenue for fostering dialogue and connection during this developmental phase. Regulation skills can be enhanced both indirectly through modeling [39] and directly through active caregiver involvement and discussions about media content [40], which can then foster critical thinking and self-regulation in adolescents [41].

Suggested Call to Action

Taken together, social media is an integral part of modern adolescent life. We encourage the following 4 calls to action as a summary of this viewpoint:

- *Continue research into the specific and differential effects of social media platforms and algorithms*. Social media platforms and usage patterns influence adolescent mental health outcomes differently, but regulations often overlook these distinctions. More research is needed to identify how various platforms, usage patterns, and algorithms specifically impact adolescent mental health.
- *Discontinue ineffective (and potentially harmful) social media bans*. Policy makers should explore alternative approaches to address adolescent mental health concerns.
- *Recognize social media benefits for adolescents*. Acknowledge the potential positive impacts on adolescent mental health, especially for vulnerable and minoritized groups. A balanced perspective will foster opportunities for healthy usage.
- *Promote emotion regulation strategies on social media*. Encourage the development of skills that enable adolescents to navigate digital spaces and foster positive social media experiences. The process model of emotion regulation provides an organized way to apply emotion regulation specifically to social media use behaviors. Parents,

educators, and app developers should create environments and tools that help adolescents practice and strengthen their emotion regulation skills in the context of social media use.

Conclusion

Shifts in psychosocial development throughout adolescence and navigating the social media landscape present unique challenges for adolescents, all of which can significantly impact adolescents' emotional well-being. Yet, social media bans and severe restrictions that rely on external regulation offer little improvement in addressing these unique challenges. Further, social media bans and severe restrictions neglect the positive experiences that promote social connectedness and improve mental well-being among adolescents. Emotion regulation, which can be self-directed, offers a way forward to address these challenges while maintaining the autonomy and interests of adolescents. Parental or caregiver guidance helps adolescents become discerning consumers of media, better equipped to handle the emotional challenges posed by digital interactions. The process model of emotion regulation can serve as an initial guide to orient the identification and usage of emotion regulation skills within specific social media use behaviors. Taken together, emotion regulation in social media use offers a promising avenue forward that maintains adolescent agency while addressing growing concerns regarding adolescent mental health.

Conflicts of Interest

All authors are employees of Fit Minded, Inc. The views expressed in this manuscript are those of the authors and do not necessarily reflect the official position of Fit Minded, Inc.

References

1. Health advisory on social media use in adolescence. American Psychological Association. URL: https://www.apa.org/topics/social-media-internet/health-advisory-adolescent-social-media-use [accessed 2024-10-19]
2. Social media and teens. The American Academy of Child and Adolescent Psychiatry. 2018 Mar. URL: https://www.aacap.org/AACAP/Families_and_Youth/Facts_for_Families/FFF-Guide/Social-Media-and-Teens-100.aspx [accessed 2024-10-19]
3. Teens spend average of 4.8 hours on social media per day. Gallup, Inc. 2023 Oct 13. URL: https://news.gallup.com/poll/512576/teens-spend-average-hours-social-media-per-day.aspx [accessed 2024-07-19]
4. Coyne SM, Rogers AA, Zurcher JD, Stockdale L, Booth M. Does time spent using social media impact mental health? An eight year longitudinal study. Comput Human Behav 2020 Mar;104:106160. [doi: 10.1016/j.chb.2019.106160]
5. Campbell M, Edwards EJ, Pennell D, et al. Evidence for and against banning mobile phones in schools: a scoping review. J Psychol Couns Schools 2024 Sep;34(3):242-265. [doi: 10.1177/20556365241270394]
6. Abrahamsson S. Smartphone bans, student outcomes and mental health. SSRN. 2024 Feb 22. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4735240 [accessed 2024-10-19]
7. Pickering G. "Harmful to minors": how book bans hurt adolescent development. Ser Libr 2023;84(1-4):32-45. [doi: 10.1080/0361526X.2023.2245843]
8. Celano M, Hazzard A, Rivera VA. Supporting children's healthy ethnic-racial identities in the context of banned books: perspectives from mental health professionals. In: Supporting Students' Intellectual Freedom in Schools: The Right to Read: IGI Global; 2024:59-83. [doi: 10.4018/978-1-6684-9655-8.ch004]
9. Barrera E, Millington K, Kremen J. The medical implications of banning transgender youth from sport participation. JAMA Pediatr 2022 Mar 1;176(3):223-224. [doi: 10.1001/jamapediatrics.2021.4597] [Medline: 34870701]
10. Byrne J, Kardefelt-Winther D, Livingstone S, Stoilova M. Global Kids Online: research synthesis, 2015-2016. Global Kids Online. 2016. URL: http://globalkidsonline.net/synthesis-report/ [accessed 2024-10-19]
11. Florida's DeSantis signs law restricting social media for people under 16. Reuters. 2024 Mar 26. URL: https://www.reuters.com/world/us/floridas-desantis-signs-law-restricting-social-media-people-under-16-2024-03-25/ [accessed 2024-10-19]
12. Social media and children 2023 legislation. National Conference of State Legislatures. URL: https://www.ncsl.org/technology-and-communication/social-media-and-children-2023-legislation [accessed 2024-10-19]
13. Guldvik MK, Kvinnsland I. Smarter without smartphones? Effects of mobile phone bans in schools on academic performance, well-being, and bullying [Master's dissertation]. : Norwegian School of Economics; 2018. [Medline: 29462127]
14. Kessel D, Hardardottir HL, Tyrefors B. The impact of banning mobile phones in Swedish secondary schools. Econ Educ Rev 2020 Aug;77:102009. [doi: 10.1016/j.econedurev.2020.102009]
15. Campbell M, Edwards EJ, Pennell D, et al. Evidence for and against banning mobile phones in schools: a scoping review. J Psychol Couns Sch 2024;34(3):242-265. [doi: 10.1177/20556365241270394]
16. McCrory A, Best P, Maddock A. The relationship between highly visual social media and young people's mental health: a scoping review. Child Youth Serv Rev 2020 Aug;115:105053. [doi: 10.1016/j.childyouth.2020.105053]
17. Yu YJ. 2nd-largest school district votes to ban cellphones and social media for students. ABC News. 2024 Jun 18. URL: https://abcnews.go.com/GMA/Family/2nd-largest-school-district-vote-cell-phone-social/story?id=111210584 [accessed 2024-07-19]

18. Macron backs ban on use of smartphones for under 11s. The Telegraph. 2024 Jun 12. URL: https://www.telegraph.co.uk/world-news/2024/06/12/macron-backs-report-calling-for-ban-on-smartphones-for-u11s/ [accessed 2024-07-19]
19. Shankleman M, Hammond L, Jones FW. Adolescent social media use and well-being: a systematic review and thematic meta-synthesis. Adolescent Res Rev 2021 Dec;6(4):471-492. [doi: 10.1007/s40894-021-00154-5]
20. Khan S, Gagné M, Yang L, Shapka J. Exploring the relationship between adolescents' self-concept and their offline and online social worlds. Comput Human Behav 2016 Feb;55:940-945. [doi: 10.1016/j.chb.2015.09.046]
21. National Academies of Sciences, Engineering, and Medicine. Social Media and Adolescent Health: National Academies Press; 2024. [doi: 10.17226/27396]
22. Meng S, Kong F, Dong W, Zhang Y, Yu T, Jin X. Mobile social media use and life satisfaction among adolescents: a moderated mediation model. Front Public Health 2023;11:1117745. [doi: 10.3389/fpubh.2023.1117745] [Medline: 38094229]
23. Prothero A. Students think social media is fine, but teachers see a mental health minefield. Education Week. 2024 Mar 25. URL: https://www.edweek.org/leadership/students-think-social-media-is-fine-but-teachers-see-a-mental-health-minefield/2024/03 [accessed 2024-07-19]
24. Berger MN, Taba M, Marino JL, Lim MSC, Skinner SR. Social media use and health and well-being of lesbian, gay, bisexual, transgender, and queer youth: systematic review. J Med Internet Res 2022 Sep 21;24(9):e38449. [doi: 10.2196/38449] [Medline: 36129741]
25. Chaiton M, Thorburn R, Chan E, Copeland I, Luphuyong C, Feng P. Evaluating web-based care for mental health and substance use issues for lesbian, gay, bisexual, transgender, queer, questioning, and 2-spirit youths in the context of the COVID-19 pandemic: community-based participatory research study. J Med Internet Res 2023 Nov 17;25:e44292. [doi: 10.2196/44292] [Medline: 37319010]
26. McInroy LB, Craig SL. "It's like a safe haven fantasy world": online fandom communities and the identity development activities of sexual and gender minority youth. Psychol Pop Media 2020;9(2):236-246. [doi: 10.1037/ppm0000234]
27. Amsalem D. Social media–based brief video interventions to support youths' mental health. Psychiatr Serv 2024 Jul 10;10:appips20240093. [doi: 10.1176/appi.ps.20240093] [Medline: 38982837]
28. Naslund JA, Aschbrenner KA, Marsch LA, Bartels SJ. The future of mental health care: peer-to-peer support and social media. Epidemiol Psychiatr Sci 2016 Apr;25(2):113-122. [doi: 10.1017/S2045796015001067] [Medline: 26744309]
29. Dobias ML, Morris RR, Schleider JL. Single-session interventions embedded within Tumblr: acceptability, feasibility, and utility study. JMIR Form Res 2022 Jul 25;6(7):e39004. [doi: 10.2196/39004] [Medline: 35877163]
30. Franz PJ, Mou D, Kessler DT, et al. Digital bibliotherapy as a scalable intervention for suicidal thoughts: a randomized controlled trial. J Consult Clin Psychol 2022 Aug;90(8):626-637. [doi: 10.1037/ccp0000752] [Medline: 36066864]
31. Morris RR, Schueller SM, Picard RW. Efficacy of a web-based, crowdsourced peer-to-peer cognitive reappraisal platform for depression: randomized controlled trial. J Med Internet Res 2015 Mar 30;17(3):e72. [doi: 10.2196/jmir.4167] [Medline: 25835472]
32. Doré BP, Morris RR, Burr DA, Picard RW, Ochsner KN. Helping others regulate emotion predicts increased regulation of one's own emotions and decreased symptoms of depression. Pers Soc Psychol Bull 2017 May;43(5):729-739. [doi: 10.1177/0146167217695558] [Medline: 28903637]
33. Koepke S, Denissen JJA. Dynamics of identity development and separation–individuation in parent–child relationships during adolescence and emerging adulthood—a conceptual integration. Dev Rev 2012 Mar;32(1):67-88. [doi: 10.1016/j.dr.2012.01.001]
34. McRae K, Gross JJ. Emotion regulation. Emotion 2020;20(1):1-9. [doi: 10.1037/emo0000703]
35. Gross JJ. Emotion regulation: current status and future prospects. Psychol Inq 2015 Jan 2;26(1):1-26. [doi: 10.1080/1047840X.2014.940781]
36. Gross JJ. The emerging field of emotion regulation: an integrative review. Rev Gen Psychol 1998 Sep;2(3):271-299. [doi: 10.1037/1089-2680.2.3.271]
37. Swart J. Experiencing algorithms: how young people understand, feel about, and engage with algorithmic news selection on social media. Soc Media Soc 2021 Apr;7(2). [doi: 10.1177/20563051211008828]
38. Doty J, Dworkin J. Parents' of adolescents use of social networking sites. Comput Human Behav 2014 Apr;33:349-355. [doi: 10.1016/j.chb.2013.07.012]
39. Vaala SE, Bleakley A. Monitoring, mediating, and modeling: parental influence on adolescent computer and internet use in the United States. J Child Media 2015 Jan 2;9(1):40-57. [doi: 10.1080/17482798.2015.997103]
40. Gabrielli J, Marsch L, Tanski S. TECH parenting to promote effective media management. Pediatrics 2018 Jul;142(1):e20173718. [doi: 10.1542/peds.2017-3718] [Medline: 29954831]
41. Walch R, Sabey A. Parental monitoring of adolescent social media use and emotional regulation. Fam Perspect 2020;1(1):Article 12.

Edited by J Torous; submitted 22.07.24; peer-reviewed by AN Daly, M Gasmi; revised version received 18.09.24; accepted 19.09.24; published 04.11.24.

Please cite as:
McAlister KL, Beatty CC, Smith-Caswell JE, Yourell JL, Huberty JL
Social Media Use in Adolescents: Bans, Benefits, and Emotion Regulation Behaviors
JMIR Ment Health 2024;11:e64626
URL: https://mental.jmir.org/2024/1/e64626
doi:10.2196/64626

Building Mutually Beneficial Collaborations Between Digital Navigators, Mental Health Professionals, and Clients: Naturalistic Observational Case Study

Carla Gorban[1,*]; Sarah McKenna[1,*], BA, BSc, PhD; Min K Chong[1], BSc (Hons); William Capon[1], BSc, MBMSc; Robert Battisti[2], BPsych (Hons), PhD; Alison Crowley[2]; Bradley Whitwell[2]; Antonia Ottavio[3], Dip (Health Sciences), MCSL, Grad Dip (Management), Grad Dip (Child and Adolescent Mental Health); Elizabeth M Scott[1], MBBS; Ian B Hickie[1], AM, MD; Frank Iorfino[1], BSc, MBMSc, PhD

1

2

3

[*]these authors contributed equally

Corresponding Author:
Frank Iorfino, BSc, MBMSc, PhD

Abstract

Despite the efficacy of digital mental health technologies (DMHTs) in clinical trials, low uptake and poor engagement are common in real-world settings. Accordingly, digital technology experts or "digital navigators" are increasingly being used to enhance engagement and shared decision-making between health professionals and clients. However, this area is relatively underexplored and there is a lack of data from naturalistic settings. In this paper, we report observational findings from the implementation of a digital navigator in a multidisciplinary mental health clinic in Sydney, Australia. The digital navigator supported clients and health professionals to use a measurement-based DMHT (the Innowell platform) for improved multidimensional outcome assessment and to guide personalized decision-making. Observational data are reported from implementation logs, platform usage statistics, and response rates to digital navigator emails and phone calls. Ultimately, support from the digital navigator led to improved data collection and clearer communications about goals for using the DMHT to track client outcomes; however, this required strong partnerships between health professionals, the digital navigator, and clients. The digital navigator helped to facilitate the integration of DMHT into care, rather than providing a stand-alone service. Thus, collaborations between health professionals and digital navigators are mutually beneficial and empower clients to be more engaged in their own care.

(JMIR Ment Health 2024;11:e58068) doi:10.2196/58068

KEYWORDS

digital navigator; digital coach; clinical technology specialist; mental health services; shared decision-making; lived experience; implementation; poor engagement; decision-making; mental health; digital mental health; digital mental health technology

Introduction

Although the demand for care has been rising, mental health care systems are falling short, with most services reporting high wait times for care, high dropout rates, and difficulties providing access to specialized care [1,2]. Digital mental health technologies (DMHTs) have the potential to facilitate highly personalized care through multidimensional assessment and more efficient care coordination [3-5]. However, implementation studies have consistently reported low uptake in real-world settings [6-8]. The barriers appear to be multifaceted, as they can be linked to individual attitudes or beliefs toward the technology, the nature of clinical practice, as well as technology-driven or organizational barriers [9-11]. Innovative solutions are needed to ensure that DMHTs can be successfully deployed into real-world settings and improve equitable access to effective mental health treatments.

Given that the barriers are complex and multifaceted, integrating DMHTs in real-world settings requires broad transformations to occur within mental health services. This calls for new supports to be integrated in clinics that can provide better training and assistance to both health professionals and clients. One solution has been the use of health technology experts in real-world services. These experts have been given various titles including "digital coaches," "clinical technology specialists," and "digital navigators"—for simplicity, this paper will refer to them as digital navigators [12-14]. Digital navigators can provide unique value to clinical teams by doing the following: (1) evaluating apps so health professionals have confidence recommending useful apps to clients; (2) offering nonclinical technical support, such as troubleshooting between sessions to improve ease of use; and (3) helping to interpret app data before visits and highlighting salient data features to improve the usefulness and clinical value of the technology to health professionals [12]. Digital navigators can expand on more traditional positions, such as community health workers or social workers, to include assisting with the use of new clinical technologies and how to utilize them in care. Moreover, digital navigators can build rapport with clients around the use of DMHTs, which is key to improving engagement and trust in any clinical intervention [12]. Overall, this ensures that health professionals and clients are actively supported to make the best use of digital technologies in health care.

Despite the promise of this approach, the implementation of digital navigators in previous research has been mixed, and there is still limited information about protocols or training [15,16]. Previous work has utilized both clinical and nonclinical staff, used scheduled and on-demand support, and trialed both face-to-face and remote interactions, with similar results [15,16]. A particularly important question to address is whether these roles should be filled by existing members of the care team, or whether a "social coaching model" (to use a term coined by Meyer, Wisniewski, and Torous [15]), which uses peer workers to encourage client engagement and provide support, is more valuable. A review of 26 DMHT trials found only 3 studies that had used a peer worker rather than a health professional, suggesting that the potential benefits of this approach are underexplored [14].

Additionally, the focus of digital navigation in existing literature is most often to support the use of self-guided online therapy modules. Another potential role of digital navigators is to support the integration of digital technologies into traditional therapies provided by health professionals. For example, measurement-based DMHTs aim to improve the identification and tracking of mental disorders over the course of care, ensuring clients receive more appropriate treatments earlier in their illness trajectory [17-19]. These platforms require health professionals to adopt new practices and processes, which often involves proactively responding to new information provided by DMHTs. Past work has consistently found that health professionals need proper training around how to use and interact with DMHTs during sessions for them to be implemented [14]. Accordingly, further work is needed to understand how digital navigators can be integrated in health services and used to enhance existing care options, such as improved outcome monitoring.

This paper presents observational evidence from the real-world implementation of a digital navigator in a multidisciplinary clinic in Sydney, Australia. This role was implemented to support the use of a measurement-based DMHT (the Innowell platform) as part of the EMPOWERED (Educate, Measurement-based, Personalised, Openness, Work collaboratively, Engage, Recovery, Enhanced Digitally) trial, a randomized controlled trial [20]. Specifically, we will focus on the period leading up to the trial in which this role was implemented and tested to ensure feasibility and viability. Although the digital navigator role has been described in our published protocol [20], we aim to provide a more detailed overview of our real-world experiences integrating this role into existing clinical practice to identify valuable, generalizable, and practical insights that could inform future efforts to utilize this role in health services.

Setting and Background

The digital navigator was introduced to Mind Plasticity (a private multidisciplinary mental health clinic in Sydney, Australia) in September 2022 to improve uptake of the Innowell platform ahead of the EMPOWERED clinical trial. Mind Plasticity is comprised of 2 clinics servicing approximately 2600 clients, and offers a range of services including psychiatry, psychology, occupational therapy, and education support, as well as a mental health nurse to provide medication management. Innowell was already being used at the service prior to the digital navigator starting. The purpose of Innowell is to facilitate routine outcome monitoring by allowing clients to complete online assessments across a range of domains, including mental and physical health, functioning, substance and alcohol use, and suicidality or self-harm, to track progress in care and guide personalized decision-making (Figures 1-3) [4,19,21]. Previous work has shown that the implementation of DMHTs (such as Innowell) in real-world settings is often limited by an unwillingness to adapt clinical practice to integrate DMHTs [22-24]. As such, there was a need to improve the integration of Innowell within the service ahead of the trial launch. Prior to the trial starting, the digital navigator (author CG) was brought in to help the service establish the system processes and best practices around using Innowell. The goal of this "pretrial" phase was to increase uptake of the platform with service stakeholders (clients and health professionals), assist with the onboarding process to the platform, enhance client engagement with Innowell, and improve overall service integration. The digital navigator continued to play a key role in supporting use of the platform during the trial as discussed in the published protocol [20].

Figure 1. Clinician view of client's dashboard in Innowell.

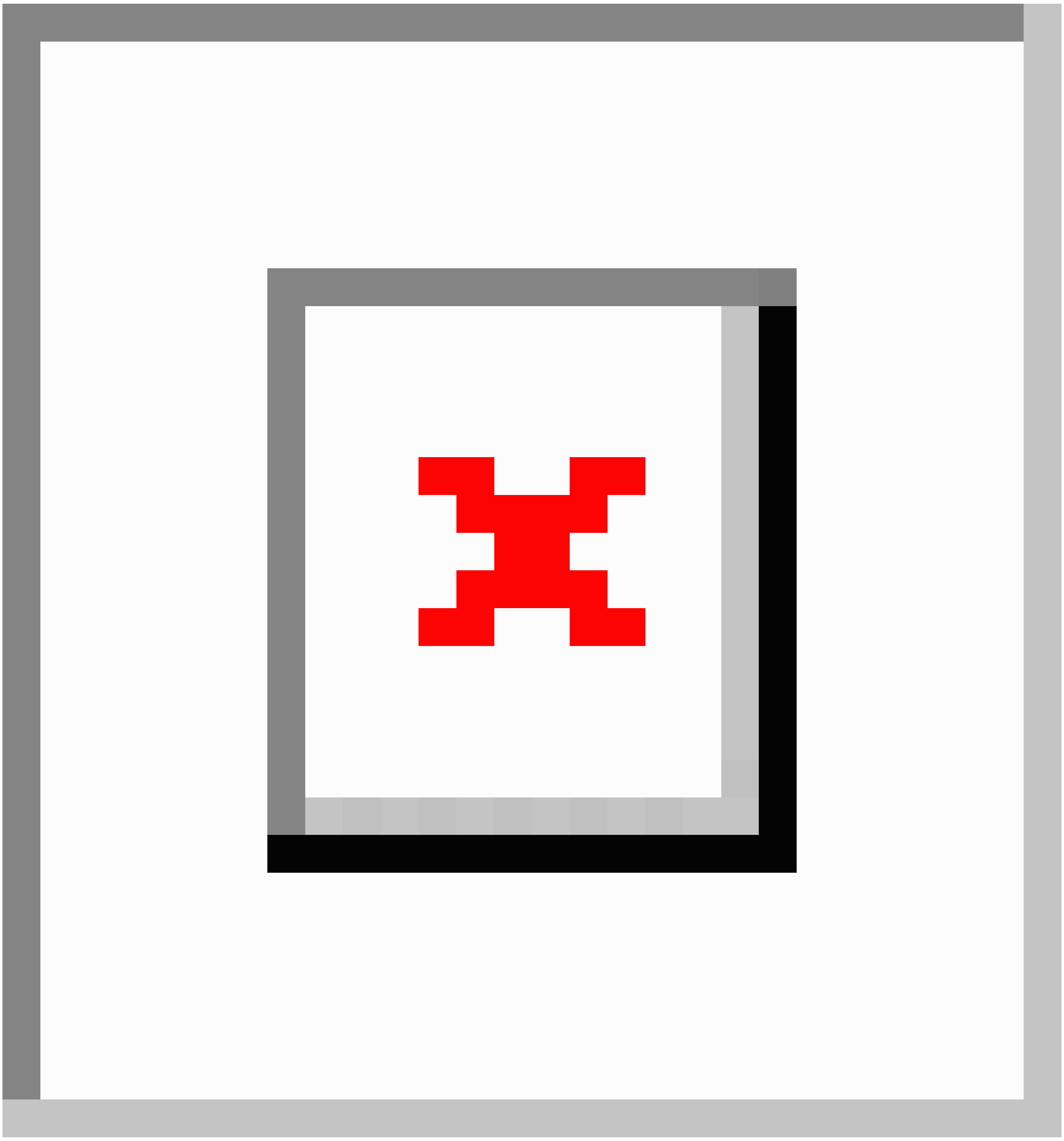


Figure 2. Client view of their own dashboard in Innowell.

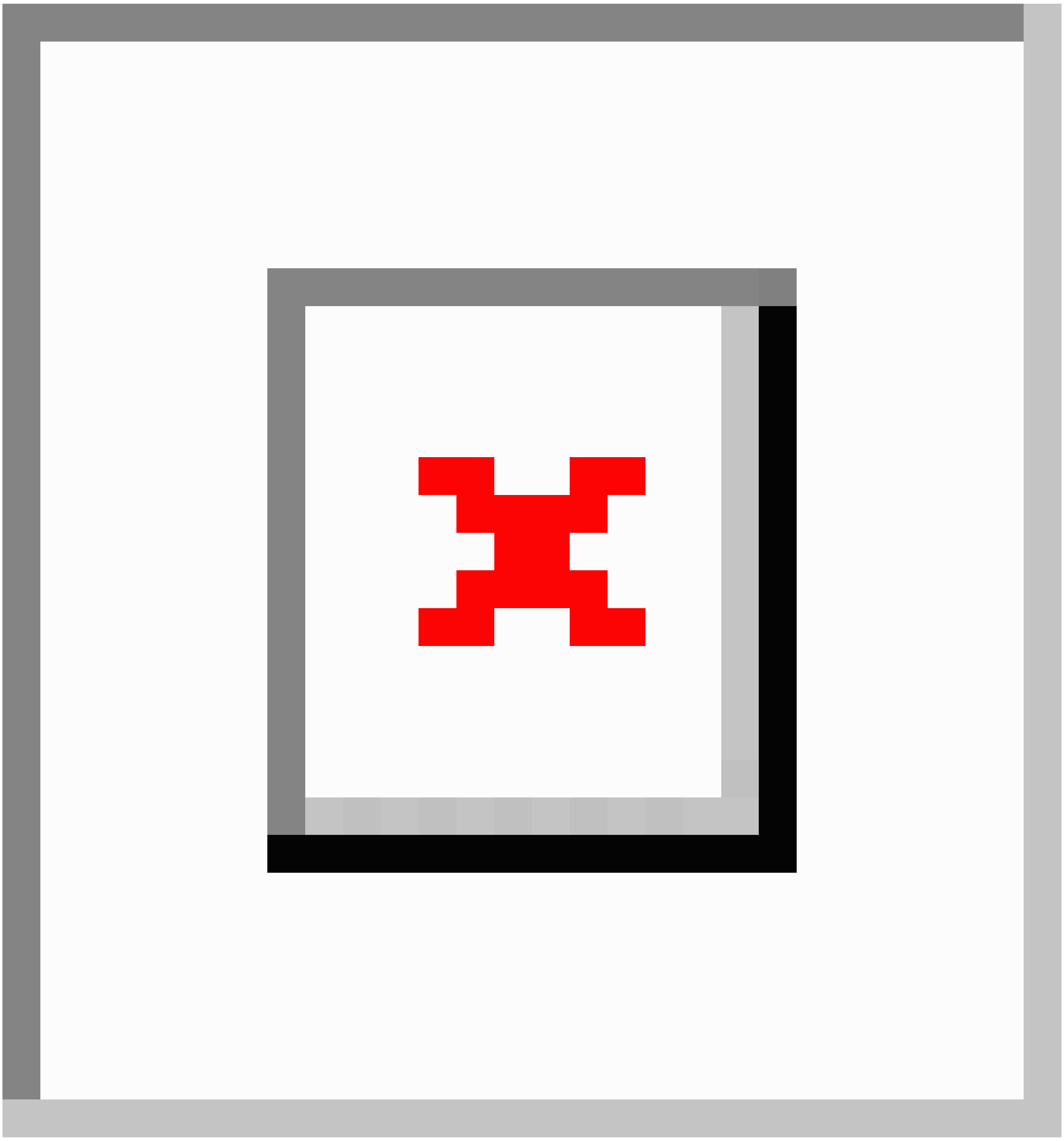
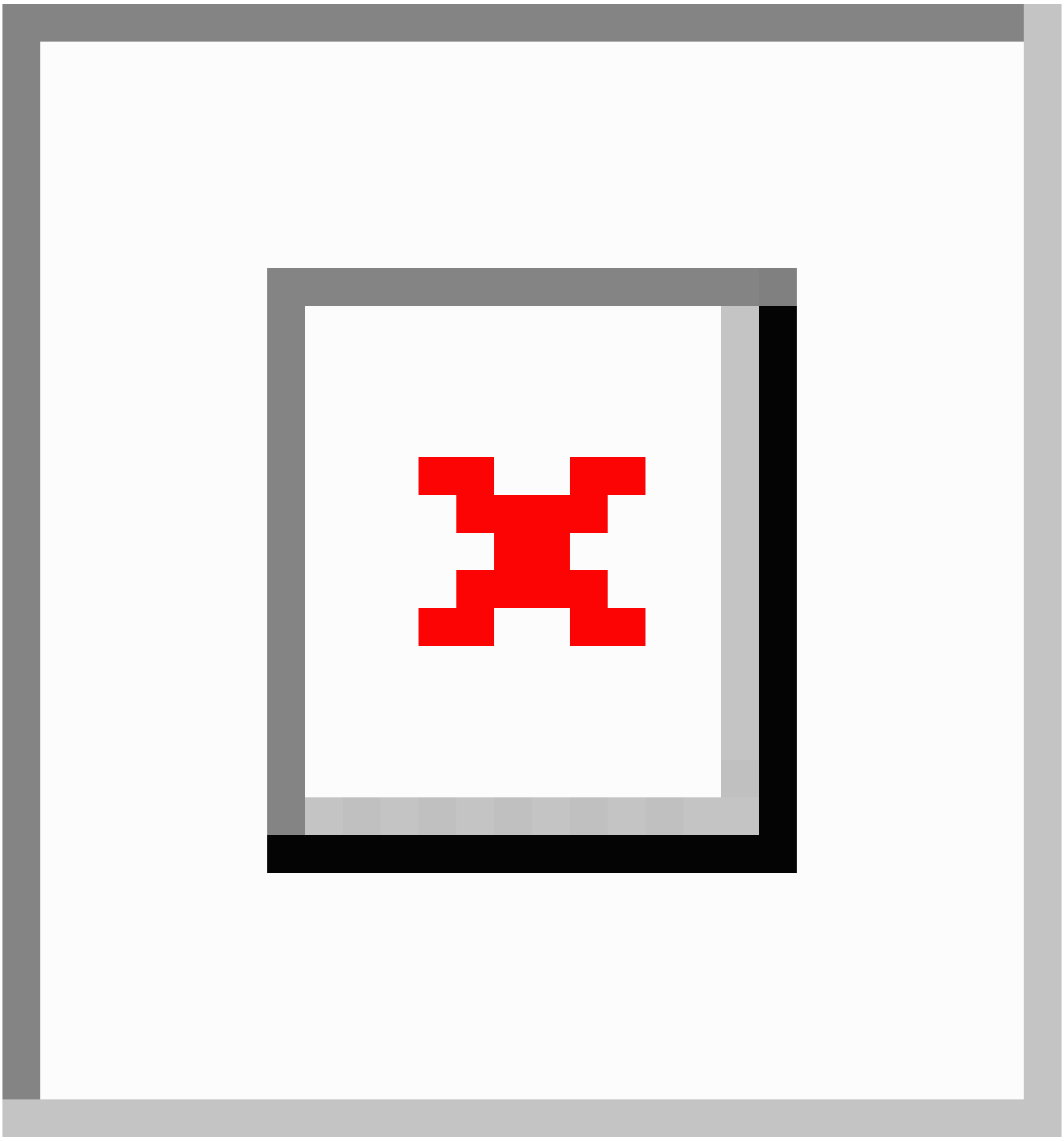


Figure 3. The longitudinal graph of the dashboard available to both clinician and client in Innowell.



Author CG was recruited to fill the role of digital navigator in the clinical trial, due to both her lived and professional expertise. Our team sought a person with lived experience of mental ill-health to fill the role of digital navigator so that this role would be more accessible and relatable to clients. Additionally, peer workers or those with lived experience are well placed in the role of digital navigators as their lived experience provides a unique perspective to walk side-by-side with a young person throughout their care-journey. CG also had professional experience from a previous implementation, which informed the current approach. This process has been outlined in detail in previous publications [4,23].

Observational Data Collection

Observational logs were used by the digital navigator to capture qualitative data about the role, including interactions with health professionals and clients (Table 1). These logs were based on feedback from face-to-face interactions, phone calls, and emails. The digital navigator also tracked client engagement with the Innowell platform and frequency of responses to calls and emails to gain further information about engagement.

Table . Observational findings from digital navigator logs on the workability and usefulness of our processes.

	What did not work?	What did?
First contact with digital navigator	• Health professionals did not have time between appointments to introduce clients to the digital navigator in person, partly because they often worked off-site. • Many clients preferred telehealth sessions and did not attend the clinic in person.	• Health professionals provided a list of clients for the digital navigator to follow up with, who had already agreed to be contacted. • One health professional made time after each session to introduce the client to the digital navigator and discuss goals for using the platform together.
Onboarding clients to the platform	• All clients were reinvited to the platform to improve service uptake, and the digital navigator sent a 2-week reminder. • Health professionals believed it was the digital navigator's role to introduce clients to the platform.	• The digital navigator sent follow-up reminders 2 days before an appointment. • Health professionals introduced the platform during the session and asked the digital navigator to provide follow-up support.
Improving ongoing use of the platform	• Clients did not generally seek ad hoc support between appointments from the digital navigator through face-to-face meetings. • Health professionals believed that clients would let them know if they wanted to discuss data from the platform in their care. • Digital navigator encouraged clients to use platform and to take initiative in asking the health professional to discuss data from the platform.	• Clients preferred to schedule communication at a time that worked for them and to choose the method (ie, emails, phone calls, Zoom, or face to face) and frequency of this communication. • Health professionals had frequent communication with the young person and digital navigator to discuss what outcomes were being tracked. • Digital navigator showed the young person the functionalities of the platform and helped them identify features that were valuable to them to assist them to get the most use out of the platform.

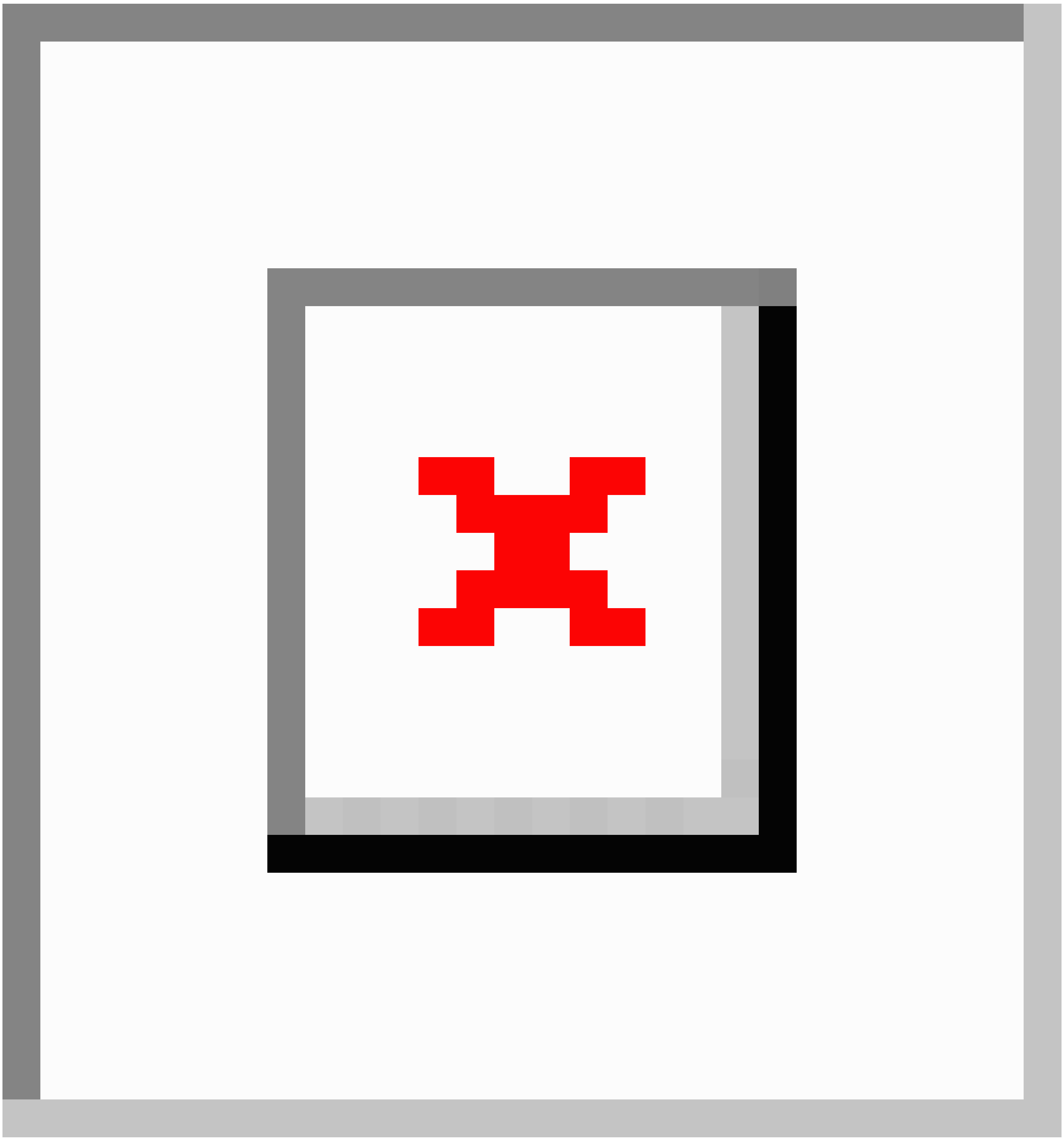
Principles of thematic analysis were used to identify common problems and solutions associated with the digital navigator role and its value in health services [25]. Preliminary analysis was conducted by the primary coder (CG), who read through the observational logs and identified common themes across the case studies. These themes were discussed with 2 secondary coders, a clinical academic researcher (colead author SM) and a senior academic researcher (author FI). CG then reviewed the case studies and identified those that reflected the main themes and that illustrated the variability in how the digital navigator role can be experienced. The case studies were then discussed with the secondary coders and the final cases and themes were agreed upon.

Principles of constructionist theory grounded our analysis, which argues that all knowledge is constructed through the experiences and subjectivities that researchers bring to the data. CG has lived experience and is a digital navigator with expertise regarding contextual factors, and SM is a clinical psychologist and is experienced in cognitive behavioral therapies that emphasize interrelationships between attitudes, experiences, and behaviors. FI is experienced in implementation of digital technologies and has knowledge of common systemic and individual-level factors that influence uptake of new technologies. These preexisting subjectivities shaped our interpretations of the data.

Description of Digital Navigator Role

Figure 4 is a detailed description of the referral process and purpose of coaching sessions. A referral-style process was used, where a health professional identified a client who would benefit from the digital navigator's support. The digital navigator then arranged unstructured "digital navigator sessions" that were tailored to the needs of the individual client. In total, there were 25 digital navigator sessions over Zoom or in-person at the clinics. The average length of time was 30 minutes per session and the digital navigator saw a range of clients, including those with attention-deficit/hyperactivity disorder and autism spectrum disorder; mood disorders such as bipolar disorder and depression; and substance misuse. The digital navigator provided feedback on the session to the health professional to ensure a constant feedback loop and encourage open communication.

Figure 4. Overview of referral process and purpose of digital navigator sessions. DN: digital navigator.



Initial meetings would be used to establish rapport with the client and introduce the Innowell platform and its functions, and provide training and guidance on how to use the platform regularly (ie, completing the summary questionnaire every 1 ‐ 2 weeks, before appointments with a health professional). During this visit, the digital navigator and the client would discuss the benefits of personalized and measurement-based care, and how Innowell can be used to support this care approach. A personalized plan would then be devised to support ongoing engagement with Innowell and its use in care. During this meeting, clients were asked how they would like to be contacted, how frequently, and for what purpose. For example, some clients preferred receiving reminder SMS text messages a couple of days before their appointment at the service, whereas other clients preferred email reminders at a set time (eg, fortnightly, on Mondays). Ongoing communication was mostly flexible and tailored to the needs of the individual. Figure 4 describes several activities that took place during these sessions.

Findings From Observational Data

Table 1 provides a summary of the aspects of the digital navigator role that were and were not successful for increasing engagement with the Innowell platform. We have summarized our experiences in more detail below.

Types of Engagement

The value of the digital navigator was not necessarily observed for increasing overall uptake of the platform but rather for improving the perceived usefulness of Innowell. Over a 2-month test period (February-April 2023), the digital navigator sent 235 emails to 204 individual clients of 4 senior health professionals, including 2 psychiatrists and 2 clinical psychologists, as part of a large-scale engagement activity. This involved a reminder to complete the onboarding process and initial Innowell questionnaire prior to their appointments at the service. However, only 18% (37/204) completed the questionnaire from this prompt by the digital navigator.

Notably, the nature and frequency of interactions between the digital navigator and clinicians seemed to influence clients' engagement with the Innowell platform and the type of contact with the digital navigator was important. Some health professionals invited the digital navigator to sessions and facilitated an introduction, whereas other clients were only contacted by email. One health professional, Clinician A, who had engaged the digital navigator as part of their care team, had a lower rate of expired links than the other clinicians, who did not have frequent contact with the digital navigator. Expired links occur when a client has not engaged with the onboarding email within 7 days. This clinician had just 5% of clients with expired links, whereas the average for other clinicians was 35% - 43%. Thus, the clients who were most likely to fill out questionnaires on the platform after contact from the digital navigator were those who had health professionals that strongly promoted using the DMHT and engaging with the digital navigator.

This is further demonstrated by case studies 1 and 2. Case study 1 (Textbox 1) describes how a client who had never before used the Innowell platform in their care had a positive experience meeting with the digital navigator and appeared to be engaged for several weeks; however, they eventually stopped using the platform. By contrast, case study 2 (Textbox 2) describes how the digital navigator supported the client to communicate their concerns to their health professional and ensured that the DMHT was used more frequently in care according to the client's wishes. In this way, digital navigators can empower clients to make shared decisions about the use of the DMHT in their care and even strengthen therapeutic relationships.

In other words, implementing digital navigators in services is not enough to improve uptake. The frequency with which clients use DMHT is most likely to depend on health professionals' attitudes and behaviors. Instead, the digital navigator can remove some of the burdens of integrating new technology in care and help both clients and health professionals to understand how the technology can be integrated effectively.

Textbox 1. Case study 1: The digital navigator may not improve long-term engagement when working in isolation.

Client details

The client and their psychologist had discussed using Innowell to help monitor mood and their symptoms and the client needed support with the onboarding process to the platform and how to use the technology. The client had never used Innowell before.

Initial session with the digital navigator

The digital navigator met with the client for roughly 1 hour via Zoom. The digital navigator assisted the client with the onboarding process (accessing secure invite links; creating and saving profiles; bookmarking the login page to their web browser, etc). The client asked questions such as how to use Innowell in sessions with their health professional and how frequently they should answer the summary questionnaire.

The digital navigator explained the purpose of the platform—how their psychologist can review their Innowell dashboard prior to appointments, making sessions more tailored to the client's unique changes and experiences based on the client's data inputs in the platform. The client and clinician can then review Innowell data together during appointments by having the client's Innowell dashboard open on the computer screen in the consult room. The digital navigator explained that answering the summary questionnaire weekly to fortnightly could be valuable as it coincided with the client's regular appointments.

At the end of the coaching session, the client said they felt confident using the technology and requested the digital navigator send text message reminders weekly to complete the summary questionnaire before attending their appointment with the psychologist. The client requested this method of contact as it would be like having "an accountability buddy," and they would reply to the digital navigator's text message once they had completed the summary questionnaire.

The client asked for the digital navigator to send the reminder on a specific day and time during the week, as this was suited to their schedule. This also allowed the client to plan by setting aside time dedicated to completing the questionnaire. The client thanked the digital navigator, commenting that they had found the coaching session to be very helpful and that they felt more confident using the Innowell platform in the future.

Summary of ongoing contact

After the coaching session, the digital navigator let the psychologist know that the client had completed the full questionnaire and informed them of the accountability plan. The client and digital navigator also scheduled a second Zoom session, for a month later.

Outcome

This plan was successful for 4 weeks, in that the client remained engaged with Innowell and completed the summary questionnaire when reminded by the digital navigator. The client was also responsive to the digital navigator's text messages and would reply as agreed.

However, the client disengaged with the digital navigator's reminders and stopped using Innowell after this 4-week period. The digital navigator kept the clinician informed and let them know when the client had disengaged. The client missed the second Zoom with the digital navigator and informed the clinician of the missed appointment.

The client reengaged with Innowell 3 months later, for 2 entries, before disengaging again. The client did not reach out directly to the digital navigator during their reengagement.

Textbox 2. Case study 2: Value of troubleshooting and improving comfortability.

Client details

The client had been using Innowell on-and-off with their psychiatrist for approximately 2 years.

Initial session with digital navigator

The client said they had been trying to use the platform as best as they could, but they were uncertain if they were using it correctly and asked the digital navigator, "how do I re-do it? I don't know what to do, so a lot of it [using Innowell] is through my own initiative." They also wanted to provide feedback regarding the functionality of the platform.

Although their psychiatrist was encouraging and incorporated the technology during some sessions, very minimal instructions were provided to them at the start on how to use the technology, and the client noted that it was "not explained to me on how to use it." The client had previously asked the administrative staff at the service for instructions but "they didn't know either."

The client had been completing the full questionnaire each time they engaged with the platform but said this was not always possible due to other commitments and they did not always have the time (as it can take 30 minutes to complete).

The digital navigator showed the client how to access the shorter summary questionnaire on the personalized home dashboard, which takes approximately 5 minutes to complete, and explained they could access more specific questionnaires by clicking on individual health cards on the dashboard.

The client provided feedback to the digital navigator on how they thought other people might be less likely to persevere and engage with DMHT, like Innowell, if it was implemented in a "wish-washy way by practices."

The client reflected that, while not always discussed or reviewed at every session, their psychiatrist's uptake of Innowell was an encouraging factor for their own use. The client also saw the value and benefits that routine outcome monitoring provided them with managing their own symptoms.

At the end of the coaching session, the client told the digital navigator, "It would have been really helpful to have you at the start to help set up the profile. There's a lot of words to understand in the full questionnaire, how does someone understand who might be suffering severely? I think it would be good to have a mentor for this. Having someone sit with you in person or online would be amazing, it's been really helpful."

Outcome

This client continues to use Innowell every 1-2 weeks.

Mutually Beneficial Relationships Between the Health Professionals, Digital Navigator, and Client

Similarly, the benefits of the DMHT for measurement-based and personalized care are dependent on the strength of collaboration between health professionals, digital navigators, and clients. When all 3 parties worked together to create shared goals for using the DMHT and to troubleshoot barriers, the perceived benefits of the platform were greater.

In one instance described in case study 3 (Textbox 3), both the health professional and the client reported that they had previously stopped discussing the platform during sessions because they believed that the other party was not interested in the data, without addressing why this was the case. They were not necessarily avoiding discussing the technology but did not see it as a priority during sessions when other problems needed to be discussed.

In this case, the digital navigator was able to empower the client to raise their concerns with their health professional and was able to provide simple solutions that helped to integrate the DMHT into care, ensuring that the client was using Innowell appropriately. It is important that health professionals and clients make shared decisions about how and why DMHTs are going to be used in care. Digital navigators can enhance this process by explaining the intended purpose, improving trust, and troubleshooting common barriers. This case study shows that the digital navigator can effectively empower clients to have specific discussions with health professionals about how they would like to use DMHTs.

Clients use Innowell for outcome measurement, yet this requires health professionals to review and respond to the data provided by clients. As shown in case study 4 (Textbox 4), some clients regularly completed platform data because they had been asked to by health professionals at the beginning of treatment, yet became frustrated when the data were not discussed during sessions. This suggests that health services should not simply introduce DMHT without providing adequate support to fully integrate them into care review processes. The time and effort for clients to engage with the digital navigator and platform may actually be unrealized if health professionals are not utilizing the information provided to guide personalized decision-making.

Taken together, our experiences suggest that the best way to implement a digital navigator is to make sure that they are seen as part of the care team and are helping to improve traditional mental health treatments alongside health professionals. Health professionals and clients will need assistance to build trust and comfortability with using new digital tools. Thus, the biggest improvements to mental health treatments are likely to be seen when health professionals, digital navigators, and clients are working together to integrate DMHT into care processes, and making joint decisions about the goals for using this technology and how barriers will be overcome.

Textbox 3. Case study 3: The digital navigator can strengthen shared decision-making between health professionals and clients.

Initial session with the digital navigator

The client met with the digital navigator in person. They were familiar with the purpose of Innowell (routine outcome monitoring and self-tracking) but found it difficult to stay motivated due to their clinician's lack of engagement with reviewing Innowell, despite the clinician's encouragement to use Innowell. The client also explained that they found it burdensome to complete the full questionnaire.

The client said they had not been shown how to use the platform in an ongoing capacity and explained that they had lost interest in dedicating the time to complete Innowell as their health professional was not reviewing or discussing it during their appointments, saying, "it was difficult to find half an hour to dedicate to answering Innowell and [health professional] weren't looking at it anyway, it's like what's the point then?" The client explained they felt disheartened by this as they had been putting in effort to complete the full questionnaire prior to appointments, and they said they were frustrated that "[Clinician] asks me the same questions as Innowell does, if they looked at it beforehand, that would be better for me. It would be a better use of my appointments too, because [clinician] would already know how I am, or if they reviewed it with me at least."

The digital navigator explained and showed the client that they can complete the shorter summary questionnaire rather than the full questionnaire each time.

Outcome

The client and digital navigator developed a plan for reminding the client to complete the summary questionnaire 2 days prior to their appointments and informing the clinician to review Innowell before and during this client's appointments.

In the 8 months since this meeting, the client has maintained consistent use of the platform, with at least 16 entries.

Textbox 4. Case study 4: Clients prefer the digital navigator to work in partnership with health professionals, not separately.

Client details

Client had previously met with the digital navigator for a coaching and onboarding session. After some time had passed, the client asked the digital navigator for online resources, specifically to help with studying, reduce procrastination, and improve organizational skills and habits.

Session with digital navigator

The client and digital navigator met over Zoom for approximately 1 hour. The client was engaged with their Innowell and had completed the summary questionnaires regularly (fortnightly to monthly) to coincide with appointments.

However, the client explained that their clinician was not engaged with Innowell and it was not discussed during sessions.

After speaking with the digital navigator, the client commented that, while useful, Innowell requires support from both the clinician and a digital navigator to optimize the client's care and use of the platform.

"Because I'm talking to you, it's very helpful. But Innowell alone isn't very helpful because it's done by itself."

The digital navigator discussed and suggested some digital and online resources with the client that could assist with their goals (as mentioned in the "Client details" section). At the conclusion of the Zoom call, the client said they found meeting with the digital navigator to discuss their Innowell and find other digital online resources to be very helpful, as was the digital navigator's encouragement of their continued engagement with Innowell.

Summary of ongoing contact

The digital navigator provided feedback to the service about the client's use of Innowell and their request for more support with studying and groups.

Outcome

The client continues to engage with Innowell, although it is unclear whether their clinician has begun incorporating Innowell into the client's appointments. The client has completed the summary questionnaire 4 times over a 4-month period since the session with the digital navigator.

The client continues to seek additional support from the service for studying and groups. The client continues to use Innowell in the hope that their clinician will start to engage with the platform to provide the more personalized support they are seeking.

Providing Virtual Support to Increase Feasibility and Accessibility of the Digital Navigator Role

Our experiences suggest that having a digital navigator "on-site" does not necessarily improve the value of this role, and that having a remote presence may help to improve both feasibility and acceptability of the role. Initially, the service requested that the digital navigator be available in person in a central area, such as the waiting room. This had a number of purposes including: (1) to prompt clients to complete questionnaires on the platform while waiting for their appointment to start; (2) to provide a visual reminder for health professionals and clients to use Innowell; and (3) to provide more opportunities for organic and ad hoc interactions between the digital navigator and health professionals (service staff reported that they did not have time for formal scheduled meetings with the digital navigator).

However, in practice, this did not produce the intended benefits. Health professionals were frequently off-site or unable to leave their consultation rooms, meaning they did not have time for ad hoc interactions with the digital navigator. In most cases, the health professional did not walk their client back out to reception, limiting opportunities for joint face-to-face interactions with the digital navigator. Moreover, some health professionals worked off-site through telehealth, or worked predominantly in an outreach setting. Similarly, in the wake of the COVID-19 pandemic, patients were more likely to attend sessions via telehealth (ie, phone or Zoom), meaning the benefits of face-to-face interactions could not be realized. Ultimately, we found that referrals were most likely to occur via health professionals providing the digital navigator with a list of names to follow-up with via phone or Zoom. This also allowed the digital navigator to contact clients at a time that was convenient to the client. Thus, given the busy nature of health clinics and the time commitment that clients are already making to attend appointments, there are clear advantages to providing remote digital navigation support.

Discussion

Digital mental health technologies have significant potential to improve personalized and measurement-based care in mental health services, but they have struggled to realize this potential due to persistent challenges of implementing novel tools in real-world settings [6-8]. This paper explored the value of implementing a new role in services, a "digital navigator," to improve the uptake and quality use of a novel multidisciplinary assessment platform, Innowell.

> *Digital mental health tools give clients a voice in their care. Digital Navigators ensure that voice is heard. However, a coordinated relationship between the health professional, client, and Digital Navigator has shown the clearest indicators of success in engagement and uptake of these tools.* [CG, digital navigator]

Overall, we found that the digital navigator was particularly useful when they were introduced to clients by health professionals as a member of the care team and assisted with troubleshooting problems and demonstrating the functionalities of the platform. By contrast, when health professionals had not embraced Innowell into their day-to-day practices, had limited interactions with the digital navigator, and did not regularly discuss Innowell data with clients during care, clients were unlikely to continue engaging with Innowell or the new support role. Therefore, digital navigators should not provide a stand-alone service to clients or health professionals but should instead work closely with both parties to enhance existing mental health treatments. Our observations suggest that, when this occurs, the quality of the care experience increases through greater transparency and shared decision-making.

This case study contributes to existing literature by providing practical recommendations from real-world experiences. There are limited guidelines or protocols that inform how digital navigator roles should be implemented in real-world settings [12,15,16,24]. Previous research has predominantly focused on using digital coaching to support self-guided online modules, and only 3 previous studies have explored the potential value of using peer workers to support digital transformation of services [15]. By contrast, we focused on implementing a measurement-based DMHT that is focused on tracking multidimensional client outcomes and assisting with more personalized and efficient treatment planning [26]. In addition, a peer worker added significant value to the role because they were seen as partners and advocates by clients. This enabled the digital navigator to identify how each client was most likely to benefit from the DMHT and to make shared decisions with them and their health professionals about how the DMHT could add value.

Our experiences also have some broader implications for the implementation of new DMHTs in mental health services. Successful implementation of DMHTs is reliant on health services creating optimal environments for digital transformation to occur. Noel and colleagues [14] theorized that digital navigators would help "to enhance the therapeutic bond between the client and the clinician" by working with both parties to identify recovery goals that will be the focus of monitoring, and by ensuring that all interactions are focused on these goals; this would ultimately improve clients' "sense of control over their mental health." Consistently, our experiences demonstrate that the relationship between the health professional, client, and digital navigator is mutually beneficial. When health professionals integrated Innowell into care and made shared decisions about the goals for using the platform with the client and the digital navigator, clients reported having more positive experiences and more frequent engagement with the platform. Thus, while these roles have predominantly been implemented to assist with recommending self-help apps used outside of therapy, health professionals and clients alike may receive more value from DMHT when they are used to enhance existing care.

Despite these contributions, our paper also had several limitations. Importantly, our findings were based on observational data recorded by the digital navigator themselves. As such, they are likely to be biased by the digital navigator's own experiences, attitudes, and reflections. Thus, we have presented our experiences as an observational case study because of the need to improve transparency and guidelines about how digital navigator roles are being implemented. However, future research should focus on more detailed evaluations that will serve as a more robust exploration of the value and effectiveness of this role [20]. Relatedly, the digital navigator only worked with a small number of health professionals and clients at the health service. Future work should focus on quantifying in more detail the time investment and costs associated with this role to better understand the utility and scalability in health services. There is also scope for future research into how the role could additionally be used to facilitate and support fidelity to a range of intervention approaches (eg, cognitive behavioral therapy). Communication with the digital navigator was voluntary; thus, those who did interact with them were most likely to view this role favourably. As such, longitudinal large cohort research in real-world settings is needed to fully understand the benefits of this role.

Taken together, our experiences suggest that when a digital navigator works closely alongside health professionals and clients to create goals for new DMHTs, provide encouragement to engage with the tool, build on users' skills to increase accessibility and enhance confidence, and provide technical troubleshooting, clients better understand the purpose of the technology, are more likely to complete ongoing assessments, and are more likely to see value from using the DMHT in their care. Thus, digital navigators have an important place in the digital transformation of health services but should not be seen as a stand-alone role.

Acknowledgments

The funding sources of this study have had no input into the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. This work was supported by the Medical Research Future Fund Applied Artificial Intelligence in Health Care grant (MRFAI000097). CG was supported by The Ainsworth 4 Foundation. SM was supported by the Cottle Family Fellowship in Youth Mental Health. MC was supported by the Australian Government Research Training Program Scholarship. WC was supported by the Australian Government Research Training Program Scholarship. IBH was supported by a National Health and Medical Research Council (NHMRC) L3 Investigator Grant (GNT2016346). FI was supported by the Bill and Patricia Richie Foundation, and an NHMRC EL1 Investigator Grant (GNT2018157). We thank all the young people who participated in this study and all the staff in the Youth Mental Health Team at the University of Sydney's Brain and Mind Centre, past and present, who contributed to this work.

Authors' Contributions

CG, SM, and FI were involved in conceptualization, design, analysis, interpretation, and writing. CG was involved in data collection. MC, WC, YS, EMS, and IBH assisted with conceptualization, review and editing, supervision, and funding acquisition. All authors (CG, SM, MC, WC, RB, AC, BW, AO, EMS, IBH, and FI) substantively edited and revised the manuscript, and approved the final version of the manuscript.

Conflicts of Interest

IBH is a Professor of Psychiatry and the Co-Director of Health and Policy, Brain and Mind Centre, University of Sydney. He has led major public health and health service development in Australia, particularly focusing on early intervention for young people with depression, suicidal thoughts and behaviors, and complex mood disorders. He is active in the development through co-design, implementation, and continuous evaluation of new health information and personal monitoring technologies to drive highly personalized and measurement-based care. He holds a 3.2% equity share in Innowell Pty Ltd, which is focused on digital transformation of mental health services. EMS is Principal Research Fellow at the Brain and Mind Centre, The University of Sydney. She is Discipline Leader of Adult Mental Health, School of Medicine, University of Notre Dame, and a Consultant Psychiatrist. She was the Medical Director, Young Adult Mental Health Unit, St Vincent's Hospital Darlinghurst until January 2021. She has received honoraria for educational seminars related to the clinical management of depressive disorders supported by Servier, Janssen, and Eli-Lilly pharmaceuticals. She has participated in a national advisory board for the antidepressant compound Pristiq, manufactured by Pfizer. She was the National Coordinator of an antidepressant trial sponsored by Servier.

References

1. Department of Health, KMPG. Evaluation of the National headspace Program. : Department of Health; 2022 URL: https://www.health.gov.au/resources/publications/evaluation-of-the-national-headspace-program [accessed 2024-10-25]
2. Royal Commission into Victoria's Mental Health System. Final Report - Volume 1 - A new approach to mental health and wellbeing in Victoria. : Victorian Government; 2021 URL: https://www.vic.gov.au/royal-commission-victorias-mental-health-system-final-report [accessed 2024-10-22]
3. Capon W, Hickie IB, McKenna S. Characterising variability in youth mental health service populations: a detailed and scalable approach using digital technology. Australas Psychiatry 2023 Jun;31(3):295-301. [doi: 10.1177/10398562231167681] [Medline: 37035873]
4. LaMonica HM, Iorfino F, Lee GY, et al. Informing the future of integrated digital and clinical mental health care: synthesis of the outcomes from Project Synergy. JMIR Ment Health 2022 Mar 9;9(3):e33060. [doi: 10.2196/33060] [Medline: 34974414]
5. Hickie IB. Implementing 21st century "end-to-end" and technology-enhanced care for young people. World Psychiatry 2022 Feb;21(1):79-81. [doi: 10.1002/wps.20928] [Medline: 35015351]
6. Mohr DC, Lyon AR, Lattie EG, Reddy M, Schueller SM. Accelerating digital mental health research from early design and creation to successful implementation and sustainment. J Med Internet Res 2017 May 10;19(5):e153. [doi: 10.2196/jmir.7725] [Medline: 28490417]
7. Borghouts J, Eikey E, Mark G, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387. [doi: 10.2196/24387] [Medline: 33759801]
8. de Jong K, Conijn JM, Gallagher RAV, Reshetnikova AS, Heij M, Lutz MC. Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: a multilevel meta-analysis. Clin Psychol Rev 2021 Apr;85:102002. [doi: 10.1016/j.cpr.2021.102002] [Medline: 33721605]

9. Carpenter-Song E, Acquilano SC, Noel V, Al-Abdulmunem M, Torous J, Drake RE. Individualized intervention to support mental health recovery through implementation of digital tools into clinical care: feasibility study. Commun Ment Health J 2022 Jan;58(1):99-110. [doi: 10.1007/s10597-021-00798-6] [Medline: 33611684]
10. Torous J, Powell AC, Rodriguez-Villa E. Health information technology resources to support measurement-based care. Child Adolesc Psychiatr Clin N Am 2020 Oct;29(4):763-773. [doi: 10.1016/j.chc.2020.06.011] [Medline: 32891375]
11. Torous J, Jän Myrick K, Rauseo-Ricupero N, Firth J. Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow. JMIR Ment Health 2020 Mar 26;7(3):e18848. [doi: 10.2196/18848] [Medline: 32213476]
12. Wisniewski H, Torous J. Digital navigators to implement smartphone and digital tools in care. Acta Psychiatr Scand 2020 Apr;141(4):350-355. [doi: 10.1111/acps.13149] [Medline: 31930477]
13. Ben-Zeev D, Drake R, Marsch L. Clinical technology specialists. BMJ 2015 Feb 19;350:h945. [doi: 10.1136/bmj.h945] [Medline: 25697164]
14. Noel VA, Carpenter-Song E, Acquilano SC, Torous J, Drake RE. The technology specialist: a 21st century support role in clinical care. NPJ Digit Med 2019;2(1):61. [doi: 10.1038/s41746-019-0137-6] [Medline: 31388565]
15. Meyer A, Wisniewski H, Torous J. Coaching to support mental health apps: exploratory narrative review. JMIR Hum Factors 2022 Mar 8;9(1):e28301. [doi: 10.2196/28301] [Medline: 35258468]
16. Gershkowitz BD, Hillert CJ, Crotty BH. Digital coaching strategies to facilitate behavioral change in type 2 diabetes: a systematic review. J Clin Endocrinol Metab 2021 Mar 25;106(4):e1513-e1520. [doi: 10.1210/clinem/dgaa850] [Medline: 33206975]
17. Gelkopf M, Mazor Y, Roe D. A systematic review of patient-reported outcome measurement (PROM) and provider assessment in mental health: goals, implementation, setting, measurement characteristics and barriers. Int J Qual Health Care 2021 Mar 5;34(Suppl 1):ii13-ii27. [doi: 10.1093/intqhc/mzz133] [Medline: 32159763]
18. Mackrill T, Sørensen KM. Implementing routine outcome measurement in psychosocial interventions – a systematic review. Eur J Soc Work 2020 Sep 2;23(5):790-808. [doi: 10.1080/13691457.2019.1602029]
19. Iorfino F, Cross SP, Davenport T, et al. A digital platform designed for youth mental health services to deliver personalized and measurement-based care. Front Psychiatry 2019;10:595. [doi: 10.3389/fpsyt.2019.00595] [Medline: 31507465]
20. Hickie IB, Iorfino F, Rohleder C, et al. EMPOWERED trial: protocol for a randomised control trial of digitally supported, highly personalised and measurement-based care to improve functional outcomes in young people with mood disorders. BMJ Open 2023 Oct 11;13(10):e072082. [doi: 10.1136/bmjopen-2023-072082] [Medline: 37821139]
21. Davenport TA, LaMonica HM, Whittle L, et al. Validation of the InnoWell platform: protocol for a clinical trial. JMIR Res Protoc 2019 May 31;8(5):e13955. [doi: 10.2196/13955] [Medline: 31152524]
22. McKenna S, Piper S, Capon W, et al. The polarization of clinician and service staff perspectives after the use of health information technology in youth mental health services: implementation and evaluation study. JMIR Hum Factors 2023 Jul 25;10:e42993. [doi: 10.2196/42993] [Medline: 37490321]
23. LaMonica HM, Milton A, Braunstein K, et al. Technology-enabled solutions for Australian mental health services reform: impact evaluation. JMIR Form Res 2020 Nov 19;4(11):e18759. [doi: 10.2196/18759] [Medline: 33211024]
24. Rodriguez-Villa E, Rauseo-Ricupero N, Camacho E, Wisniewski H, Keshavan M, Torous J. The digital clinic: implementing technology and augmenting care for mental health. Gen Hosp Psychiatry 2020;66:59-66. [doi: 10.1016/j.genhosppsych.2020.06.009] [Medline: 32688094]
25. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
26. Hickie IB, Scott EM, Cross SP, et al. Right care, first time: a highly personalised and measurement - based care model to manage youth mental health. Med J Aust 2019 Nov;211(S9) [FREE Full text] [doi: 10.5694/mja2.50383]

Abbreviations

DMHT: digital mental health tools
EMPOWERED: Educate, Measurement-based, Personalised, Openness, Work collaboratively, Engage, Recovery, Enhanced Digitally

Edited by J Torous; submitted 05.03.24; peer-reviewed by NH Zainal, T Kuo; revised version received 17.07.24; accepted 02.08.24; published 06.11.24.

Please cite as:
Gorban C, McKenna S, Chong MK, Capon W, Battisti R, Crowley A, Whitwell B, Ottavio A, Scott EM, Hickie IB, Iorfino F
Building Mutually Beneficial Collaborations Between Digital Navigators, Mental Health Professionals, and Clients: Naturalistic Observational Case Study
JMIR Ment Health 2024;11:e58068
URL: https://mental.jmir.org/2024/1/e58068
doi:10.2196/58068

The Urgent Need for an Evidence-Based Digital Mental Health Practice Model of Care for Youth

Brad Ridout[1], MAPS, PhD; Rowena Forsyth[1], PhD; Krestina L Amon[1], PhD; Pablo Navarro[2], MHlthSc, MAppPsych, MClinPsych, PhD; Andrew J Campbell[1], MAPS, PhD

1

2

Corresponding Author:
Brad Ridout, MAPS, PhD

Abstract

Australian providers of mental health services and support for young people include private and public allied health providers, government initiatives (eg, headspace), nongovernment organizations (eg, Kids Helpline), general practitioners (GPs), and the hospital system. Over 20 years of research has established that many young people prefer to seek mental health support online; however, clear client pathways within and between online and offline mental health services are currently lacking. The authors propose a Digital Mental Health Practice model of care for youth to assist with digital mental health service mapping. The proposed model offers accessible pathways for a client to engage with digital mental health services, provides clear navigation to access support for individual needs, and facilitates a seamless connection with offline mental health services using a transferable electronic health records system. This future-looking model also includes emerging technologies, such as artificial intelligence and the metaverse, which must be accounted for as potential tools to be leveraged for digital therapies and support systems. The urgent need for a user-centered Digital Mental Health Practice model of care for youth in Australia is discussed, highlighting the shortcomings of traditional and existing online triage models evident during the COVID-19 pandemic, and the complex challenges that must be overcome, such as the integration of diverse mental health care providers and establishment of a robust electronic health records system. Potential benefits of such a model include reduced pressure on emergency rooms, improved identification of immediate needs, enhanced referral practices, and the establishment of a cost-efficient national digital mental health care model with global applicability. The authors conclude by stressing the consequences of inaction, warning that delays may lead to more complex challenges as new technologies emerge and exacerbate the long-term negative consequences of poor mental health management on the economic and biopsychosocial well-being of young Australians.

(JMIR Ment Health 2024;11:e48441) doi:10.2196/48441

KEYWORDS

mental health; internet; digital health; telecounselling; social networking; telehealth; telemedicine; counseling; counselling; digital health; service; services; healthcare delivery; youth; model

Introduction

The COVID-19 pandemic highlighted what was already well-established about Australia's mental health system: that demand and ease of access were not being successfully met [1-5], especially for individuals younger than 25 years, who are the most at-risk age group [6,7]. Over the past 20 years, the Australian youth mental health system has slowly transitioned from being fragmented and piecemeal, with poor integration between inpatient and outpatient services, to include online and offline offerings nationally [8]. However, these offerings are not always clearly connected for clients to navigate [1,7] and are often difficult to access in regional and remote areas (especially for young indigenous and marginalized populations of Australia), where online and offline services are seldom integrated or inclusive [9].

Young people are increasingly using the web as the first step in seeking support and information regarding their mental health [10,11]; yet, the digital services currently on offer often poorly integrate with existing offline mental health and hospital systems of care [1,12,13]. There is also a lack of sharing of clients' mental health records between private and nongovernment organization (NGO) providers who offer online support (eg, ReachOut, BeyondBlue, Kids Helpline, and SANE), telephone centric services (eg, Lifeline and Kids Helpline), and traditional face-to-face clinical support (eg, headspace). This is due to several ethical and practical challenges associated with sharing electronic health records (EHRs) in mental health care [14], including issues of interoperability, missing data, privacy and confidentiality concerns, and legal compliance.

Currently, EHRs of young people are typically not part of digital mental health service offerings, with data often only collected upon entrance to determine the level of care required (ie, the triage stage). This means that when young people later return to online services at differing stages of need, or via a different modality, they often must “start again” with programs or offerings that may no longer suit their needs, rather than being directed to more suitable supports that reflect their prior engagement and the current stage of their mental health journey. Furthermore, the current lack of transferable EHRs between service providers implies that young people often have to retell their story multiple times. This can lead to incomplete patient histories, duplicate assessments, fragmented care, and safety risks associated with delayed or inappropriate treatment [15].

We propose a Digital Mental Health Practice model of care for youth (DMHP-Y) to aid digital mental health service mapping. It provides clear and accessible pathways for a young person to enter a digital mental health system, and then accurately navigate it to access support for their individual needs, beyond initial triage. This model proposes transferable EHRs regarding mental health service use to inform clinicians of historical online and offline service access, including any digital mental health assessments and treatments previously engaged in by a client. This will provide clinicians with essential background information on a patient’s mental health journey to date, assist in decision-making about future assessments and treatments, and facilitate continuity of care.

A Brief Overview of Current Digital Mental Health Practices in Australia

Established in 2006, headspace is an NGO funded by the Australian Government as part of its national approach to service the mental health needs of young people (younger than 25 years) [16]. While headspace has been widely lauded for providing no-gap Medicare access to psychologists under a general practitioner (GP)–referred mental health care plan, the service has been under strain, given its high client demand nationwide [17,18] and the limitations to ongoing mental health management beyond the “Better Access” to mental health Medicare system [19-21]. “eheadspace,” the digital provision of mental health support, was established to provide online resources for young people to connect with mental health information and support through forums, digital tools, and one-to-one counseling via web chat and telephone. While not an emergency service, eheadspace duplicated some of the resources offered by other NGOs’ online services (eg, ReachOut and Kids Helpline), which have also struggled in connecting online resource offerings to offline clinician services (often because clients may choose to contact these services anonymously). Furthermore, the lack of integration between eheadspace and face-to-face headspace services demonstrates the challenges involved in integrating online and offline services, even within the same service environment. Herein lies a major systemic issue: the lack of secure, transferable, digital mental health recordkeeping within and between services [22-24].

Notwithstanding this challenge, new online platforms have emerged in recent years that leverage health ITs (HITs) to integrate digital mental health tools with existing face-to-face mental health services. The “Innowell” platform [25] is a customizable digital dashboard for mental health services, designed for assessing, monitoring, and managing the mental health of their clients. It collects and reports personal information to clients and their health professionals to promote collaborative and measurement-based care, and offers a range of online clinical content and assessment tools that are made available to clients as determined by their face-to-face mental health service. Another Australian example is Orygen’s “Moderated Online Social Therapy” (MOST) platform [26,27], a free online digital mental health service that young people can connect to following a referral from participating offline services. The MOST model of care is designed to support young people with, and in between, their face-to-face sessions with a clinician, or while they are waiting for care. It offers a range of on-demand digital services, including one-on-one support, self-directed online therapy, and moderated peer-to-peer online social networking.

These 2 platforms reflect emerging models of digital mental health care internationally, which also use HIT to augment face-to-face sessions and facilitate measurement-based care. A notable example from the United States is the “Digital Clinic” [28,29], which uses smartphone apps to augment and extend care at the Beth Israel Deaconess Medical Center in Boston, Massachusetts. This evolving hybrid model of care emphasizes therapeutic alliance, measurement-based care, and shared decision-making between clients and clinicians. A key feature of the model is mindLAMP (Learn, Assess, Manage, Prevent), an open-source app that provides customizable education, assessment, and management tools, and facilitates data sharing with patients and clinician support. mindLAMP can also be used to capture metadata regarding clients’ use of the app, as well as health and physical activity patterns via a smartwatch (eg, step count and heart rate) to enable “digital phenotyping.” Digital phenotyping is a new and evolving multidisciplinary field of science that uses data from smart devices to create a holistic digital picture of behavior and health, and has the potential to augment clinical care at the individual level (eg, by identifying clinically significant behavior changes) [30].

By integrating HIT with existing mental health services, the models described above represent the beginning of a new era of digital mental health service delivery. Something these models have in common is the requirement for clients to first connect with a face-to-face mental health service to access online services. However, young people are often reluctant to seek professional face-to-face help and face barriers to treatment such as cost, stigma, confidentiality concerns, poor mental health literacy, and inaccessibility to or lack of knowledge of resources [31,32]. These factors contribute to an increasing trend for young people to use the web as the first step in seeking support and information regarding their mental health [10,11], usually by typing a text-based query into an internet search engine [12]. Discussions of Rickwood et al’s [33] help seeking model in the context of the online environment emphasize the importance of internet-based resources and interventions in addressing the mental health concerns of young people [34] and ensuring that young people are guided to appropriate services [35]. Reputable digital mental health services that do not require prior face-to-face referrals, such as ReachOut and Kids Helpline, therefore play an important role in early intervention and as gateway services for young people in Australia who are seeking mental health information and support [34,36,37]. They also provide opportunities for young people to anonymously engage with peers online about mental health concerns, using moderated social media platforms, which itself has been shown to increase the likelihood of seeking formal mental health treatment [38,39]. A DMHP-Y is therefore required to map client-led online help seeking and engagement pathways to and within digital mental health services, and help clients navigate access to support for their individual needs, including seamless connections with offline mental health services.

Considerations in Evidence-Basing a DMHP-Y

Digital mental health approaches are typically evidence-based as stand-alone interventions (eg, apps, video counseling, websites, and forums) rather than as holistic models of care. For a DMHP-Y to be effective and well-used, more research is needed to build an evidence base for the following, in order to inform design decisions that reflect youths' help seeking behaviors and preferences, and the navigation of their mental health journey.

Understand and Monitor Digital Engagement With Mental Health Care Over Time

This is necessary to map how and at what stages of their mental health journey young people engage with the digital mental health system over time, as use of digital mental health information and tools is often not linear. It also needs to be understood why some demographics might be more likely to arrive at a digital mental health service via an internet search, rather than via clinician referral or social networks. Further, in relation to online search words, search engine optimization (SEO) strategies must account for the different search terms individuals use to find information on mental health, given that the vocabulary for mental health help seeking has been shown to differ based on developmental age [40].

Understanding Preferred Digital Tools of Young People for Supporting Their Mental Health

The online support services and modalities offered as part of a DMHP-Y must reflect and keep up with changing preferences and use patterns among young people. For example, there is growing evidence suggesting that young people want to use dynamic social media environments to seek information about their mental health and treatment options, and give and receive peer support from others facing similar mental health challenges [41-43]. While public social media platforms have significant privacy and security risks, recent studies have shown that custom-built social media–based services such as Kids Helpline's "My Circle" (formerly "KHL Circles") can be a safe and effective way for young people to engage in counselor-guided peer support for their mental health [25,44-46]. Next on the horizon is the increasing incorporation of virtual, augmented, and mixed reality technologies into social media, giving rise to the metaverse, which presents both opportunities and challenges for the future of digital mental health treatment and support [47,48]. A future-looking DMHP-Y model of care must therefore be dynamic enough to adapt quickly as technology and platform preferences continue to evolve, yet ensure that new service offerings are tested and evidence-based before implementation, and user privacy and safety remain the first priority.

Ensuring an Optimum and Inclusive User Experience

It is well established that participatory design research methodologies involving all stakeholders and end users in the design and development of digital mental health systems and interventions are needed to ensure an optimum user experience for clients, clinicians, and other support staff [49,50]. However, further research into understanding individual human factors across different ages, genders, education, locations, abilities, neurotypes, and cultural needs is crucial to providing an inclusive and dynamic experience, leading to better service matching, more engagement, and less dropout from services. Furthermore, all DMHP-Y web content should be created using "universal design" principles to ensure usability by people with the widest possible range of abilities in the widest possible range of situations [51]. For example, Kids Helpline has already adopted a mobile-first approach and has implemented a range of inclusivity features to allow for people with hearing or sight impairment to access their services, such as "voice to text" across all modalities, ensuring that all website text is accessible via screen readers, and that all video content includes subtitles and audio cues. Kids Helpline has also introduced design initiatives to support children and young people with limited access to devices, data, and network coverage.

Ensuring Continuity of Care

As discussed, a DMHP-Y model of care must provide for ongoing connections and transferability of information between online modalities and with offline services via EHRs. The need for and benefits of a transferable EHR should be made clear to clients upon first engagement with a digital mental health system. For example, so that they can easily navigate their mental health journey across modalities (ie, within and between online and offline services) and avoid having to repeat initial onboarding processes in the future. Participatory action research is needed to ensure that such a transferable EHR system is developed on the basis of user-centered design principles so that it meets the needs of both clinicians and clients, and effectively facilitates continuity of care (especially for nonlinear mental health treatment journeys). Once the evidence base for this is established, a standardized EHR system should ideally be implemented nationally across all community and private mental health services to ensure comprehensive and seamless record keeping between online and offline services.

My Health Records (Australia's national digital health record system managed by the Australian Digital Health Agency) provides a case study in how difficult it can be to roll out an EHR system nationally. Challenges include low uptake, incomplete information, and concerns around privacy and security [52]. While it is possible for My Health Records to be used for mental health care, it is currently underused for this purpose, with concerns raised about the inclusion of sensitive and potentially stigmatizing mental health information alongside physical health information [24]. Furthermore, psychologists and other allied health professionals are excluded from engaging meaningfully with My Health Records, as conformant software is not currently available for allied health practices. My Health Records also does not allow for the inclusion of records of engagement with NGOs' digital mental health services. Therefore, our perspective is that a DMHP-Y model of care requires a national EHR system specifically for mental health, which facilitates the transferring of records between digital mental health services and other health providers. While the considerations are complex and beyond the scope of this paper, any EHR system that seeks to support the mental health journeys of young people needs to include records of engagement with digital mental health services (most of which in Australia, at present, are NGOs and are outside of the health care system).

Proposing a Future-Looking DMHP-Y

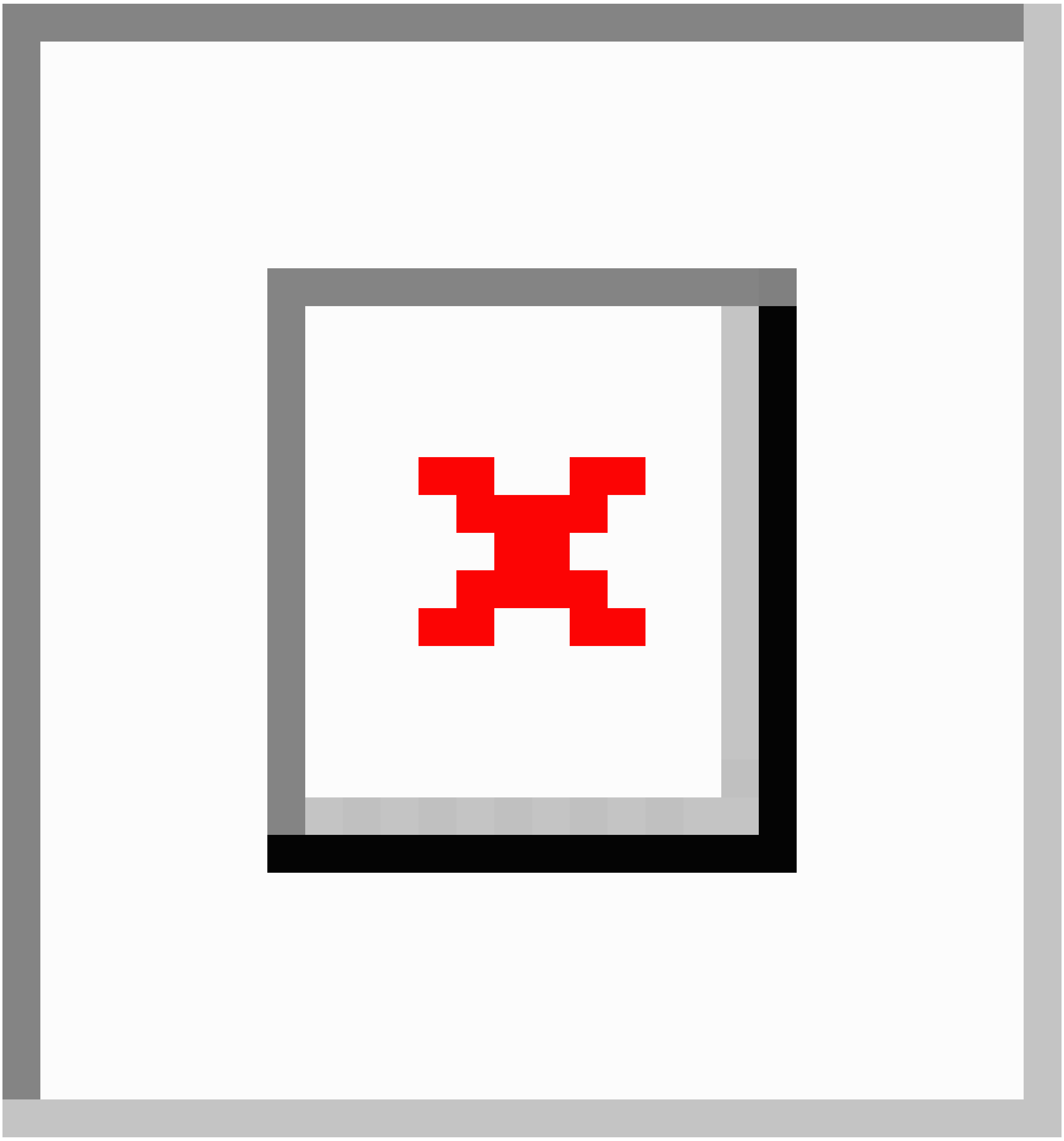
Overview

As outlined, current service provision for youth mental health in Australia offers both offline (eg, headspace and private practice allied health) and online support services (eg, eheadspace, Kids Helpline, and ReachOut). At the time of writing, the only emerging evidence-based digital platforms in Australia connecting clinicians and clients across online and offline service modalities are Innowell [25] and MOST [26,27]. While these models of care provide some basis for the model proposed below in Figure 1, they can only be accessed by clients after obtaining a referral from a participating offline mental health service.

In contrast, the DMHP-Y model of care proposed by the authors is focused on client-led online help seeking and engagement pathways. It aims to provide a map of the client-led engagement pathways that would be possible from the point at which a user first engages with a digital mental health service and is provided with a transferable EHR, regardless of whether they have previously engaged with offline mental health services or how far along their overall mental health journey they might be. While the model proposes that initial access to digital services (with the exception of phone counseling) should be via a common landing point for onboarding, it accounts for there being numerous engagement and re-engagement points (such as following engagement with offline services). The model also accounts for technologies that will change the way we interact online over the coming years, such as the incorporation of natural language processing and other artificial intelligence (AI) modalities, and avatar and social network use in the metaverse. The model presented is, therefore, a future-looking starting point for a model of care that would improve connections between services and transferability of client information, as technological advances and user preferences continue to evolve.

The proposed DMHP-Y model aims for what is herein dubbed the 3 Is of digital mental health care: *inclusivity* (equal access to all), *individuality* (choice based on preferences), and *integrity* (reliable, ethical, and evidence-based care). The 3 Is are addressed, where applicable, across each step of a client's movement through the system of care.

Figure 1. The future-looking Digital Mental Health Practice model of care for youth (version 1.0).



Step 1: First Engagement

This step represents the *individual* journeys and modalities that lead young people to first engage with a digital mental health service. As stated previously, young people often use the web as their first step in seeking support and information regarding their mental health [10,11], usually via a text-based query in an internet search engine, which is the most common help seeking approach among young people aged <25 years [12]. Other potential initial engagement channels include suggested links on websites or social media, recommendations from text-based chat services and social media, as well as clinician referral and word of mouth.

Research on young people's mental health–related search word practices suggests that digital mental health services need to ensure that they can be easily found not only using diagnostic search terms (eg, depression, anxiety, etc), but also using simple affective phrases (eg, sad, scared, or lonely) [40]. SEO strategies also need to account for differences in demographics, ability, neurotype, language, and cultural backgrounds to ensure *inclusivity*. Moreover, ongoing analysis of common search phrases for mental health issues to inform SEO updates is key to a DMHP-Y model remaining a primary online source of help (*integrity*) over less evidence-based or predatory mental health offerings.

Step 2: Landing Page

No matter what leads a potential client to first engage with a DMHP-Y system, they should be directed to a common landing page that quickly explains the service modality options available to them (*individuality*), as well as a clearly identifiable link to "immediate help" (ie, a link to a national mental health crisis line [53]) to reduce cognitive load on what could be a distressed individual (*integrity*). Often, landing pages are cluttered with dense information or vague navigation. Studies clearly show that time spent engaging with a digital platform is directly related to intuitive design and consideration of user digital literacy [54,55]. The design of the landing page is therefore crucial and should not be dictated solely by the service provider but rather be designed in consultation with young people from a diverse range of backgrounds and with differing levels of mental health literacy (*inclusivity*). This is to ensure that the right balance is achieved between the need to provide essential information to the client and the need of the client to progress to an appropriate service quickly and easily. This essential information should include explaining the need for and benefit of creating a transferable EHR to navigate their mental health journey, as the next step in the model may require this for progress into the service.

Step 3: Onboarding and Virtual Triage

The next step is to create a transferable EHR for the client (if they do not already have one) so they can onboard into the DMHP-Y system and complete a mental health screening assessment to inform "virtual triage." It has been proposed that virtual triage could also leverage AI, specifically natural language processing, which shows promise for adaptation for mental health services [55-58] and is currently being used successfully in corporate operations and retail services. It is not the authors' suggestion that such a sensitive system and service model be reliant on AI alone, rather, a hybrid model of triage engagement be offered (ie, human and AI), depending on the level of severity of a client upon initial screening. An example of this would be when a client has completed the initial onboarding for an EHR and their mental health screen indicates high levels of distress (addressing *individual* urgent need), they are flagged as such and directed immediately to a human counselor contact (eg, text-based or voice-based, depending on the preference of the client) as the first port of call for immediate triage (ie, *integrity* of priority care). For clients who decline contact with a human counselor, the system may provide them with targeted resources and alternate options for support (*inclusivity*), while flagging their profile within the ecosystem for therapeutic monitoring, follow-up, and support. Clients identified as "severe" at screening can, if required, be directed to emergency services, or referred to community or private mental health services via a contact number or URL for their closest service.

Step 4: Online Support

Once level of care requirements are established, the DMHP-Y model can offer a personalized (ie, *individual* and *inclusive*) menu of text- and voice-based choices from a suite of digital therapies and supports. Low-risk first-time clients may be directed to multimedia psychoeducation, or virtual counselor sessions can be scheduled in an environment of choice (eg, an initial telecounseling session via phone, web, or metaverse avatar counseling). This initial entrance level of support could also include evidence-based AI chatbots for mental health information [56,58] or asynchronous text-based guidance from counselors [59]. In turn, this would aid them to learn more about their mental health needs and options for individual care, both within the online service and via offline services. With appropriate technology and client consent in place, there is also potential for digital phenotyping to play a role in informing clinical support needs, by integrating data from a client's smart devices with their engagement trends with the DMHP-Y model over time [30].

For clients who understand their mental health needs (ie, have previously received diagnosis or therapies), the option of self-guided psychoeducation and digital programs, such as cognitive behavioral therapy or acceptance and commitment therapy, may be offered. These forms of digital therapies are evidence-based for clients experiencing depression, anxiety, and stress symptoms and wanting to self-manage [60,61].

Clients who are seeking continuous social support, especially for chronic mental health conditions, may benefit from community engagement via counselor-moderated 24/7 peer-to-peer social networks. This level of ongoing social support has been found to benefit young people in managing their mood and learning new coping strategies, and offers social support outside of specific therapies [26,44-46]. Connecting for peer-to-peer support in counselor-moderated social networks such as Kids Helpline's My Circle has also been shown to reduce mental health stigma and increase help seeking for mental health issues in the future [44-46]. Kids Helpline's My Circle is an example of how peer-to-peer social networks can integrate and link directly to other text-based resources such as self-directed online psychoeducation modules, website content, and text-based one-to-one counseling in an easy-to-navigate digital ecosystem. The integration of digital mental health resources and interventions into an ecosystem further allows them to be used strategically and dynamically by digital counselors, by conceptualizing them on dichotomies from low to high intensity (eg, website vs one-to-one text-based counseling) or solo to assisted interventions. For example, clients presenting with "severe" symptomatology who are struggling to see improvements by using self-directed online psychoeducation may conceptually benefit from using higher-intensity and assisted interventions, which may involve online counselors directing them to peer-to-peer support or one-to-one interventions that match their willingness, motivation, and clinical needs.

Offline Support

Though not a "step" in the DMHP-Y model, referral to offline face-to-face support may be deemed appropriate or necessary at any stage of a client's engagement with a digital mental health service. In this situation, digital counselors should provide "warm referrals" for clients; that is, they should discuss options and gain client consent to introduce them to an offline service. A system for sharing of EHRs must therefore be established between a DMHP-Y model and offline services for seamless recordkeeping and best-practice care for clients in need of outpatient or inpatient care. This, of course, goes both ways, as clients may transition to digital services having already engaged with offline services. An EHR system for mental health should therefore ideally be implemented nationally across all community, private mental health, and digital mental health services and allow clients who may have an existing EHR established with offline services to connect it to digital mental health services (as many seek digital services themselves rather than via a referral from an offline service).

At the time of writing, there are multiple health record software programs on offer in Australia, but none provide seamless recordkeeping between online and offline services. There are, however, digital mental health services in Australia that seek to connect young people to and from offline services via direct referrals, albeit without fully transferable EHR systems. For example, Kids Helpline connects young people to offline services via warm referrals to community and specialist mental health services and emergency and child protection services where clients are at risk of significant harm. Conversely, as discussed, following a referral from participating offline services, young people can connect to Orygen's MOST platform [26,27] to access online support either while they are waiting for face-to-face care in order to augment their face-to-face sessions or after discharge. It is these types of systems of care that need to be further developed and tested at the national level.

Conclusion

Australia presently overrelies on traditional triage models of care for mental health (eg, hospital emergency departments and crisis support telephone services), which came close to systemic failure during the peak of the COVID-19 pandemic as they struggled to address increased demand for mental health support during the health crisis [62]. Given that the demand for online services is growing despite many digital mental health services currently having a poor evidence base [63], it is now even more necessary to develop a road map, testing phase, and rollout plan for a DMHP-Y model of care for service provision.

User-preferred models for mental health that provide EHRs to enable seamless cross-referrals between online and offline allied health and psychiatric care will reduce pressures on existing mental health providers in Australia and facilitate continuity of care for clients. This will be achieved by maintaining detailed records of the young person's mental health journey—tracking engagement with both clinical treatment and digital support tools—which will allow seamless repeat entrance to both online and offline support when needed. The goal should be for young people to be able to enter services for support without having to retell their story wherever possible.

The primary challenge is the "start up" phase for a DMHP-Y model of care, which must address four significant hurdles: (1) synchronization of existing systems and models of mental health care, online and offline, across different provider organizations (NGOs, government, private sectors, GPs, and hospitals); (2) establishing a transparent, secure, and agile EHR system for mental health; (3) tracking and managing clients that present within digital systems under multiple identities (eg, due to stigma or privacy concerns); and (4) upskilling mental health workers and professionals in digital health and digital therapy skills to extend evidence-based practices across existing (eg, videoconference counseling) and future technologies (eg, metaverse counseling and AI). Developing a dynamic and future-looking DMHP-Y model of care that incorporates the introduction of new automated technologies such as AI does not imply the loss of quality mental health professional care. On the contrary, a DMHP-Y model of care that incorporates AI facilitates the economic and professional release of clinicians to dedicate time and expertise to clients who are in urgent need of help or require long-term care across hybrid offerings of traditional community face-to-face resources and digital resources.

The investment and development in DMHP-Y models of care will improve services in Australia from "treatment only" to "treatment and user-choice support" systems. In turn, this will (1) alleviate pressure on emergency rooms (which have seen a dramatic increase in mental health presentations over the past decade) [64]; (2) improve online to offline referral practices (and vice versa); (3) better identify and prioritize those who need immediate face-to-face help from those who need ongoing social and psychoeducation support; (4) greatly improve the cost efficiency of the current mental health system in Australia; and (5) pioneer a national model of digital mental health care that is replicable globally.

The most significant risk to our mental health system is inaction. If we do not start implementing changes in digital systems in mental health care at the national level, the knock-on effects of poor mental health care management for young people will continue into adulthood. This will contribute to significant long-term economic and biopsychosocial problems for Australia's population in the future, compounding national health burdens of disability and disease. Moreover, the longer the delay to develop, test, and refine such digital models for mental health services, the more complex the task will become as new technologies and platforms such as AI and the metaverse are introduced at an increasingly rapid pace.

Conflicts of Interest

None declared.

References

1. Duggan M, Harris B, Chislett WK, Calder R. Nowhere Else to Go: Why Australia's Health System Results in People With Mental Illness Getting 'Stuck' in Emergency Departments: Mitchell Institute, Victoria University; 2020.

2. Griffiths KM, Mendoza J, Carron-Arthur B. Whereto mental health reform in Australia: is anyone listening to our independent auditors? Med J Aust 2015 Mar 2;202(4):172-174. [doi: 10.5694/mja14.01034] [Medline: 25716589]
3. Rosenberg S, Salvador-Carulla L. Perspectives: accountability for mental health: the Australian experience. J Ment Health Policy Econ 2017 Mar 1;20(1):37-54. [Medline: 28418836]
4. Rosenberg S, Hickie I. No gold medals: assessing Australia's international mental health performance. Australas Psychiatry 2019 Feb;27(1):36-40. [doi: 10.1177/1039856218804335] [Medline: 30293454]
5. Kelshaw S, Boyd J, Gallagher I, Sara G. Consumer experience of mental health services during the COVID-19 pandemic: evidence from an Australian mental health system. Patient Exp J 2022;9(1):72-81. [doi: 10.35680/2372-0247.1658]
6. Kwan B, Rickwood DJ. A systematic review of mental health outcome measures for young people aged 12 to 25 years. BMC Psychiatry 2015 Nov 14;15:279. [doi: 10.1186/s12888-015-0664-x] [Medline: 26573269]
7. McGorry PD, Goldstone SD, Parker AG, Rickwood DJ, Hickie IB. Cultures for mental health care of young people: an Australian blueprint for reform. Lancet Psychiatry 2014 Dec;1(7):559-568. [doi: 10.1016/S2215-0366(14)00082-0] [Medline: 26361315]
8. McGorry PD, Mei C. Early intervention in youth mental health: progress and future directions. Evid Based Ment Health 2018 Nov;21(4):182-184. [doi: 10.1136/ebmental-2018-300060] [Medline: 30352884]
9. Kilian A, Williamson A. What is known about pathways to mental health care for Australian aboriginal young people?: a narrative review. Int J Equity Health 2018 Jan 27;17(1):12. [doi: 10.1186/s12939-018-0727-y] [Medline: 29374482]
10. Bradford S, Rickwood D. Young people's views on electronic mental health assessment: prefer to type than talk? J Child Fam Stud 2015;24(5):1213-1221. [doi: 10.1007/s10826-014-9929-0] [Medline: 25960628]
11. Gibson K. What Young People Want From Mental Health Services: A Youth Informed Approach for the Digital Age: Routledge; 2021. URL: https://www.taylorfrancis.com/books/9780429322457 [accessed 2024-03-15] [doi: 10.4324/9780429322457]
12. Pretorius C, Chambers D, Coyle D. Young people's online help-seeking and mental health difficulties: systematic narrative review. J Med Internet Res 2019 Nov 19;21(11):e13873. [doi: 10.2196/13873] [Medline: 31742562]
13. Roland J, Lawrance E, Insel T, Christensen H. The Digital Mental Health Revolution: Transforming Care Through Innovation and Scale-Up: World Innovation Summit for Health; 2020.
14. Schwarz J, Bärkås A, Blease C, et al. Sharing clinical notes and electronic health records with people affected by mental health conditions: scoping review. JMIR Ment Health 2021 Dec 14;8(12):e34170. [doi: 10.2196/34170] [Medline: 34904956]
15. MacDonald K, Ferrari M, Fainman-Adelman N, Iyer SN. Experiences of pathways to mental health services for young people and their carers: a qualitative meta-synthesis review. Soc Psychiatry Psychiatr Epidemiol 2021 Mar;56(3):339-361. [doi: 10.1007/s00127-020-01976-9] [Medline: 33206200]
16. Rickwood DJ, Telford NR, Parker AG, Tanti CJ, McGorry PD. Headspace—Australia's innovation in youth mental health: who are the clients and why are they presenting. Med J Aust 2014 Feb 3;200(2):108-111. [doi: 10.5694/mja13.11235] [Medline: 24484115]
17. Hilferty F, Cassells R, Muir K, Katz I. Is Headspace Making a Difference to Young People's Lives?: Social Policy Research Centre, UNSW Australia; 2016. URL: https://apo.org.au/node/71950 [accessed 2024-03-15]
18. Kisely S, Looi JC. Latest evidence casts further doubt on the effectiveness of headspace. Med J Aust 2022 Oct 17;217(8):388-390. [doi: 10.5694/mja2.51700] [Medline: 36182662]
19. Hiscock H, Mulraney M, Efron D, et al. Use and predictors of health services among Australian children with mental health problems: a national prospective study. Aust J Psychol 2020 Mar 1;72(1):31-40. [doi: 10.1111/ajpy.12256]
20. Meadows GN, Enticott JC, Inder B, Russell GM, Gurr R. Better access to mental health care and the failure of the medicare principle of universality. Med J Aust 2015 Mar 2;202(4):190-194. [doi: 10.5694/mja14.00330] [Medline: 25716601]
21. Rosenberg SP, Hickie IB. The runaway giant: ten years of the better access program. Med J Aust 2019 Apr;210(7):299-301. [doi: 10.5694/mja2.50068] [Medline: 30929255]
22. Ennis L, Rose D, Callard F, Denis M, Wykes T. Rapid progress or lengthy process? Electronic personal health records in mental health. BMC Psychiatry 2011 Jul 26;11:117. [doi: 10.1186/1471-244X-11-117] [Medline: 21791069]
23. Robotham D, Mayhew M, Rose D, Wykes T. Electronic personal health records for people with severe mental illness; a feasibility study. BMC Psychiatry 2015 Aug 6;15:192. [doi: 10.1186/s12888-015-0558-y] [Medline: 26245805]
24. Kariotis TC, Harris KM. Clinician perceptions of my health record in mental health care: medication management and sharing mental health information. Aust J Prim Health 2019 Mar;25(1):66-71. [doi: 10.1071/PY17181] [Medline: 30636668]
25. Davenport TA, Cheng VWS, Iorfino F, et al. Flip the clinic: a digital health approach to youth mental health service delivery during the COVID-19 pandemic and beyond. JMIR Ment Health 2020 Dec 15;7(12):e24578. [doi: 10.2196/24578] [Medline: 33206051]
26. Alvarez-Jimenez M, Rice S, D'Alfonso S, et al. A novel multimodal digital service (Moderated Online Social Therapy+) for help-seeking young people experiencing mental ill-health: pilot evaluation within a national youth e-mental health service. J Med Internet Res 2020 Aug 13;22(8):e17155. [doi: 10.2196/17155] [Medline: 32788151]
27. Clinical services - MOST. Orygen. 2023. URL: https://www.orygen.org.au/Clinical-Care/Clinical-services/most [accessed 2024-03-07]

28. Macrynikola N, Nguyen N, Lane E, Yen S, Torous J. The digital clinic: an innovative mental health care delivery model utilizing hybrid synchronous and asynchronous treatment. NEJM Catalyst 2023 Aug 16;4(9). [doi: 10.1056/CAT.23.0100]
29. Rodriguez-Villa E, Rauseo-Ricupero N, Camacho E, Wisniewski H, Keshavan M, Torous J. The digital clinic: implementing technology and augmenting care for mental health. Gen Hosp Psychiatry 2020;66:59-66. [doi: 10.1016/j.genhosppsych.2020.06.009] [Medline: 32688094]
30. Wisniewski H, Henson P, Torous J. Using a smartphone app to identify clinically relevant behavior trends via symptom report, cognition scores, and exercise levels: a case series. Front Psychiatry 2019;10:652. [doi: 10.3389/fpsyt.2019.00652] [Medline: 31607960]
31. Wright A, Jorm AF. Labels used by young people to describe mental disorders: factors associated with their development. Aust N Z J Psychiatry 2009 Oct;43(10):946-955. [doi: 10.1080/00048670903179129]
32. Brown A, Rice SM, Rickwood DJ, Parker AG. Systematic review of barriers and facilitators to accessing and engaging with mental health care among at-risk young people. Asia Pac Psychiatry 2016 Mar;8(1):3-22. [doi: 10.1111/appy.12199] [Medline: 26238088]
33. Rickwood D, Deane FP, Wilson CJ, Ciarrochi J. Young people's help-seeking for mental health problems. Aust E-J Adv Ment Health 2005 Jan;4(3):218-251. [doi: 10.5172/jamh.4.3.218]
34. Yung AR, Cotter J, McGorry PD. Help-seeking in young people. In: Yung AR, McGorry PD, Cotter J, editors. Youth Mental Health: Approaches to Emerging Mental Ill-Health in Young People: Routledge; 2021:35-52 URL: https://www.taylorfrancis.com/books/9781000292589 [accessed 2024-03-15] [doi: 10.4324/9780429285806]
35. Rickwood DJ, Mazzer KR, Telford NR. Social influences on seeking help from mental health services, in-person and online, during adolescence and young adulthood. BMC Psychiatry 2015 Mar 7;15:40. [doi: 10.1186/s12888-015-0429-6] [Medline: 25886609]
36. Collin PJ, Metcalf AT, Stephens-Reicher JC, et al. ReachOut.com: the role of an online service for promoting help-seeking in young people. Adv Ment Health 2011 Oct;10(1):39-51. [doi: 10.5172/jamh.2011.10.1.39]
37. Insights 2020: Insights into young people in Australia. : Kids Helpline; 2020 URL: https://www.yourtown.com.au/sites/default/files/document/Kids-Helpline-Insights-2020-Report-Final.pdf [accessed 2024-03-15]
38. Ridout B, Campbell A. The use of social networking sites in mental health interventions for young people: systematic review. J Med Internet Res 2018 Dec 18;20(12):e12244. [doi: 10.2196/12244] [Medline: 30563811]
39. Lawlor A, Kirakowski J. Online support groups for mental health: a space for challenging self-stigma or a means of social avoidance. Comput Hum Behav 2014 Mar;32:152-161. [doi: 10.1016/j.chb.2013.11.015]
40. Campbell AJ, Ridout BF, Linden M, Collyer B, Dalgleish J. A preliminary understanding of search words used by children, teenagers and young adults in seeking information about depression and anxiety online. J Technol Hum Serv 2018 Oct 2;36(4):208-221. [doi: 10.1080/15228835.2018.1518186]
41. Naslund JA, Grande SW, Aschbrenner KA, Elwyn G. Naturally occurring peer support through social media: the experiences of individuals with severe mental illness using YouTube. PLoS One 2014;9(10):e110171. [doi: 10.1371/journal.pone.0110171] [Medline: 25333470]
42. Lal S, Nguyen V, Theriault J. Seeking mental health information and support online: experiences and perspectives of young people receiving treatment for first-episode psychosis. Early Interv Psychiatry 2018 Jun;12(3):324-330. [doi: 10.1111/eip.12317] [Medline: 26810026]
43. Gowen K, Deschaine M, Gruttadara D, Markey D. Young adults with mental health conditions and social networking websites: seeking tools to build community. Psychiatr Rehabil J 2012;35(3):245-250. [doi: 10.2975/35.3.2012.245.250] [Medline: 22246123]
44. Campbell A, Ridout B, Amon K, Navarro P, Collyer B, Dalgleish J. A customized social network platform (Kids Helpline Circles) for delivering group counseling to young people experiencing family discord that impacts their well-being: exploratory study. J Med Internet Res 2019 Dec 20;21(12):e16176. [doi: 10.2196/16176] [Medline: 31859671]
45. Amon K, Ridout B, Forsyth R, Campbell A. Online group counseling for young people through a customized social networking platform: phase 2 of Kids Helpline Circles. Cyberpsychol Behav Soc Netw 2022 Sep;25(9):580-588. [doi: 10.1089/cyber.2022.0044] [Medline: 35951017]
46. Ridout B, Amon K, Forsyth R, Campbell A. Development and Evaluation of Kids Helpline's My Circle: The University of Sydney; 2023.
47. Petrigna L, Musumeci G. The metaverse: a new challenge for the healthcare system: a scoping review. J Funct Morphol Kinesiol 2022 Aug 30;7(3):63. [doi: 10.3390/jfmk7030063] [Medline: 36135421]
48. Benrimoh D, Chheda FD, Margolese HC. The best predictor of the future-the metaverse, mental health, and lessons learned from current technologies. JMIR Ment Health 2022 Oct 28;9(10):e40410. [doi: 10.2196/40410] [Medline: 36306155]
49. Petersen C, Adams SA, DeMuro PR. mHealth: don't forget all the stakeholders in the business case. Med 2 0 2015 Dec 31;4(2):e4. [doi: 10.2196/med20.4349] [Medline: 26720310]
50. Ospina-Pinillos L, Davenport T, Iorfino F, et al. Using new and innovative technologies to assess clinical stage in early intervention youth mental health services: evaluation study. J Med Internet Res 2018 Sep 10;20(9):e259. [doi: 10.2196/jmir.9966] [Medline: 30201602]

51. Henry SL, Abou-Zahra S, Brewer J. The role of accessibility in a universal web. In: W4A ’14: Proceedings of the 11th Web for All Conference: Association for Computing Machinery; 2014:1-4 URL: https://dl.acm.org/doi/10.1145/2596695.2596719 [accessed 2024-03-15] [doi: 10.1145/2596695.2596719]
52. Pang PI, McKay D, Chang S, Chen Q, Zhang X, Cui L. Privacy concerns of the Australian My Health Record: implications for other large-scale opt-out personal health records. Inf Process Manag 2020 Nov;57(6):102364. [doi: 10.1016/j.ipm.2020.102364]
53. Healthdirect: mental health helplines. Healthdirect Australia Limited. URL: https://www.healthdirect.gov.au/mental-health-helplines [accessed 2024-03-07]
54. Yu N, Kong J. User experience with web browsing on small screens: experimental investigations of mobile-page interface design and homepage design for news websites. Inf Sci 2016 Feb;330:427-443. [doi: 10.1016/j.ins.2015.06.004]
55. van der Schyff EL, Ridout B, Amon KL, Forsyth R, Campbell AJ. Providing self-led mental health support through an artificial intelligence-powered chat bot (Leora) to meet the demand of mental health care. J Med Internet Res 2023 Jun 19;25:e46448. [doi: 10.2196/46448] [Medline: 37335608]
56. Abd-Alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M. An overview of the features of chatbots in mental health: a scoping review. Int J Med Inform 2019 Dec;132:103978. [doi: 10.1016/j.ijmedinf.2019.103978] [Medline: 31622850]
57. Denecke K, Abd-Alrazaq A, Househ M. Artificial intelligence for chatbots in mental health: opportunities and challenges. In: Househ M, Borycki E, Kushniruk A, editors. Multiple Perspectives on Artificial Intelligence in Healthcare: Springer; 2021:115-128. [doi: 10.1007/978-3-030-67303-1]
58. Boucher EM, Harake NR, Ward HE, et al. Artificially intelligent chatbots in digital mental health interventions: a review. Expert Rev Med Devices 2021 Dec;18(sup1):37-49. [doi: 10.1080/17434440.2021.2013200] [Medline: 34872429]
59. Sindahl TN, Fukkink RG, Helles R. SMS counselling at a child helpline: counsellor strategies, children’s stressors and well-being. Br J Guid Couns 2020 Mar 3;48(2):263-275. [doi: 10.1080/03069885.2019.1580676]
60. Kelson J, Rollin A, Ridout B, Campbell A. Internet-delivered acceptance and commitment therapy for anxiety treatment: systematic review. J Med Internet Res 2019 Jan 29;21(1):e12530. [doi: 10.2196/12530] [Medline: 30694201]
61. Twomey C, O’Reilly G. Effectiveness of a freely available computerised cognitive behavioural therapy programme (MoodGYM) for depression: meta-analysis. Aust N Z J Psychiatry 2017 Mar;51(3):260-269. [doi: 10.1177/0004867416656258] [Medline: 27384752]
62. McGorry P. Mental health and COVID-19: are we really all in this together? Med J Aust 2020 Nov;213(10):454-455. [doi: 10.5694/mja2.50834] [Medline: 33107074]
63. Philippe TJ, Sikder N, Jackson A, et al. Digital health interventions for delivery of mental health care: systematic and comprehensive meta-review. JMIR Ment Health 2022 May 12;9(5):e35159. [doi: 10.2196/35159] [Medline: 35551058]
64. Mental health services provided in emergency departments. Australian Institute of Health and Welfare. 2023. URL: https://www.aihw.gov.au/mental-health/topic-areas/emergency-departments [accessed 2024-03-08]

Abbreviations

AI: artificial intelligence
DMHP-Y: Digital Mental Health Practice model of care for youth
EHR: electronic health record
GP: general practitioner
HIT: health IT
mindLAMP: Learn, Assess, Manage, Prevent
MOST: Moderated Online Social Therapy
NGO: nongovernment organization
SEO: search engine optimization

Edited by J Torous; submitted 24.04.23; peer-reviewed by D Rickwood, J Matar; revised version received 07.01.24; accepted 25.01.24; published 22.03.24.

Please cite as:
Ridout B, Forsyth R, Amon KL, Navarro P, Campbell AJ
The Urgent Need for an Evidence-Based Digital Mental Health Practice Model of Care for Youth
JMIR Ment Health 2024;11:e48441
URL: https://mental.jmir.org/2024/1/e48441
doi:10.2196/48441

Viewpoint

Data Integrity Issues With Web-Based Studies: An Institutional Example of a Widespread Challenge

Blandine French[1], PhD; Camilla Babbage[1], PhD; Katherine Bird[1], MA; Lauren Marsh[1], PhD; Mirabel Pelton[1], PhD; Shireen Patel[1], MA; Sarah Cassidy[1], PhD; Stefan Rennick-Egglestone[1], PhD

School of Psychology, University of Nottingham, Nottingham, United Kingdom

Corresponding Author:
Blandine French, PhD
School of Psychology
University of Nottingham
University Park Campus
Psychology, Pharmacy, Life Sciences, East Dr
Nottingham, NG7 2RD
United Kingdom
Phone: 44 0115 748 6970
Email: blandine.french@nottingham.ac.uk

Related Article:

This is a corrected version. See correction statement: https://mental.jmir.org/2024/1/e67286

Abstract

This paper reports on the growing issues experienced when conducting web-based–based research. Nongenuine participants, repeat responders, and misrepresentation are common issues in health research posing significant challenges to data integrity. A summary of existing data on the topic and the different impacts on studies is presented. Seven case studies experienced by different teams within our institutions are then reported, primarily focused on mental health research. Finally, strategies to combat these challenges are presented, including protocol development, transparent recruitment practices, and continuous data monitoring. These strategies and challenges impact the entire research cycle and need to be considered prior to, during, and post data collection. With a lack of current clear guidelines on this topic, this report attempts to highlight considerations to be taken to minimize the impact of such challenges on researchers, studies, and wider research. Researchers conducting web-based research must put mitigating strategies in place, and reporting on mitigation efforts should be mandatory in grant applications and publications to uphold the credibility of web-based research.

(JMIR Ment Health 2024;11:e58432) doi:10.2196/58432

KEYWORDS

web-based research; web-based studies; qualitative studies; surveys; mental health; data integrity, misrepresentation

Introduction

Web-based research, such as web-based surveys or qualitative interviews via videoconferencing platforms, has grown in popularity and usability in the last 2 decades but more specifically since 2020 and the COVID-19 global pandemic [1-3]. Web-based studies have enabled researchers to continue conducting research studies when in-person testing was not possible [4], facilitating the recruitment of large samples, specific samples, or populations often excluded from research (described as underserved communities, or "hard to reach") [5,6]. It also offers other benefits such as minimizing recruitment costs [7] or allowing for anonymity, often favored by specific populations [8]. Web-based anonymity is attractive for certain groups as the researcher does not meet the participants and may even allow participation from individuals who might otherwise feel excluded from research; however, this lack of in-person validation poses a distinct threat to the validity, quality, and integrity of the data [9].

While web-based research has been heavily used by researchers due to its many advantages, this increased use has also led to researchers questioning whether their web-based participant pool is genuine. Web-based research describes "any research involving the remote acquisition of data from or about human participants using the internet and its associated technologies" [10], including both quantitative (eg, surveys or questionnaires) and qualitative methodologies (eg, focus groups and interviews). The following 5 categories of behavior, by participants and others, can significantly impact data quality and integrity in web-based studies:

1. Nongenuine participants: participants lying about their lived experiences or identity.
2. Repeat responders: participants taking part more than once.
3. Misrepresentation: exaggeration of specific details.
4. Lack of engagement: participants answering quickly, not paying attention and not reading the questions fully.
5. Bots: automated software application that performs repetitive tasks over a network.

These categories encompass not only different methodologies (qualitative and quantitative) but also different samples (general population and specific samples). While historically web-based studies were used by behavioral scientists (often aimed at the general population), health research (often aimed at specific groups) has also increasingly transitioned online [11]. Primarily issues relate to repeat responders, lack of engagement, and bots [12,13]. To remedy and minimize these issues many "secure" web-based behavioral platforms have been developed, such as Amazon Mechanical Turk [14] or Prolific Academic [15]; however, their reliability is often contested [12,16]. These issues have a significant impact on data integrity with lack of engagement, for instance, as high as 40% [17].

When conducting our research with specific participant groups with a range of health and neurodevelopment conditions, we have often experienced issues with nongenuine participants, repeat responders, and misrepresentation. Many studies indicate that a subset of participants in web-based surveys tends to falsely assert eligibility to obtain entry into the study, particularly when the incentive for participation is monetarized [18,19]. Hydock (2017) [20] reported that a "small but nontrivial portion of participants in online survey studies misrepresented their identity for the chance of financial gain."

This publication focuses solely on health web-based research and its most common barriers, namely, nongenuine participants, repeat responders, and misrepresentation. Challenges and strategies to overcome these difficulties are presented, most relevant to researchers in this field of work, which requires necessary caution given the complex health conditions and participants being regarded.

The motivations for repeat responders or nongenuine participants are varied, including monetary incentive [19,20], accessing a special intervention or treatment [21], or be politically motivated [22]. Indeed, 1 study experienced purposeful interference and registration in their study from specific groups with vested interests [22]. Finally, repeat registration can also be due solely to human error, for example, not realizing that they had already signed up, though insincere behaviors could be to receive further financial incentives.

Many studies have reported the significant impacts of these issues on their studies. In a series of 3 studies, Chandler and Paolacci [19] found that between 3% and 40% of participants misrepresented themselves, for example, saying that they were US residents when they were not. Bowen et al [23] recorded the numbers of repeated responses (627/1900, 33%) and demonstrated the wide extent by classifying them into 4 categories: infrequent (2-5 responses from same person), persistent (6-10), repeater (11-30), and hackers (45-67). These issues have also been reported in qualitative web-based interviews [24-26], for example, Roehl and Harland [24] reported 5 out of 14 participants as inauthentic. Finally, many studies report great difficulties after advertising on social media. Salinas [27] found that only 9.8% (50/512) of respondents from Facebook adverts were genuine while Pozzar et al [28] reported that after 7 hours of advertising their study on Twitter, 94.5% (256/271) were inauthentic and 16.2% (44/271) of them showed evidence of bot automation.

The impact of these issues on data integrity is significant. Chandler et al [29] demonstrated that inauthentic data have a significant impact on experiments—by increasing between-group variance, on individual differences—by providing false information, and on the association between variables—by suppressing, inflating, or reversing correlation. For example, Ysidron et al [16] found that nongenuine participants in a study on diabetes (150/307, 49% of the sample) reported significantly poorer physical and mental health issues than the clinical group, suggesting substantial exaggeration of adverse health. This leads to incorrect conclusions, potentially creating inappropriate recommendations for both practice and future research. The impact of these issues also greatly increases research costs both in terms of financial cost and time [30].

After having experienced many of these issues in their health research studies, the research team collated their individual experiences. Strategies and tips were exchanged, leading to institution-wide presentation and awareness to inform any future studies and initiating institutionally recognized strategies. This publication reports a sample of case studies experienced within our UK-based institution at the University of Nottingham and reflections on the methodological challenges and strategies extracted from the high-quality publications in the area.

Approach

Case Studies

We present 7 case studies experienced within our institution, all relating to mental health research. It is not possible to summarize all issues experienced within our institution in 1 paper. Consequently, this list is not exhaustive of issues experienced but rather it describes some examples of the breadth of this issue and how it impacted each research team. Table 1 shows the different case studies included in this manuscript.

Table 1. University of Nottingham case studies.

Case study	Methodology used	Category of issues experienced	Participants impacted, n (%)	Impact on study
1	Qualitative: interviews	Nongenuine Participants Repeat responders	8/12 (66.67) were nongenuine data.	Loss of funds, inability to finish project or publish, time wasted by PI[a] and researcher to identify data.
2	Randomized controlled trial	Repeat responders	100/1123 (8.9) accounts were suspended due to repeat registration.	Human resources to develop and administer repeat registration protocol and suspend accounts suspected of repeat registration. Additional burden and distress for participants who repeatedly registered in error.
3	Randomized controlled trial	Nongenuine participants Repeat responders	Approximately 482/483 (99.79) were suspected as nongenuine participants.	Impact on study researchers in terms of increased workload but it did not impact on recruitment figures.
4	Quantitative: survey	Nongenuine participants Repeat responders	349/391 (89.26) were suspected as nongenuine participants.	The process of cleaning the data was difficult in decision-making, time-consuming, and led to delays in completing the project. Reluctance to widely advertise afterward.
5	Qualitative: interviews	Nongenuine participants	54/54 (100) were suspected as nongenuine; 7 were identified as nongenuine. No genuine participants were recruited.	Loss of time, had to widen recruitment to the general population to deliver the project.
6	Qualitative: focus groups	Nongenuine participants; repeat responders	Approximately 115 suspected nongenuine participants applied. 83% of selected participants were nongenuine.	Loss of funds, loss of time, and loss of data integrity.
7	Quantitative: administration of online task	Nongenuine participants	20/31 (64.52) nonautistic participants consented to the study were nongenuine.	Loss of comparison group, unable to fully deliver funder expectations, time, and stress.

[a]PI: principal investigator.

Case Study 1: Web-Based Interviews on Cognitive Behavior Therapy and Attention-Deficit/Hyperactive Disorder

What Happened?

The study aimed to establish the experience of cognitive behavior therapy (CBT) in adults with attention-deficit/hyperactive disorder (ADHD) through semi-structured interviews as part of a masters' student project. The research study invited adults with ADHD from the principal investigator's research database and received multiple and prompt expressions of interest. She interviewed 12 participants in a couple of weeks but after finishing the data collection, she felt that there was "something wrong" with some interviews.

How Did You Find Out?

Out of 12 interviews, 8 were suspicious. The interviewer felt that the responses were very brief and not in-depth, some stories were very similar if not identical, and the participants followed a similar pattern. All 8 interviews had similar format of Gmail email addresses, refused to put their camera on, related similar stories, had identical non-British accents, and were actively asking about payments. In addition, from our decade-long experience in this topic, we know that most adults with ADHD have very difficult experiences with CBT, and this group all reported very positive experiences of CBT which is very unusual. As only participants from the database had been invited, all participants (n=6) who were not on the database were excluded automatically after asking how they heard about the study. The remaining 2 narratives were too similar to some of the excluded ones and were therefore also excluded.

How Did It Impact Your Research (Financial and Time Costs)?

The whole student project was compromised by this issue. As participants were paid £20 (US $26) per interview, no extra budget was available to conduct additional interviews as well as no extra time within the constraint of the master's dissertation submission. The analysis had to be conducted on 4 interviews alone, which was acceptable for the dissertation purposes (after explaining the issue to the head of teaching) but not for publication. The student was very disappointed by the amount of time that had been wasted on these and the impact it had on her opportunity to get her project published.

Case Study 2: Randomized Controlled Trials, Narrative Experiences Online

What Happened?

The Narrative Experiences Online study conducted 2 definitive pragmatic web-based trials of a web-based digital health intervention, which provided access to a collection of mental health recovery narratives. To participate in the trial, UK residents completed a web-based informed consent form and registered a web-based account using a personal email address. Formal identity verification was not required by our regulator, and we chose not to use it, to avoid contributing to paranoia for a trial population who had a personal experience of psychosis (for the Narrative Experiences Online Trial) [31]. We were prepared for the possibility of repeat registration of accounts due to published accounts of previous web-based trials [21], and so our protocol included an outline procedure to suspend accounts suspected of repeat registrations [31]. As recruitment to our trials progressed, we observed a range of instances of suspected repeat registrations. We decided to formalize our decision-making procedures on account suspension.

How Did You Find Out?

The web application that delivered our trials included an administrative control panel providing access to information about all registered accounts. Through regular monitoring, we saw instances of sequential registrations, using email addresses that appeared related, with little time between each registration, or from the same IP address. We communicated with some participants and learnt that some instances were accidental repeat registrations due to confusion with trial procedures and others were deliberate (including to obtain payment vouchers). Many participants did not respond to our messages. In some cases, there were legitimate reasons for these indicators. These included registration by more than 1 person from the same household, often on the same device, where trial information had been shared among the household.

How Did It Impact Your Research?

We spent a substantial effort on a protocol for decision-making about account suspensions (Multimedia Appendix 1), which was necessary given ambiguous decision-making (eg, where some participants did not respond to contact). Our protocol developed by the research team was amended and authorized by a trial management group and subsequently authorized by a program steering committee. We produced our protocol for the important purposes of (1) supporting trial integrity, and (2) enabling ethical conduct in communication with participants. Enacting the protocol required a substantial amount of time on behalf of administrators, who collected information on possible repeat registrations in "Repeat Registration Reports" (Multimedia Appendix 2). It also required time from the decision makers who reviewed those reports. For analytical integrity, further effort was spent on articulating a modified intention-to-treat principle, which excluded accounts suspended due to repeat registrations [32], and on developing a modified CONSORT diagram to report on suspensions [32,33].

Case Study 3: Randomized Controlled Trial, Mindful Life-Well at Work

What Happened?

The Mindful Life-Well at Work study is a randomized controlled trial (RCT) assessing whether Mindfulness-Based Cognitive Therapy—for Life is more effective than stress reduction psychoeducation. Participants are recruited from health care, social care, and teaching sectors. To participate in the trial, clinical staff who receive the study flyer circulated by their participating NHS trust sites complete a web-based eligibility screening and informed consent form. All forms and questionnaires are completed independently by participants using the web-based REDCap (Research Electronic Data Capture) platform. During the second recruitment wave we received an influx of emails; 483 emails were received over 5 days. The emails were sent out at strange times (eg, 3 AM) and a large number of emails were received within a few minutes of each other. The number received was considerably larger than in earlier recruitment periods.

How Did You Find Out?

Initially, the study researchers did not suspect any dubious activity. However, on the second day when a study researcher was recording the data for the noneligible participants, the researcher noticed that all inclusion or exclusion criteria had been selected for most cases. This was peculiar, as data from previous participants implied that usually 1 or 2 criteria were selected. Upon closer scrutiny, they noticed that the email account names did not match the names provided. The study team also observed that signatures on the consent forms indicated that the same person was completing them. The first batch of emails was sent out from Gmail and the second batch was from Hotmail. The data on REDCap suggested that many tabs had been opened at the same time and forms completed one at a time, that is, submitted each consent form approximately 5 seconds after each other.

How Did It Impact Your Research (Financial and Time Costs)?

Despite the study team feeling quite convinced that these were not emails from actual participants, they investigated further to ensure that actual participants were not being missed. The senior database manager was consulted for advice. He informed the study team that as IP addresses were not collected because they are identifiable; all cases would need to be treated as actual participants until it was verified that they were not genuine. All those not deemed to be eligible were emailed and provided with alternative sources of support. All those who had consented were emailed to ask whether they would be happy to be contacted when program dates became available. Of the 483 emails sent out only 1 reply was received. In terms of resource implications, this required additional time from study researchers as they had to go through numerous emails, ascertain whether it may be a fake participant, record this on a separate log, and reply to all emails. We were pleased that this had been identified early on before any nongenuine participants were randomized to the trial.

Case Study 4: Web-Based Survey to Assess the Acceptability of SPARX

What Happened?

This project assessed whether any changes needed to be made to SPARX [34], a game developed in New Zealand that delivers CBT to young people with low mood and anxiety, ahead of a pilot and feasibility RCT in the United Kingdom [35]. To do this, a Joint Information Systems Committee web-based survey was promoted via relevant networks and organizations, including social media. Our target was to reach 100-200, 11- to 19-year-olds. The survey was designed so that consent (for those older than 16 years) or consent (from a parent or legal carer) and assent (from those younger than 16 years) were initially obtained, which would direct young people to complete the study survey. After consent, on the final page of the survey, young people were directed to another web-based survey via a link where they could register for a prize draw. Questions were open-ended, asking the age of the responders, what they liked about playing computer games (question 8), and what they thought the advantages of e-therapy were (question 14). In total 415 young people gave consent; 391 participants completed the web-based survey and 372 registered for the prize draw.

How Did You Find Out?

The first suspicious activity noted was the speed of recruitment, receiving 299 responses to the survey in 1 day. On collating responses, it became clear that many participants (184/391, 47%) were not eligible (ie, being outside of the age of 11-19 years). Responses to question 8 were exact, case-sensitive duplicates of other responders, with 63.8% (132/391) answers being replicas and not answering the question. From this, only 28.5% (59/207) eligible responses were not noted as suspicious. A further 17 responses were noted as suspicious as when reviewing question 14. Many responses seen in question 8 were also made here and did not make sense in the context of the question asked. For example, the words “the Forest” appeared throughout the survey responses. From all responses to the survey, we were left with 10.7% (42/391) responses believed to be genuine.

How Did It Impact Your Research?

The process of cleaning the survey data was time-consuming and led to delays in completing the project. In addition, despite wanting to recruit more young people due to the team being unsure where the suspicious activity was coming from, the team was reluctant to continue pushing for recruitment. Furthermore, it has been difficult to decide whether these data should be published as there are no guarantees that the final included responses are genuine. Perhaps most importantly, given the reliance on this research in informing an upcoming RCT, caution was required in how the data could be applied to the study design. Finally, in reflecting on the issue, the experience brought about interesting feelings about one’s trust in the data that they have collected.

Case Study 5: Web-Based Interviews on Self-Harm in Transgender People

What Happened?

We recruited transgender people for a web-based interview study regarding factors for self-harm. The study offered a £10 (US $13) voucher for transgender views on an existing research tool and was advertised on social media (ie, Reddit), alongside other recruitment pathways. Once the study was advertised online, we received more than 50 responses expressing interest in 24 hours. The sheer number of respondents raised suspicions as respondent numbers far exceeded previous recruitment drives. At this stage, it was noted that all emails, without exception, were gmail.com accounts and followed the same pattern: first name, last name, and then a series of numbers. Also, some participants used terminology which, in our experience, is not used by the transgender community (ie, “I am a transgender”). These raised suspicions, but we did not want initial suspicions to result in genuine participants being missed, so follow-up emails were sent with study and consent details to all.

How Did You Find Out?

The nature of the study required potential participants to complete a well-being plan alongside standard informed consent. This entailed providing contact details of their general practitioner (GP: family doctor) and a trusted person and the address from which they will take part. People who regularly take part in self-harm research are familiar with ethical requirements and complete necessary forms in an appropriate manner. However, here respondents either did not complete the well-being plan or omitted necessary information. Some did complete well-being plans, but these raised further suspicions that these were fake participants. For example, GP surgery details included a mobile phone and a Gmail address, neither of which is typical for UK GP surgeries. Because of this, we performed cursory Google searches of both the GP and home addresses provided in the well-being plan to ensure that they were legitimate addresses. In all cases, it transpired that the addresses provided were either entirely false (ie, did not exist at all) or were commercial addresses. In one instance, for example, the “home address” was a hotel in central London and another was an industrial property. These addresses may be temporary residential properties or used during work hours to take part; however, they did not feel authentic. The false addresses, false or suspicious GP details, poorly completed well-being plans, suspicious terminology, and all respondents using Gmail accounts with similar email addresses left the researcher believing that these respondents were inauthentic, and they were not recruited to the study.

How Did It Impact Your Research (Financial and Time Costs)?

The process was disheartening, and the researcher spent hours responding to and identifying nongenuine responses, but early suspicions meant that the researcher avoided wasting significant time and resources interviewing and having discarded data. However, no genuine participants were recruited for this study, and we had to widen the study to the general population.

Case Study 6: Web-Based Focus Group on Inclusion at Work for Autistic Adults

What Happened?

We advertised for autistic participants to take part in a web-based focus group. The advert was posted by a project partner on Twitter. Approximately 150 potential participants responded within 24 hours, which was more than anticipated and more than we could invite to take part. The advert was removed, and the research team was suspicious that there were some nongenuine respondents. To tackle this, we asked respondents for demographic and diagnostic information and their motivation for being involved. We received 114 responses, the majority of which we believe were inauthentic. We invited 6 respondents to the focus group who we thought were most likely genuine participants.

We ran the web-based focus group with 6 participants for 1 hour. All participants turned on their cameras and provided verbal answers to the discussion questions. However, responses from 5 out of the 6 participants were general and provided little depth or personal perspective. When prompted, they did not expand their points or give specific examples. All 3 researchers present strongly suspected that 5 out of 6 participants did not have lived experience of autism. Although this was impossible to prove, the research team has more than 30 years of combined experience working closely with autistic adults and this experience led them to suspect that these participants were being insincere about their lived experience.

How Did You Find Out?

The large volume of volunteers in the short time window was the first indication that the initial recruitment phase had attracted nongenuine participants. This was confirmed by the request for further information as many motivation statements were duplicated between respondents, or were superficial, textbook-like responses. Like previous case studies, most respondents had Gmail addresses in the form of first name, last name, and a random number. Despite carefully selecting participants based on the quality of their responses and removing responses where inconsistent or suspicious information was provided, we were still not confident that our included participants were genuine.

How Did It Impact Your Research (Financial and Time Costs)?

Participants received a £30 (US $39) voucher, regardless of their contribution. We experienced an ethical dilemma about whether to include the data. Given we had no proof of inauthentic activity, we chose to analyze all the data and also collect and include additional data to identify converging themes from a wider range of people and experiences. Aside from potentially compromising our data, this experience was financially costly (paying additional participants), time-consuming (dealing with many emails, filtering and identifying genuine responses, and running and analyzing data from inauthentic focus groups), and had an emotional impact on the researchers.

Case Study 7: Exploring Patterns of Self-Harm in Autistic and Nonautistic Adults

What Happened?

We recruited autistic and nonautistic adults for a web-based interview study comparing proximal and distal risk markers for self-harm between the 2 groups, using the Card Sort Task for Self-Harm. Participants were offered a £10 (US $13) voucher for participating. We advertised the study through social media, charities, and volunteer research databases. Although many genuine participants signed up for our study, we became suspicious that a small group of participants recruited from 1 mental health research platform were not genuine participants.

How Did You Find Out?

Our study required participants to have an initial meeting with the researcher to complete a web-based demographics questionnaire to check eligibility, create a well-being plan, and obtain informed consent to take part. The well-being plan (identical to the one in case study 6) entailed providing multiple contact details and addresses. Several indicators raised our suspicions that some participants were not genuine. First, several indicators raised our suspicions that some participants were not interested in the research process with initial emails focusing on payment. Participants appeared to lack knowledge about the subject and could not provide details of their experiences after being probed. Second, there was evidence of participants attempting to register and take part in the study more than once such as refusing to turn their camera on. Despite this, it was clear to the researchers that the person had already taken part in the study. Furthermore, it was also clear that demographic data had been entered more than once, provided neighboring or fake addresses, fake phone numbers, and incomplete information (eg, no details of health care provider or trusted person to contact). Together, these indicated a small group of participants who were attempting to register and take part in the study more than once, presumably for financial gain.

How Did It Impact Your Research (Financial and Time Costs)?

The small group of participants who attempted to take part in the study more than once, and provided unreliable data, had to be excluded from the data set. It took a significant amount of time to be confident that our study had been infiltrated by nongenuine participants. By the time we had identified an issue and taken action, we had run out of time and funding to complete data collection for the nonautistic group. This meant that we had to revise the aims of the research to explore patterns of self-harm in the autistic group only—we were not able to make comparisons with nonautistic adults as we had planned. We had to submit a report explaining this to the funder on completion of the study.

The multiple case studies reflect a snippet of some of the institutional experiences within our health research teams. While we could not incorporate all the different case studies presented to us within our institution as they are far too many, these brief experiences summarize the many different ways that it can impact research project from multiple different methodological approaches.

Methodological or Research Challenges and Strategies

Overview

To minimize the impact of issues with web-based studies (qualitative or quantitative), it is essential to have a systematic process for “determining the level of suspicion required to remove potentially unreliable data” [36]. This includes many strategies that have been summarized in publications [36-38] such as a framework outlining the importance to Reflect, Expect, Analyze, and Label [36].

To think about the different challenges, it is important to think about these issues in relation to the whole research cycle and ask different questions at different stages of the research process [25]:

1. During funding application: What problems are likely to occur? What resources do I need to include to manage these problems?

2. Before recruitment: What is likely to happen? Can I develop a protocol that identifies and considers as many potential issues as possible?
3. During recruitment: Can I verify that the participant met the inclusion criteria? How confident am I in this information?
4. During data collection: Was the participant hesitant or flustered when asked probing questions for additional detail? What were my first impressions of honesty in my reflexive journal? Did I note any nonverbal cues that might be a clue to participant dishonesty?
5. During data analysis: Did I find places where the participants contradicted themselves? Were a participant's answers detailed enough that the participant seemed knowledgeable about the topic?

There are many noteworthy strategies put forward in this literature [36-38] and we bring together these challenges and strategies with our case studies to support applications for how one might implement them within their web-based research.

Pre–Data Collection

Pre–data collection refers to 2 different stages of the research cycle: ethics application and recruitment.

Challenges

The challenges experienced at the ethics application stage involve thinking in advance about what the potential threats to data integrity are. This will apply to threats throughout the research project and how the research team plans on dealing with these. In terms of recruitment, these threats could be multiple sign-ups from the same participants or nongenuine participants signing up (pretending to have specific health conditions or eligibility criteria). Some of the patterns that have been observed in detecting those include numerous entries very quickly, numerous consecutive entries with the same email format (name.surname999@gmail.com), entries from countries outside the recruitment area, similar IP addresses, refusal to provide a phone number or other key details, fast response to communication, or inconsistencies.

Strategies

We present some of the strategies our research team has put in place. It is important to note that some strategies might not be accepted by certain ethics committees and lived-experience groups and liaising with them is essential prior to recruitment.

Some useful strategies that we have put in place and found useful include the following:

1. *Protocol*: Creating a protocol before recruitment mitigating these issues as much as possible and where possible, coproduced with a lived-experience group to ensure the acceptability of the chosen strategies. To develop this protocol, it is useful to think about the following questions: How will you deal with inauthentic data? And how will I establish inauthentic data?
2. *Transparency*: Include all steps to be taken for nongenuine or multiple registration in the participant information sheet.
3. *Social media*: Close or strongly limit recruitment from particular social media platforms. Once your advert is on public platforms, you can no longer control how far and where it is distributed with many people sharing it, leading to nongenuine registrations. If advertising on social media, do not use terms such as "payment" or "gift cards" in adverts.
4. *Incentives*: One strategy is to not offer any incentives which will not be attractive to any nongenuine participants. Alternatively, offering "in-kind" incentives that would only be attractive to your target group can be helpful. For example, in a sleep study, participants were offered a sleep training course upon completion.
5. *Technical considerations*: Monitor responses from the same IP addresses. Installing cookies detector and CAPTCHA. Creating single-use links for each response. Having experts, for example, computer science, social media, and database experts as part of the team.

During Data Collection

Challenges

It is also important to check these issues once participants are recruited onto the study. Our team has experienced many challenges at this stage despite inputting strategies before data collection. Some of the challenges included people providing contradicting or inconsistent responses (between eligibility criteria and consent form), providing similar stories or responses to the study multiple times, vague answers or those who cannot elaborate when prompted, and shorter than average time in responses; in qualitative studies, refusing to put cameras on video calls, poor quality, or technical problems; and in quantitative studies, giving straight-line answers, high levels of nonresponse, answers that do not make sense, and empty free text boxes.

Strategies

Following recruitment but before data collection, a few strategies can be put in place.

1. *Interviews*: Asking participants to do a brief phone interview to check eligibility criteria before taking part. Providing a valid phone number as well as asking to briefly turn on the camera will often rebut nongenuine participants.
2. *Identity*: For participants with specific conditions, insider knowledge is useful—asking them to describe their lived experiences could help identify nongenuine participants. In addition, asking for a valid ID or documentation on the condition (for those who have received a diagnosis) can help minimize multiple and nongenuine responses.

It is also important to have strategies in place as data are collected.

1. *Data*: Keep checking data as they come through. Check for survey or interview duration, duplicate responses, or look for inconsistencies in responses. Keep an eye out for responses that do not "feel right." Roll out recruitment gradually to have time to check and stop any potential issues with the data collection process.
2. *Implement data screening*: Follow the data-screening plan and the protocol on reimbursement and data inclusion.

3. *Reflexive notes*: Keep track of your decision-making process and any challenges occurred at this stage for full transparency.

Post–Data Collection

Challenges

Once data collection is undertaken and multiple checks have been performed, it is also important to think about what will become of any data that you establish as nongenuine. For example, it is possible that participants sign up multiple times or do not complete certain questions, but these might be due to legitimate reasons. Issues around transparency in reporting the extent of data removed can arise in terms of how much to disclose in limited word count or potential concerns from reviewers.

Strategies

1. *Checking*: Contacting participants who responded multiple times or whose responses seem inconsistent to give them a chance to explain any personal difficulties with the study. Conducting regular debriefs with the research team to make decisions as a group and support the researchers' well-being.
2. *Incentives*: Aside from offering no incentives or in-kind incentives, check data before giving incentives, avoiding automated payments.
3. *Transparency*: Notify ethics or funder of issues with nongenuine participants and their impact on the study. Be transparent in reports or papers, establishing the ratio of this impact is important and should be disclosed in further publications.

It is important to note that while these strategies are often very useful, they also present challenges to researchers in terms of inclusivity, accessibility, or engagement. There are no foolproof strategies and researchers need to assess the benefits and limitations each provides.

Discussion

This publication reflects the significant impact and presence of issues with web-based studies by disclosing some of our experiences and the challenges faced with the strategies implemented. Primarily focused on mental health research, our experiences report a range of difficulties with repeated responders and nongenuine participants. Careful considerations and strategies are presented to help mitigate the threats these experiences can have on data integrity. These experiences and suggestions are given alongside those put forward in the literature, along the research cycle to give clear suggestions for consideration in future web-based research. We have shown that the issues around web-based data collection are broad and widely experienced, and our case studies reflect the increasing threat to data integrity. While we were not able to capture all the case studies in this publication, we know of at least 20 individual cases experienced on top of our reported studies.

It is important to note that a lot of the strategies presented have limitations. For example, tracking IP addresses can be easily changed by the use of a VPN which prevents the researcher from tracking genuine country of origin and it is also impossible to track through Gmail accounts [39]. In addition, it restricts members of the same households from joining, potentially excluding genuine participants. Other strategies can also be perceived as unethical or counterproductive. For instance, not recompensing participants for their time and effort [40] or assessing eligibility either in person or on the phone [41] might not be acceptable for certain participant groups and does not capitalize on the full benefits and appeal of web-based research. Some strategies are also not friendly to all groups. For example, while some conditions can be easily "confirmed" with a diagnostic letter, many mental health conditions can be trickier to demonstrate or are stigmatized conditions that may prevent people from getting involved (eg, anxiety and self-harm). Questioning participants' lived experiences through some form of justification might be extremely insensitive and unacceptable.

Many considerations must be taken when thinking about strategies through the research cycle. Some strategies can become quickly obsolete, for example, as technology evolves, new ways of bypassing existing strategies are always emerging. For example, the recently created platform of ChatGPT can easily replicate a believable experience of living with certain conditions. In addition, experienced users have been known to exchange tips on social media platforms on how to bypass study criteria and maximize rewards over engagement (eg, ProlificAC on Reddit).

In terms of the research cycle, it is important to remember that it impacts all areas of the research cycle (from grant application to writing up) and to think about all the different threats to data integrity as early as possible. Most strategies also require resources (time and money). Hence, appropriate resources must be requested in grants, or provided directly by institutions from central funding. Institutions might consider putting in place structural support for researchers doing web-based studies, such as through continuous professional development of staff that support studies, for example, research librarians, and research data managers.

In addition, we are unaware of any study yet that reports successfully addressing all issues of nongenuine participants, despite the efforts of developing strategies and seeking ethical approval. In a survey of fraud detection, Ballard et al [8] implemented 9 different strategies to detect and deter nongenuine responses and their final results demonstrated that only 38.9% (161/414) of responses were genuine.

These issues impact the quality and integrity of the data and add significant financial and time costs to a research project. In terms of financial cost, projects may waste funding on paying nongenuine participants. Timewise, hours and days are spent putting strategies into place and checking the data, which takes away essential time spent on other aspects of the project. It also has a cost in terms of mental well-being. We have found, as researchers, that spending time investigating sensitive health care topics to find out that participants have misrepresented their experiences, invalidating others' experiences, makes you doubtful of the data and is potentially heartbreaking. An agreed decision-making protocol can facilitate the process of including or excluding participants and reduce pressure on researchers to make decisions, which could impact study findings [38,42]. It is also important to note that issues affecting data integrity have a significant impact on participants as well as researchers. In our experience of conducting advisory groups and focus groups to identify key research topics, having even just 1 member in this group pretending to have a condition and listening to other's experiences can be very distressing, and navigating the facilitation of these situations requires high-level expertise. Especially if the extent of the fraud leads to nonpublication of results and essentially what could be perceived as a "waste of time."

Finally, as suggested by recent and insightful publications on the impact of these issues on specific groups, it is tricky but essential to strike a balance between data integrity and participants' vulnerability [38]. It is essential to maintain trust with participants, especially from potentially vulnerable backgrounds in the case of health research. Therefore, exclusion should be dealt with very carefully as while it is important to ensure that nongenuine participants' results are not included in the analysis to minimize its impact on results, it is also essential to give genuine participants the benefit of the doubt and not exclude genuine mistakes or difficulties in communications.

In conclusion, many important considerations need to be given throughout the project to minimize, as much as possible, the impact of multiple responses, bots, and nongenuine participants on our data. This is not easy and while many strategies exist and are useful, their efficacy highly depends on the project's methodology and population of choice. Careful considerations need to be taken when implementing these strategies, ensuring that they are acceptable and feasible within the remits of each project. This is not a quick process and involves time and resources. However, these are essential in conducting web-based studies as without these checks, it is very unlikely that the data collected will be reliable and representative. Reporting how these risks have been mitigated should become compulsory in upcoming grants and publications to ensure data integrity and credibility for web-based research.

Acknowledgments

This work and Blandine French were supported by an ESRC grant (ES/X000141/1). Stefan Rennick-Egglestone was supported by the National Institute for Health and Care Research (NIHR), through the Narrative Experiences Online (NEON) program (RP-PG-0615-20016) and the NIHR Nottingham Biomedical Research Centre (NIHR203310). Katherine Bird is supported by a grant from Mental Health Research UK. Camilla Babbage acknowledges funding support from the NIHR MindTech Medtech Cooperative, the NIHR Nottingham Biomedical Research Centre, and the UK Research and Innovation (UKRI) Digital Youth Programme award which was part of the AHRC/ESRC/MRC Adolescence, Mental Health and the Developing Mind program (MRC project reference: MR/W002450/1 and MR/T046864/1). The views expressed here are those of the authors and not necessarily those of the NIHR or the UK Department of Health and Social Care. Mirabel Pelton is supported by a grant from Autism Centre of Excellence at Cambridge (grant 124306).

Conflicts of Interest

BF reports personal fees and nonfinancial support from Takeda and Medice. All other authors report no conflict of interest.

Multimedia Appendix 1
Repeat registration procedures.
[DOCX File , 42 KB - mental_v11i1e58432_app1.docx]

Multimedia Appendix 2
Repeat registration report.
[DOCX File , 27 KB - mental_v11i1e58432_app2.docx]

References

1. Hlatshwako TG, Shah SJ, Kosana P, Adebayo E, Hendriks J, Larsson EC, et al. Online health survey research during COVID-19. Lancet Digit Health 2021;3(2):e76-e77 [FREE Full text] [doi: 10.1016/S2589-7500(21)00002-9] [Medline: 33509387]
2. De Man J, Campbell L, Tabana H, Wouters E. The pandemic of online research in times of COVID-19. BMJ Open 2021;11(2):e043866 [FREE Full text] [doi: 10.1136/bmjopen-2020-043866] [Medline: 33622948]
3. Jones A, Caes L, Rugg T, Noel M, Bateman S, Jordan A. Challenging issues of integrity and identity of participants in non-synchronous online qualitative methods. Methods Psychol 2021;5:100072. [doi: 10.1016/j.metip.2021.100072]
4. Kramer J, Rubin A, Coster W, Helmuth E, Hermos J, Rosenbloom D, et al. Strategies to address participant misrepresentation for eligibility in Web-based research. Int J Methods Psychiatr Res 2014;23(1):120-129 [FREE Full text] [doi: 10.1002/mpr.1415] [Medline: 24431134]
5. Routen A, Bambra C, Willis A, Khunti K. Hard to reach? Language matters when describing populations underserved by health and social care research. Public Health 2022;205:e28-e29. [doi: 10.1016/j.puhe.2022.02.006] [Medline: 35282902]

6. Eysenbach G, Wyatt J. Using the Internet for surveys and health research. J Med Internet Res 2002;4(2):E13 [FREE Full text] [doi: 10.2196/jmir.4.2.e13] [Medline: 12554560]
7. Lefever S, Dal M, Matthíasdóttir Á. Online data collection in academic research: advantages and limitations. Br J Educ Technol 2007;38(4):574-582 [FREE Full text] [doi: 10.1111/j.1467-8535.2006.00638.x]
8. Ballard AM, Cardwell T, Young AM. Fraud detection protocol for web-based research among men who have sex with men: development and descriptive evaluation. JMIR Public Health Surveill 2019;5(1):e12344 [FREE Full text] [doi: 10.2196/12344] [Medline: 30714944]
9. Gagné N, Franzen L. How to run behavioural experiments online: best practice suggestions for cognitive psychology and neuroscience. Ubiquity Press 2003;3(1):1. [doi: 10.5334/spo.34]
10. The British Psychological Society. Ethics guidelines for internet-mediated research. URL: https://www.bps.org.uk/guideline/ethics-guidelines-internet-mediated-research [accessed 2024-02-23]
11. Mortensen K, Hughes TL. Comparing Amazon's mechanical turk platform to conventional data collection methods in the health and medical research literature. J Gen Intern Med 2018;33(4):533-538 [FREE Full text] [doi: 10.1007/s11606-017-4246-0] [Medline: 29302882]
12. Peer E, Rothschild D, Gordon A, Evernden Z, Damer E. Data quality of platforms and panels for online behavioral research. Behav Res Methods 2022;54(4):1643-1662 [FREE Full text] [doi: 10.3758/s13428-021-01694-3] [Medline: 34590289]
13. Yarrish C, Groshon L, Mitchell J, Appelbaum A, Klock S, Winternitz T, et al. Finding the signal in the noise: minimizing responses from bots and inattentive humans in online research. Behavior Therapist 2019;42(7):235-242.
14. Buhrmester M, Kwang T, Gosling SD. Amazon's mechanical Turk: a new source of inexpensive, yet high-quality, data? Perspect Psychol Sci 2011;6(1):3-5. [doi: 10.1177/1745691610393980] [Medline: 26162106]
15. ScienceDirect. Prolific.ac—a subject pool for online experiments. URL: https://www.sciencedirect.com/science/article/pii/S2214635017300989?via%3Dihub [accessed 2024-01-15]
16. Ysidron DW, France CR, Yang Y, Mischkowski D. Research participants recruited using online labor markets may feign medical conditions and overreport symptoms: caveat emptor. J Psychosom Res 2022;159:110948. [doi: 10.1016/j.jpsychores.2022.110948] [Medline: 35644084]
17. Osborne JW, Blanchard MR. Random responding from participants is a threat to the validity of social science research results. Front Psychol 2010;1:220 [FREE Full text] [doi: 10.3389/fpsyg.2010.00220] [Medline: 21833275]
18. Dennis SA, Goodson BM, Pearson CA. Online worker fraud and evolving threats to the integrity of MTurk data: a discussion of virtual private servers and the limitations of IP-based screening procedures. Behav Res Account 2020;32(1):119-134. [doi: 10.2308/bria-18-044]
19. Chandler JJ, Paolacci G. Lie for a dime: when most prescreening responses are honest but most study participants are impostors. 2017. URL: https://journals.sagepub.com/doi/full/10.1177/1948550617698203 [accessed 2024-01-15]
20. Hydock C. Assessing and overcoming participant dishonesty in online data collection. Behav Res Methods 2018;50(4):1563-1567. [doi: 10.3758/s13428-017-0984-5] [Medline: 29274007]
21. Murray E, Khadjesari Z, White IR, Kalaitzaki E, Godfrey C, McCambridge J, et al. Methodological challenges in online trials. J Med Internet Res 2009;11(2):e9 [FREE Full text] [doi: 10.2196/jmir.1052] [Medline: 19403465]
22. Fronek P, Briggs L. Faking participant identity: vested interests and purposeful interference. 2018. URL: https://journals.sagepub.com/doi/full/10.1177/1747016117740177 [accessed 2024-01-15]
23. Bowen AM, Daniel CM, Williams ML, Baird GL. Identifying multiple submissions in Internet research: preserving data integrity. AIDS Behav 2008;12(6):964-973 [FREE Full text] [doi: 10.1007/s10461-007-9352-2] [Medline: 18240015]
24. ProQuest. Imposter participants: overcoming methodological challenges related to balancing participant privacy with data quality when using online recruitment and data collection. URL: https://www.proquest.com/docview/2737499952?pq-origsite=gscholar&fromopenview=true&sourcetype=Scholarly%20Journals [accessed 2024-01-15]
25. Ridge D, Bullock L, Causer H, Fisher T, Hider S, Kingstone T, et al. 'Imposter participants' in online qualitative research, a new and increasing threat to data integrity? Health Expect 2023;26(3):941-944 [FREE Full text] [doi: 10.1111/hex.13724] [Medline: 36797816]
26. Pellicano E, Adams D, Crane L, Hollingue C, Allen C, Almendinger K, et al. Letter to the editor: a possible threat to data integrity for online qualitative autism research. Autism. 2023. URL: https://journals.sagepub.com/doi/full/10.1177/13623613231174543 [accessed 2024-01-15]
27. Salinas MR. Are your participants real? Dealing with fraud in recruiting older adults online. West J Nurs Res 2023;45(1):93-99. [doi: 10.1177/01939459221098468] [Medline: 35587721]
28. Pozzar R, Hammer MJ, Underhill-Blazey M, Wright AA, Tulsky JA, Hong F, et al. Threats of bots and other bad actors to data quality following research participant recruitment through social media: cross-sectional questionnaire. J Med Internet Res 2020;22(10):e23021 [FREE Full text] [doi: 10.2196/23021] [Medline: 33026360]
29. Chandler J, Sisso I, Shapiro D. Participant carelessness and fraud: consequences for clinical research and potential solutions. J Abnorm Psychol 2020;129(1):49-55. [doi: 10.1037/abn0000479] [Medline: 31868387]
30. Strickland JC, Stoops WW. Utilizing content-knowledge questionnaires to assess study eligibility and detect deceptive responding. Am J Drug Alcohol Abuse 2020;46(2):149-157. [doi: 10.1080/00952990.2019.1689990] [Medline: 31810399]

31. Rennick-Egglestone S, Elliott R, Smuk M, Robinson C, Bailey S, Smith R, et al. Impact of receiving recorded mental health recovery narratives on quality of life in people experiencing psychosis, people experiencing other mental health problems and for informal carers: Narrative Experiences Online (NEON) study protocol for three randomised controlled trials. Trials 2020;21(1):661 [FREE Full text] [doi: 10.1186/s13063-020-04428-6] [Medline: 32690105]
32. Robinson C, Newby C, Rennick-Egglestone S, Llewellyn-Beardsley J, Ng F, Elliott RA, et al. Statistical analysis plans for two randomised controlled trials of the Narrative Experiences Online (NEON) Intervention: impact of receiving recorded mental health recovery narratives on quality of life in people experiencing psychosis (NEON) and people experiencing non-psychosis mental health problems (NEON-O). Trials 2023;24(1):343 [FREE Full text] [doi: 10.1186/s13063-023-07246-8] [Medline: 37210551]
33. Slade M, Rennick-Egglestone S, Elliott R, Newby C, Robinson C, Gavan SP, NEON study group. Effectiveness and cost-effectiveness of online recorded recovery narratives in improving quality of life for people with non-psychotic mental health problems: a pragmatic randomized controlled trial. World Psychiatry 2024;23(1):101-112 [FREE Full text] [doi: 10.1002/wps.21176] [Medline: 38214639]
34. Merry SN, Stasiak K, Shepherd M, Frampton C, Fleming T, Lucassen MFG. The effectiveness of SPARX, a computerised self help intervention for adolescents seeking help for depression: randomised controlled non-inferiority trial. BMJ 2012;344:e2598 [FREE Full text] [doi: 10.1136/bmj.e2598] [Medline: 22517917]
35. Health Research Authority. SPARX-UK: pilot and feasibility trial. URL: https://tinyurl.com/3jpb294j [accessed 2024-02-23]
36. Lawlor J, Thomas C, Guhin AT, Kenyon K, Lerner MD, Drahota A. Suspicious and fraudulent online survey participation: introducing the REAL framework. Methodological Innov 2021;14(3):205979912110504. [doi: 10.1177/20597991211050467]
37. Teitcher JEF, Bockting WO, Bauermeister JA, Hoefer CJ, Miner MH, Klitzman RL. Detecting, preventing, and responding to "fraudsters" in internet research: ethics and tradeoffs. J Law Med Ethics 2015;43(1):116-133 [FREE Full text] [doi: 10.1111/jlme.12200] [Medline: 25846043]
38. Sansfaçon AP, Gravel E, Gelly MA. Dealing with scam in online qualitative research strategies and ethical considerations. 2024. URL: https://tinyurl.com/4u7s3c6c [accessed 2024-02-08]
39. Liu E, Sun L, Bellon A, Voelker G, Savage S, Munyaka I. Understanding the viability of Gmail's origin indicator for identifying the sender. URL: https://tinyurl.com/mjp8h248 [accessed 2024-08-17]
40. Wilson PM, Petticrew M, Calnan M, Nazareth I. Effects of a financial incentive on health researchers' response to an online survey: a randomized controlled trial. J Med Internet Res 2010;12(2):e13 [FREE Full text] [doi: 10.2196/jmir.1251] [Medline: 20457556]
41. Bromberg J, Wood ME, Black RA, Surette DA, Zacharoff KL, Chiauzzi EJ. A randomized trial of a web-based intervention to improve migraine self-management and coping. Headache 2012;52(2):244-261 [FREE Full text] [doi: 10.1111/j.1526-4610.2011.02031.x] [Medline: 22413151]
42. Bauermeister JA, Pingel E, Zimmerman M, Couper M, Carballo-Diéguez A, Strecher VJ. Data quality in web-based HIV/AIDS research: handling invalid and suspicious data. Field methods 2012;24(3):272-291 [FREE Full text] [doi: 10.1177/1525822X12443097] [Medline: 23180978]

Abbreviations

ADHD: attention-deficit/hyperactive disorder
CBT: cognitive behavior therapy
GP: general practitioner
RCT: randomized controlled trial
REDCap: Research Electronic Data Capture

Edited by J Torous; submitted 15.03.24; peer-reviewed by N Josifovski, S Gould; comments to author 15.06.24; revised version received 03.07.24; accepted 13.07.24; published 16.09.24.

Please cite as:
French B, Babbage C, Bird K, Marsh L, Pelton M, Patel S, Cassidy S, Rennick-Egglestone S
Data Integrity Issues With Web-Based Studies: An Institutional Example of a Widespread Challenge
JMIR Ment Health 2024;11:e58432
URL: https://mental.jmir.org/2024/1/e58432
doi:10.2196/58432
PMID:

Original Paper

The Implementation of Recommender Systems for Mental Health Recovery Narratives: Evaluation of Use and Performance

Emily Slade[1,2], BA; Stefan Rennick-Egglestone[1], PhD; Fiona Ng[1], PhD; Yasuhiro Kotera[1,3], PhD; Joy Llewellyn-Beardsley[1], PhD; Chris Newby[4], PhD; Tony Glover[5], PhD; Jeroen Keppens[6], PhD; Mike Slade[1,7], PhD

[1]School of Health Sciences, Institute of Mental Health, University of Nottingham, Nottingham, United Kingdom
[2]Department of Computer Science, University of Oxford, Oxford, United Kingdom
[3]Center for Infectious Disease Education and Research, Osaka University, Osaka, Japan
[4]School of Medicine, University of Nottingham, Nottingham, United Kingdom
[5]DRT Software Ltd., Nottingham, United Kingdom
[6]Department of Informatics, King's College London, London, United Kingdom
[7]Nord University, Faculty of Nursing and Health Sciences, Namsos, Norway

Corresponding Author:
Emily Slade, BA
School of Health Sciences
Institute of Mental Health
University of Nottingham
University of Nottingham Innovation Park
Triumph Road
Nottingham, NG7 2TU
United Kingdom
Phone: 44 0115 823 0812
Email: emily_slade@outlook.com

Abstract

Background: Recommender systems help narrow down a large range of items to a smaller, personalized set. NarraGive is a first-in-field hybrid recommender system for mental health recovery narratives, recommending narratives based on their content and narrator characteristics (using content-based filtering) and on narratives beneficially impacting other similar users (using collaborative filtering). NarraGive is integrated into the Narrative Experiences Online (NEON) intervention, a web application providing access to the NEON Collection of recovery narratives.

Objective: This study aims to analyze the 3 recommender system algorithms used in NarraGive to inform future interventions using recommender systems for lived experience narratives.

Methods: Using a recently published framework for evaluating recommender systems to structure the analysis, we compared the content-based filtering algorithm and collaborative filtering algorithms by evaluating the accuracy (how close the predicted ratings are to the true ratings), precision (the proportion of the recommended narratives that are relevant), diversity (how diverse the recommended narratives are), coverage (the proportion of all available narratives that can be recommended), and unfairness (whether the algorithms produce less accurate predictions for disadvantaged participants) across gender and ethnicity. We used data from all participants in 2 parallel-group, waitlist control clinical trials of the NEON intervention (NEON trial: N=739; NEON for other [eg, nonpsychosis] mental health problems [NEON-O] trial: N=1023). Both trials included people with self-reported mental health problems who had and had not used statutory mental health services. In addition, NEON trial participants had experienced self-reported psychosis in the previous 5 years. Our evaluation used a database of Likert-scale narrative ratings provided by trial participants in response to validated narrative feedback questions.

Results: Participants from the NEON and NEON-O trials provided 2288 and 1896 narrative ratings, respectively. Each rated narrative had a median of 3 ratings and 2 ratings, respectively. For the NEON trial, the content-based filtering algorithm performed better for coverage; the collaborative filtering algorithms performed better for accuracy, diversity, and unfairness across both gender and ethnicity; and neither algorithm performed better for precision. For the NEON-O trial, the content-based filtering algorithm did not perform better on any metric; the collaborative filtering algorithms performed better on accuracy and unfairness across both gender and ethnicity; and neither algorithm performed better for precision, diversity, or coverage.

Conclusions: Clinical population may be associated with recommender system performance. Recommender systems are susceptible to a wide range of undesirable biases. Approaches to mitigating these include providing enough initial data for the recommender system (to prevent overfitting), ensuring that items can be accessed outside the recommender system (to prevent a feedback loop between accessed items and recommended items), and encouraging participants to provide feedback on every narrative they interact with (to prevent participants from only providing feedback when they have strong opinions).

(JMIR Ment Health 2024;11:e45754) doi:10.2196/45754

KEYWORDS

recommender system; mean absolute error; precision; intralist diversity; item space coverage; fairness across users; psychosis; Narrative Experiences Online trial; NEON trial; lived experience narrative; recovery story

Introduction

Background

Recommender systems create personalized recommendations within a specific domain, suggesting items that may be of use to a user and helping quickly narrow down a potentially overwhelming number of options [1]. Recommender systems are used on global platforms such as Netflix—a movie streaming service—which uses other people's movie ratings to recommend movies, Amazon—an e-commerce company—which uses frequently-bought-together items to recommend purchases, and Pandora—a music streaming service—which uses 450 musical attributes to recommend songs [2].

A range of health care applications for recommender systems have been examined, including the use of recommender systems to suggest prompts for counselors in a suicide prevention helpline chat [3], tailor care preference assessments in nursing homes [4], and identify expert physicians for specific diseases [5].

In this paper, we present an evaluation of NarraGive, the first recommender system for providing web-based recommendations from a collection of mental health recovery narratives.

Lived Experience Narratives

Mental health recovery narratives are a subset of lived experience narratives, which are representations of a person's experiences of physical or mental health and how that person has lived through and responded to those experiences [6]. The uses of lived experience narratives in health research have been extensively studied but with little focus on *which* narratives people engage with.

Studies have explored the use of lived experiences to encourage people to seek and sustain treatment, such as using narratives to improve health care participation in patients with breast cancer [7], promote smoking cessation in the African American community [8], and promote diabetes self-management [9] and diabetes medication adherence [10]. The use of lived experiences in support groups has also been studied, such as sharing stories in diabetes education in minority ethnic groups [11]. Some studies have provided medical students with narratives to facilitate learning and improve subsequent medical practices, such as using patient stories during practice placements [12] and learning about cancer pathology using narratives of patients who have experienced cancer [13].

Other studies have explored the use of lived experiences as a therapeutic tool for individuals, such as student nurses creating digital stories to challenge the "reality shock" of beginning clinical practice [14], young women telling their stories to reduce stress [15], women with eating disorders accessing recovery stories [16], service users with psychosis watching lived experience videos [17], incarcerated women telling their stories [18], patients with dementia using storytelling as a therapeutic tool [19], adults with diabetes engaging in lived experience support groups to reduce diabetes-related distress [20], painting trees to symbolize periods of one's life as a starting point for telling a life story to treat depression and anxiety [21], and young people watching digital stories to reduce the prevalence of binge drinking [22].

Lived experience narratives have the potential to be used for a wide variety of purposes and, as a result—as documented previously—are frequently used in interventions. However, so far, the focus of health lived experience–based interventions has been solely on examining the effects of engaging with these narratives, with less focus on which *specific* narratives the participants are exposed to (though a few studies have placed emphasis on providing representative narratives [23] or particularly engaging and high-quality narratives [8]). Thus, while there have been studies evaluating the use of recommender systems in health care settings and, separately, evaluating the use of lived experience narratives, there have not been any lived experience narrative recommender systems developed before this study.

The Problem Being Addressed

This is the first evaluation of a lived experience narrative recommender system. The design of such a recommender system has distinct challenges. For example, narratives are sensitive types of data that impose ethical requirements to protect both the narrator and the recipient. Therefore, the use of recommender systems needs to be informed by considerations about the curation and use of narratives [24-26]. The goal of our evaluation was to develop preliminary evidence to inform the future use and evaluation of recommender systems with lived experience narratives.

The Narrative Experiences Online Intervention

Overview

The Narrative Experiences Online (NEON) study [27,28] evaluated whether having web-based access to people's real-life stories of recovery from mental illness can be helpful for people who are experiencing psychosis or other mental health problems. This builds on the evidence base that indicates that receiving recovery narratives can support mental health [27]. In the NEON intervention, participants interact with a web application through which they can access a web-based collection of mental health recovery narratives (henceforth, narratives)—the NEON Collection.

Narrative Characterization

The development of the NEON Collection, including the narrative inclusion criteria, has been reported elsewhere [29]. In brief, recorded recovery narratives were obtained, always with consent, from existing collections and individual donations to the study. Only narratives that could be presented on the web in a single electronic file (eg, PDF, JPEG, and WAV) were included. Within these files, narratives were presented in a range of forms, including prose, poetry, audio recordings, video recordings, individual images, and sequential art. Each was presented by a single narrator only—there were no composite narratives. The narratives were deliberately chosen to be diverse [30]. All narratives in the NEON Collection were characterized using the Inventory of Characteristics of Recovery Stories (INCRESE) [31] to capture 77 different features of the narratives related to narrator characteristics, narrative content, and turning points. While we used selected INCRESE characteristics in our recommender system, the greater breadth of characteristics collected will support future secondary analyses. The trials opened with 348 narratives and closed with 659 narratives available.

Narrative Request Routes

There are 6 ways for participants to request narratives through the NEON intervention, which are internally documented as 1 of 8 request methods.

Textbox 1 summarizes the external and internal narrative request routes.

The NEON intervention home page has buttons corresponding to 4 of the 6 external narrative request routes: “Match me to a story (recommended),” “Get me a random story,” “Browse stories,” and “My stories.”

The first option uses NarraGive to recommend a single narrative that the participant has not seen before. NarraGive is a hybrid recommender system (meaning that it uses a combination of recommendation strategies [32]) that uses both content-based filtering (recommending narratives based on their content) and collaborative filtering (recommending narratives based on how other participants have rated them) to recommend narratives to participants.

The second option presents a randomly selected narrative that the participant has not seen before.

The third option allows participants to browse narratives grouped into categories based on the narratives’ INCRESE characteristics (Figures S1 and S2 in Multimedia Appendix 1)—some categories are based on the value of a single characteristic (eg, the narrator’s gender is “female”), and some are based on the value of multiple characteristics (eg, a positive narrative, defined as having an “upbeat” tone and an “escape” or “enlightenment” genre; Table S1 in Multimedia Appendix 1). Not all narratives are accessible through the category option.

The fourth option allows participants to access narratives that they have previously bookmarked or rated highly.

In addition, the internal request routes include whether NarraGive produced the recommendation using content-based filtering or collaborative filtering and whether a narrative selected from the “My stories” page was previously rated highly for hopefulness or manually bookmarked by the participant. One important benefit of having different narrative request routes is to prevent exposure bias, a well-known issue in recommender systems where participants are only presented with a subset of the available items, so they only provide ratings for that subset, with recommender systems unable to distinguish between disliked and unrated items and unknown and unrated items [33]. For example, the “Get me a random story” button might allow participants to access narratives that they would not otherwise be exposed to but that nonetheless may be beneficial.

Textbox 1. Narrative request mechanisms that participants use to access narratives (external routes) and the corresponding logs made by the intervention (internal routes).

External and internal narrative request routes

- Participant clicks on the "Match me to a story (recommended)" button
  - Participant accesses a narrative recommended via content-based filtering.
  - Participant accesses a narrative recommended via collaborative filtering.
- Participant clicks on the "Get me a random story" button
  - Participant requests a random narrative.
- Participant clicks on the "Browse stories" button and selects a narrative
  - Participant makes a category-based request for a narrative.
- Participant clicks on the "My stories" button and selects a narrative
  - Participant requests a narrative that they have rated as hopeful.
  - Participant requests a narrative that they have marked as a favorite.
- Participant uses the intervention for the first time and is presented with their first narrative
  - Participant accesses their "first" narrative.
- Participant clicks on a narrative from a Narrative Experiences Online (NEON) communication
  - Participant accesses the suggested narrative in a reminder message aimed at prompting them to use the NEON intervention.

Narrative Feedback

After a participant has accessed a narrative through any request route, they are presented with 5 feedback questions (Table 1), and their responses to these questions are time-stamped and logged. The focus (hope, similarity, learning, and empathy) is based on the NEON Impact Model [29] developed through a systematic review [34] and qualitative [35] and experimental studies [36]. The measurement approach has been previously validated [29]. To maximize response rates, the first question is marked as compulsory. The other 4 questions are marked as optional, and the participant has the choice to answer either all or none of the optional questions. A set of 5 response values (for the 1 compulsory and 4 optional questions) forms a single rating, as does a single response value for the compulsory question. Ratings with optional questions answered are also referred to as optional ratings. Table 1 shows the questions, answer options, and numerical ranges (not visible to participants) of the questions and whether they are mandatory.

If a narrative is rerated, this overrides the previous rating (but the time-stamped logs of previous ratings are not deleted).

One benefit of recommender systems *requiring* a rating for each narrative is that this helps minimize selection bias, which occurs when participants are allowed to choose whether to rate the items, leading to ratings that are typically biased toward higher or more homogeneous ratings [33,37]. Selection bias is a well-known problem in recommender systems relying on explicit data.

Table 1. Questions, answer options, numerical ranges, and mandatory nature of narrative response data.

Question	Answer options	Range	Mandatory
How hopeful did the story leave you feeling? (hopeful-ness)	"Less hopeful than before," "no change," "a bit more hopeful," and "much more hopeful"	−1 to 2	Yes
How similar was the storyteller to you? (similarity to the narrator)	"Not at all," "a bit," "quite a lot," and "very much"	0 to 3	No
How similar was the storyteller's life to your life? (similarity to the narrative)	"Not at all," "a bit," "quite a lot," and "very much"	0 to 3	No
How much did you learn from the story? (learning)	"Not at all," "a bit," "quite a lot," and "very much"	0 to 3	No
How emotionally connected did you feel with the story? (empathy)	"Not at all," "a bit," "quite a lot," and "very much"	0 to 3	No

The NarraGive Recommender System

NarraGive is a hybrid recommender system. It uses one content-based and 2 collaborative filtering algorithms to allow for comparison of performance of the 3 algorithms using 2 distinct approaches to inform this new field of lived experience narrative recommendation. NarraGive was assembled using the Simple Python Recommendation System Engine library (SurPRISE; version 1.1.1; Nicolas Hug) for Python (version 3.6 and above), integrating implementations of filtering algorithms provided in these libraries [38]. NarraGive does not recommend previously requested narratives, types of narratives that a user has previously blocked, or individual narratives that a user has blocked.

The content-based filtering algorithm is based on the SurPRISE implementation of the k-nearest neighbor (kNN) algorithm. Although kNN is traditionally used as a collaborative filtering algorithm, NarraGive used an adapted version to measure the similarity between narratives, in which it uses their INCRESE characteristics to cluster together narratives in "neighborhoods" and recommend to participants unseen narratives that are similar to their other highly rated narratives. Narrative similarity is assessed using selected INCRESE characteristics, consisting of the INCRESE sections on narrator characteristics, narrative characteristics, narrative content, and turning points.

The selected collaborative filtering algorithms are the SurPRISE implementations of the singular value decomposition (SVD) and, to support comparison, SVD++. A broad introduction to these 2 algorithms is provided in the work by Hug [39]. These aim to capture the latent factors that determine how much a participant likes a narrative. NarraGive ran these 2 algorithms and selected the narrative with the highest predicted rating. Thus, the 2 algorithms served as distinct subsystems, so this evaluation will analyze the 2 subsystems separately to compare them. For the purposes of collaborative filtering, similarity between users is assessed using the demographic items collected in a "personal profile" created at first use and containing items describing participant demographics and format preferences. Multimedia Appendix 2 provides details on all items in the profile.

When making a narrative recommendation, narrative feedback ratings are weighted (with a hopefulness rating twice as influential as each of the individual optional ratings) and combined. This was due to the underlying theory that we developed on narratives making an impact on recipients, which emphasized hope creation as the most critical mechanism. When a participant requests a narrative from NarraGive, it internally generates 1 list per algorithm of the 10 narratives with the highest predicted rating. It then presents the highest-scoring narrative of these 30 to the participant. The participant is not shown the predicted rating, other internally generated narratives, or which of the 2 filtering mechanisms was used to generate the recommendation.

The NEON Trials

The NEON intervention has been evaluated in 3 pragmatic randomized controlled trials with different populations. The NEON trial (ISRCTN11152837; N=739) is a definitive trial for people with experience of psychosis. The NEON for other (eg, nonpsychosis) mental health problems (NEON-O) trial (ISRCTN63197153; N=1023) is a definitive trial for people experiencing any other type of mental health problem. The NEON-C trial (ISRCTN76355273; N=54) is a feasibility trial with people who informally care for people experiencing mental health problems, which is not within the scope of this study. The NEON intervention was identical in all 3 trials. A separate instance of NarraGive was used for each trial, and there was no pooling of narrative feedback or recommendations among the 3 trials.

Aims and Objectives

The aim of this study was to analyze the 3 recommender system algorithms used in NarraGive to inform future interventions using recommender systems in this new field of lived experience narrative recommendations. An evaluation of the impact of the NEON intervention using NarraGive has been reported elsewhere [40]. This study did not aim to provide an indication of NarraGive's viability but rather to inform the development of future lived experience narrative recommender systems and guide design choices on collaborative versus content-based filtering algorithms.

The objectives were as follows:

1. To describe participant characteristics and patterns of narrative requests and feedback.
2. To evaluate the algorithms used in NarraGive by comparing collaborative-based and content-based narrative recommendations to inform future implementation approaches.

Objective 1 was addressed using data from the intervention version of NarraGive, and objective 2 was addressed using data from the final evaluated version.

Methods

Overview

An evaluation of NarraGive was conducted using data from the NEON and NEON-O trials, structured using the framework for evaluating recommender systems (FEVR), which was developed through a review of recommender system evaluation work [41]. The FEVR defines a set of components intended to guide the design of a recommender system evaluation.

After the NEON trials closed, logging files describing interactions with trial procedures and the NEON intervention were downloaded for analysis. These files included trial allocation, baseline demographic characteristics, personal profiles, narrative characteristics, narratives that the participants requested and the corresponding internal narrative request route, and participants' ratings. All log entries were time-stamped.

Ethical Considerations

The NEON study trial protocol and an update have been published elsewhere [27,28]. Ethics approval was obtained in advance of trial start from a UK National Health Service Research Ethics Committee (Leicester Central Research Ethics Committee; 19/EM/0326). All participants provided web-based informed consent for the use of their data for research purposes, and all study data were pseudonymous, with each participant's data linked by a unique ID. Some participants were compensated (£20 [US $25.59] vouchers) for some data collection rounds, as described in our trial protocol.

Participants

The NEON trial included participants who (1) had experience of psychosis in the previous 5 years, (2) had experience of mental health–related distress in the previous 6 months, (3) resided in England, (4) were aged ≥18 years, (5) were capable of accessing or being supported to access the internet on a PC or mobile device or at a community venue, (6) were able to understand written and spoken English, and (7) were capable of providing web-based informed consent.

The NEON-O trial included participants who (1) had experience of mental health problems other than psychosis in the previous 5 years, (2) had experience of mental health–related distress in the previous 6 months, (3) resided in England, (4) were aged ≥18 years, (5) were capable of accessing or being supported to access the internet on a PC or mobile device or at a community venue, (6) were able to understand written and spoken English, and (7) were capable of providing web-based informed consent. It excluded participants eligible for the NEON trial.

Our study included participants from the NEON trials' intention-to-treat samples [27].

Sample Size

Both trials were powered on the mean item score for the 12 subjective items in the Manchester Short Assessment of Quality of Life (MANSA) as collected at baseline and the 52-week follow-up [42], and hence, the sample size was chosen on this basis.

For the NEON trial, a total sample size of 684 was chosen to provide 90% power to detect a minimal clinically important effect size (Cohen *d*) of 0.27 (SD 0.9 [43]; power=90%; *P*=.05), allowing for 20% attrition. The planned analyzable sample size was 546.

For the NEON-O trial, the SD of the MANSA scores for the study population was estimated from baseline data provided by the first 350 enrolled participants (see the study by Rennick-Egglestone et al [27] for the rationale). A total sample size of 994 was selected to provide 90% power to detect a minimal clinically important effect size (Cohen *d*) of 0.27 (SD 0.94; power=90%; *P*=.05), allowing for 40% attrition, which was estimated from the completion rates for interim data. The planned analyzable sample size was 596.

Both trials recruited their planned samples and were allowed to overrecruit (N=739 for the NEON trial and N=1023 for the NEON-O trial). The final attrition rates were 23.5% (NEON trial) and 44.8% (NEON-O trial).

Evaluation Framework

Table 2 describes the FEVR components that were selected to define the evaluation.

Table 2. Framework for evaluating recommender systems (FEVR) components defining the NarraGive evaluation.

FEVR component	Brief description
Evaluation objectives	
Overall goal	• To evaluate whether the recommender system NarraGive supported participants in finding helpful narratives
Stakeholders	• Participants in the NEON[a] and NEON-O[b] trials' ITT[c] samples
Properties	• Prediction accuracy, usage prediction, diversity, coverage, and unfairness across participants
Evaluation principles	
Hypothesis or research question	• Objective 1: To describe participant characteristics and patterns of narrative requests and feedback • Objective 2: To evaluate the NarraGive recommender system by comparing collaborative-based and content-based narrative recommendations
Control variables	• Randomized data set that is split 75:25 between the training set (to train the algorithms) and the testing set (to evaluate the metrics)
Generalization power	• Use of real-world data from participants with mental health problems; limited due to variation in system use
Reliability	• Cross-validation with repeated initialization of collaborative filtering algorithms
Experiment type	• Offline evaluation
Evaluation aspects	
Types of data	• Explicit ratings
Data collection	• Participant ratings (prompted after every narrative access)
Data quality and biases	• Platform bias from suggested narratives
Evaluation metrics	• Normalized mean absolute error (for prediction accuracy) • Mean average precision per participant (for usage prediction) • Intralist diversity (for diversity) • Item space coverage (for coverage) • Overestimation of unfairness (for unfairness across participants)
Evaluation system	• NEON intervention web application

[a]NEON: Narrative Experiences Online.

[b]NEON-O trial: NEON for other [eg, nonpsychosis] mental health problems) trial.

[c]ITT: intention to treat.

Recruitment

Participants were recruited across England from March 9, 2020 (both trials), to March 1, 2021 (NEON trial), or March 26, 2021 (NEON-O trial). The trials used a mixed web-based and offline approach to recruit participants. Recruitment was through paid web-based advertising on mental health websites; promotional messaging distributed by a range of community groups and health care practices; promotional messaging distributed on Facebook, Twitter (subsequently rebranded as X), and Google (with the reach of messages enhanced through payments); media appearances by the central study team; and the work of clinical research officers in 11 secondary care mental health trusts.

Clinical research officers approached participants in person and distributed promotional messaging through local authorized channels such as mailing lists of service users who had consented to be contacted about research studies. All promotional advertising and messaging conformed to principles approved in advance by the supervising research ethics committee [44].

Registration

All recruitment approaches directed potential participants to a web-based eligibility checking interface that requested responses to a series of questions specified in the trial protocol. All responses were self-rated. No formal diagnosis of a mental health condition was required for participation. Trial allocation was determined through responses, and eligible potential participants were provided with access to a tailored web-based participant information sheet. Participants subsequently completed a web-based informed consent form by providing an email address and optional telephone number.

The consent process was concluded by clicking on a link in an auto-generated email to validate the email address. After confirming consent, participants completed web-based forms to collect baseline demographic and clinical data and were then randomized using a web-based system validated by a clinical trial unit to the intervention or control arm. Demographic items were age (in years), gender (female, male, or other), ethnicity, region of residence, highest educational qualification, lifetime use of primary care mental health services, lifetime use of specialist mental health services, current use of mental health services in relation to psychosis (NEON trial only), main mental health problem in the last month, best description of recovery status, residential status, and employment status.

Intervention arm users gained immediate access to the NEON intervention until trial end (September 22, 2022), whereas control arm users gained access after completing the 52-week follow-up questionnaires and until trial end. Data on NEON intervention use by both intervention and control group users are within the scope of this study.

Analysis

Objective 1: Describe Participant Characteristics and Patterns of Narrative Requests and Feedback

Participant Characteristics

The demographic and clinical characteristics of participants randomized to each trial were described using means and SDs for normally distributed data and counts with percentages for categorical data. Descriptive statistics were calculated for all baseline demographic data items.

Following UK Data Service guidance on statistical disclosure [45], ethnicity responses were grouped into 2 categories (White British and other ethnicity) due to the small number of participants in most ethnicity categories, although recognizing that this could be perceived as a reductive approach to ethnicity data. “Current mental health problem” also comprised categories with low numbers of participants, so relevant rows were shown as “<5” with no percentage, and other rows were shown as “<10” with no percentage to avoid being able to infer other values.

Patterns of Narrative Requests and Feedback

Data on participant narrative requests and narrative feedback were taken from log files and used to calculate per-trial summary statistics for the number of participants, number of participants who requested at least one narrative, number of narratives at the start and end of the trial, number of narratives given at least one rating, number of narrative requests, number of narrative ratings, number of optional ratings, number of ratings per narrative, number of ratings per rated narrative, length of intervention use by participants, and narrative access routes.

While providing feedback on narratives was encouraged, it was possible for the participant to navigate away from the page and not submit any feedback; therefore, the number of narrative ratings may be smaller than the number of narrative requests, so these figures were reported separately.

Statistics for the number of ratings per narrative present 2 sets of figures with different selection criteria: those including only data for narratives that received at least one rating and those including data for all narratives. This breakdown shows how many ratings NarraGive had access to as it could only access rated narratives.

Nonparametric data were presented as medians and IQRs. Category data were presented as counts with percentages.

Objective 2: Evaluate the NarraGive Recommender System by Comparing Collaborative-Based and Content-Based Narrative Recommendations

Overview

The 3 algorithms (kNN, SVD, and SVD++) were trained and tested using all the available data, representing the point in time at which the trials closed. Training an algorithm involves providing it with a set of data that it can use to create predictions for missing data points. Testing an algorithm involves obtaining these predictions and measuring a feature of those predictions.

The results for objective 2 were obtained using the SurPRISE library (version 1.1.3) for Python (version 3.10.7). Only participants who provided at least one rating and narratives that were given at least one rating were included (as SurPRISE uses participant-item rating pairs as the basis for its predictions), which mirrors the information that NarraGive had access to during the intervention.

This study evaluated NarraGive using the metrics outlined in Textbox 2, applied separately to the content-based algorithm (kNN) and the collaborative filtering–based algorithms (SVD and SVD++).

There are 2 types of metrics: metrics that compare predicted ratings with actual ratings (prediction-based metrics) and metrics that measure a feature of the top-n predicted items (feature-based metrics). Prediction-based metrics include prediction accuracy, usage prediction, and unfairness across participants. Feature-based metrics include diversity and coverage. For prediction-based metrics, there is no standard data-splitting strategy [46], so the data set is split into a training set (75%) and a testing set (25%). For feature-based metrics, the entire data set is used as the training set.

NarraGive only used the first 3 sets of ratings (hopefulness, similarity to the narrator, and similarity to the narrative) to inform its recommendations as these 3 questions had been validated in a feasibility study [29] and the remaining 2 questions were added after the feasibility study. Therefore, only the first 3 sets of ratings were used in the evaluation.

The hopefulness ratings were normalized, which in this case involved shifting the ratings to use the same rating scale as that of the 4 optional questions.

The evaluated version of NarraGive presented in this paper used the same training data as the intervention version of NarraGive with 3 minor modifications. First, where the narratives’ INCRESE characteristics were updated during the trials (eg, to correct human error in inputting characteristics), this evaluation only used the final set of uploaded characteristics. Second, during the intervention, NarraGive filtered out previously requested and blocked narratives. This evaluation included these narratives as the predictions themselves were not influenced by whether a narrative was blocked or previously requested (ie, blocked and previously requested narratives were filtered out after the prediction process in the trial implementation), which could affect, for example, coverage metrics. Third, during the NEON trials, some accounts were removed due to suspected repeat registrations [27]; this evaluation removed all ratings from those participants even though NarraGive may have initially used those ratings.

The logs that were recorded during the intervention did not include NarraGive's internal recommendation lists and instead only recorded the single narrative that was selected to show to the participants. Therefore, using the intervention version of NarraGive would have prevented any comparison of its subsystems and would have allowed for only a limited analysis of its performance as a whole.

The results from objective 1 (about participants and their use of the system) used the data collected from the live intervention, whereas the results from objective 2 (about NarraGive and its subsystems) used the evaluation version of NarraGive.

During a previous feasibility study of NEON (N=25 mental health service users), 465 ratings were collected for the initial set of narratives in the NEON Collection [29]. NarraGive had access to these ratings in the NEON and NEON-O trials to reduce the "cold start" problem, where recommender systems perform poorly for new items and participants [1]. The evaluation excluded these ratings to ensure that NarraGive was only evaluated on data collected live during the NEON intervention.

The SVD and SVD++ algorithms were both randomly initialized according to a normal distribution [47], and the 75:25 split between training and testing sets was also random and calculated using NumPy (a package for scientific computing with Python) [48], where "fresh, unpredictable entropy will be pulled from the OS" [48]. To account for the randomness, cross-validation was performed. The data set was split into 4 folds, with a different fold used as the testing set each time, and the SVD and SVD++ algorithms were reinitialized each time. Medians and IQRs were reported.

An additional exploratory analysis was conducted to determine how the accuracy changed over time. For each month between June 2020 and July 2022 inclusive, data up to but not including the first day of each month were used for training and testing, and the accuracy was measured (using the same accuracy metric as for the main NarraGive evaluation).

Textbox 2. The 5 metrics for evaluating NarraGive.

Metric and metric category

- Prediction accuracy
  - Normalized mean absolute error
- Usage prediction
  - Mean average precision per participant
- Diversity
  - Intralist diversity
- Coverage
  - Item space coverage
- Unfairness across participants
  - Overestimation of unfairness

Prediction Accuracy

Prediction accuracy is the extent to which a recommender system can predict participant ratings [41]. The root-mean-square error (RMSE) and mean absolute error (MAE) [49] are 2 of the most commonly used metrics for evaluating rating prediction accuracy. The MAE uses the absolute difference between the predicted and true ratings, whereas the RMSE squares this difference, which results in the RMSE penalizing inaccurate predictions more [1].

The intervention was designed to be used over time rather than as a one-off, so the accuracy metric should primarily capture the *overall* accuracy rather than emphasizing occasional large inaccuracies (ie, an inaccurate prediction off by 2 points followed by a completely accurate prediction should be treated as no worse than 2 inaccurate predictions off by 1 point each), and this is better achieved using the MAE. Because the hopefulness ratings were normalized, the prediction accuracy metric was the normalized MAE (NMAE).

Different variations in the MAE have been reported in the literature. In particular, some versions square root the averaged summation [1], whereas others do not [47,50]. This evaluation uses SurPRISE's in-built MAE calculation, which does not use a square root.

A lower NMAE indicates greater prediction accuracy. For NarraGive, the scale ranges from 0 (greatest prediction accuracy) to 4 (equation 1 in Multimedia Appendix 3).

Usage prediction

Usage prediction is the rate of correct recommendations in a setting where recommendations are classified as 1 of 2 options: relevant or nonrelevant [41]. An item is relevant to a participant when the participant's rating for it meets a predefined numerical threshold (where the threshold is participant independent and defined per question).

There are 2 common metrics for measuring usage prediction: precision and recall. Precision measures how likely it is that a recommended item is relevant and is defined as the ratio of relevant selected items to the total number of selected items [49]. Recall, conversely, measures how likely it is that a relevant item is selected and is defined as the ratio of relevant selected items to the total number of relevant items [49].

As the length of the recommendation list increases, recall improves, whereas precision worsens [1,49]. The length of NarraGive's internal recommendation list is 10, which is relatively short (compared to, for example, a search engine that recommends tens or hundreds of web pages), meaning that it is impossible to achieve a meaningfully high recall score, so the metric for usage prediction was precision.

As usage prediction is usually used for measuring how relevant a *list* of recommendations is, this evaluation used NarraGive's internal recommendation list (consisting of a 10-narrative list produced using content-based filtering and two 10-narrative lists produced using collaborative filtering). As the participants do not see this list, only metrics that focus on the characteristics of the list as a whole—rather than focusing on the order *within* the list—were used (ie, where the list is treated more like a mathematical set than an ordered list as the ordering beyond the first item does not affect participants), and metrics that exclusively evaluate ranking order were not used.

The analysis of recommender system evaluations by Herlocker et al [49] showed that accuracy metrics can be divided into equivalence classes. One of these classes comprises all metrics that are averaged overall, and one of these classes comprises per-user correlation metrics and the mean average precision per-user metric. To ensure that this analysis of NarraGive captured its performance as widely as possible, a variation of precision that falls into a different equivalence class from that of the NMAE was used, namely, the mean average precision per participant (hereafter, precision).

As the ratings are on a 4-point scale, they need to be converted to a binary scale that classifies recommendations as either relevant or nonrelevant. For optional questions, relevance was defined as "a bit," "quite a lot," or "very much." For hopefulness, relevance was defined as "no change," "a bit more hopeful," or "much more hopeful."

Higher precision indicates a greater proportion of relevant narratives. The scale ranges from 0 (least precision) to 1 (equation 2 in Multimedia Appendix 3).

Diversity

Diversity measures how varied the recommended items are [41]. The current metrics for diversity [41,50] are intralist diversity (ILD) and variations thereof. ILD was developed by Ziegler et al [51], and variations include the rank-sensitive ILD metric by Vargas and Castells [52]. Similar to usage prediction, because the lists used to calculate diversity came from NarraGive's internal recommendation list and the ILD by Ziegler et al [51] is permutation insensitive (ie, the position of recommendations on the list does not affect the diversity score), this metric was used, with cosine similarity as the distance metric calculated using the narratives' INCRESE characteristics.

The original study defined ILD on a per-list basis (ie, for the recommendation list of one participant). This metric has been expanded in this study to be averaged over all participants' lists to produce an overall ILD value.

The lower the ILD value, the greater the diversity among the recommended items. The scale ranges from −1 (most diverse) to 1 (equation 3 in Multimedia Appendix 3).

Coverage

Coverage can be split into participant space coverage and item space coverage [41]. Participant space coverage is the proportion of participants who can be provided with recommendations by the recommender system [1]. The threshold for being provided recommendations is low—a participant needs to have rated at least one narrative (which is achieved when they first access the intervention as it is compulsory to provide a response for the first narrative); thus, participant space coverage was not used. A variation of participant space coverage assesses the proportion of participants that can be recommended *high-quality* items (ie, items with a predicted rating above a predefined threshold). This notion of variable quality among participants is addressed more thoroughly using an *unfairness across participants* metric instead.

Item space coverage is the proportion of items that the recommender system can recommend [1]. Ge et al [53] further split item space coverage into prediction coverage and catalog coverage. They defined prediction coverage as the proportion of items for which the recommender system can produce a predicted rating and catalog coverage as the proportion of items that are recommended in a series of recommendation lists. Because there is no predefined limit to when NarraGive can produce a predicted rating for a narrative, prediction coverage was used.

The definition of catalog coverage by Ge et al [53] captures the set of recommended items produced over time for a single participant (ie, the items that would have been recommended to the participant if they had asked for recommendations at that time; this is different from the set of recommended items that the participant requested and was actually presented with over time).

To capture the *overall* coverage, the proportion of narratives that are recommendable is measured, where a narrative is recommendable if, for at least one participant, the narrative appears in NarraGive's internal recommendation list.

Other versions of coverage use only the top recommendation, but as there are more narratives than there are participants, this would upper bound the item space coverage at approximately three-quarters for the NEON trial—total number of recommendations (which is equal to the number of participants who rated at least one narrative as there is 1 recommendation per participant) divided by the number of narratives that were rated at least once. For longer recommendation lists (such as 10), because recommender system algorithms cannot always produce a predicted rating for each item, a participant's list may be less than the desired length. For this evaluation, a length of 10 was sufficient to ensure that the total number of recommendations being considered across all participants was greater than the number of narratives.

A higher item space coverage value indicates greater item coverage. The scale ranges from 0 (lowest item coverage) to 1 (equation 4 in Multimedia Appendix 3).

Unfairness Across Participants

Unfairness across participants measures whether participants are treated fairly either at the group level (participants in the same group are treated fairly) or at the individual level (participants who are similar are treated fairly) [41].

NarraGive is designed for use in a health care setting—a setting in which protected characteristics such as disability are critical to attend to. It would be crude to stipulate that, for example, all participants should have an equal probability of being recommended a narrative about wheelchair users as this would be far more relevant to some participants than others (and, indeed, a recommender system's entire purpose is to provide *personalized* rather than generic recommendations). As acknowledged by Yao and Huang [54], "in tasks such as recommendation, user preferences are indeed influenced by sensitive features such as gender, race, and age. Therefore, enforcing demographic parity may significantly damage the quality of recommendations."

Thus, they proposed 4 metrics: value unfairness, absolute unfairness, underestimation of unfairness, and overestimation of unfairness. Value unfairness "occurs when one class of user is consistently given higher or lower predictions than their true preferences." Absolute unfairness "measures inconsistency in absolute estimation error across user types." Underestimation of unfairness "measures inconsistency in how much the predictions underestimate the true ratings." Overestimation of unfairness "measures inconsistency in how much the predictions overestimate the true ratings."

NarraGive is implemented in a health care context in which the principle of harm avoidance is crucial. Therefore, one of the most important factors to consider is whether NarraGive is recommending potentially harmful narratives to participants. The metric used to measure this aspect is the overestimation of unfairness.

Overestimation of unfairness measures how much NarraGive consistently overestimates the predicted rating of narratives (ie, how often a participant rates a narrative lower than NarraGive expected) within a disadvantaged subset of the participants and compares this to the overestimation in the nondisadvantaged group.

Participants were divided into groups based on their demographic characteristics. The first grouping was by ethnicity as having a minority ethnicity predicts mental health problems [55], and the second grouping was by gender, informed by Sex and Gender Equity in Research guidelines [56].

The disadvantaged group for the gender comparison was defined as either "Female" or "Other." The disadvantaged group for the ethnicity comparison was defined as "Irish," "Gypsy or Irish Traveller," "Any other White background," "White and Black Caribbean," "White and Black African," "White and Asian," "Any other Mixed/Multiple ethnic background," "Indian," "Pakistani," "Bangladeshi," "Chinese," "Any other Asian background," "African," "Caribbean," "Any other Black/African/Caribbean background," "Arab," and "Any other ethnic group."

The baseline demographic information was used for measuring unfairness between participants as the questions were compulsory, so there was higher completeness of the baseline data than of the personal profile as well as greater granularity with the range of possible answers. The overestimation of unfairness is defined according to the study by Yao and Huang [54].

A lower overestimation of unfairness value indicates that there is less disparity between overestimation among disadvantaged participants and among nondisadvantaged participants. The scale ranges from 0 (least unfair) to 4 (equation 5 in Multimedia Appendix 3).

Other Categories

Zangerle and Bauer [41] detailed 10 categories of evaluation metrics that can be used in the FEVR. Of these, 5 (discussed previously) were used in evaluating NarraGive, and the other 5—ranking, novelty, serendipity, fairness across items, and business oriented—were not used for the reasons described in Multimedia Appendix 4 [33,41,57].

Results

Objective 1: Describe Participant Characteristics and Patterns of Narrative Requests and Feedback

Participant Characteristics

The baseline sociodemographic and clinical characteristics of participants in the NEON (N=739) and NEON-O (N=1023) trials are shown in Table 3.

An exploration of the baseline differences has been reported elsewhere [58].

Table 3. Baseline sociodemographic and clinical characteristics of Narrative Experiences Online (NEON) and NEON for other (eg, nonpsychosis) mental health problems trial (NEON-O) participants.

	NEON baseline (N=739)	NEON-O baseline (N=1023)
Gender, n (%)		
Female	443 (59.9)	811 (79.3)
Male	274 (37.1)	184 (18)
Other	16 (2.2)	18 (1.8)
Age (years), mean (SD)	34.8 (12)	38.4 (13.6)
Ethnicity, n (%)		
White British	561 (75.9)	827 (80.8)
Other ethnicity	172 (23.3)	185 (18.1)
Region of residence, n (%)		
East of England	53 (7.2)	61 (6)
London	166 (22.5)	210 (20.5)
Midlands	112 (15.2)	203 (19.8)
North East and Yorkshire	80 (10.8)	102 (10)
North West	66 (8.9)	98 (9.6)
South East	133 (18)	214 (20.9)
South West	123 (16.6)	125 (12.2)
Highest educational qualification, n (%)		
No qualification	51 (6.9)	30 (2.9)
O-levels or GCSE[a]	117 (15.8)	116 (11.3)
A-levels or AS[b]-levels or NVQ[c] or equivalent	278 (37.6)	327 (32)
Degree-level qualification	207 (28)	349 (34.1)
Higher degree-level qualification	80 (10.8)	191 (18.7)
Living arrangement, n (%)		
Alone	215 (29.1)	229 (22.4)
With others	524 (70.9)	794 (77.6)
Employment status, n (%)		
Employed	277 (37.5)	586 (57.3)
Sheltered employment	10 (1.4)	6 (0.6)
Training and education	76 (10.3)	106 (10.4)
Unemployed	356 (48.2)	272 (26.6)
Retired	20 (2.7)	53 (5.2)
Current mental health problem, n (%)		
I don’t want to say	20 (2.7)	14 (1.4)
I did not experience mental health problems	19 (2.6)	31 (3)
Developmental disorder such as learning disability	15 (2)	12 (1.2)
Eating disorder	15 (2)	45 (4.4)
Mood disorder	265 (35.9)	626 (61.2)
Personality disorder	138 (18.7)	123 (12)
Schizophrenia or other psychosis	154 (20.8)	<5 (<1)
Stress-related disorder	82 (11.1)	152 (14.9)
Substance-related disorder	25 (3.4)	<10 (<1)

	NEON baseline (N=739)	NEON-O baseline (N=1023)
Lifetime user of primary care mental health services, n (%)		
Yes	698 (94.5)	949 (92.8)
No	35 (4.7)	64 (6.3)
Current use of mental health services for psychosis, n (%)		
No contact with any NHS[d] service	100 (13.5)	N/A[e]
General practitioner	234 (31.7)	N/A
Primary care counselor	59 (8)	N/A
IAPT[f]	56 (7.6)	N/A
Specialist community mental health team	261 (35.3)	N/A
Mental health inpatient in hospital	18 (2.4)	N/A
How would you best describe your recovery?, n (%)		
I don't want to say	48 (6.5)	64 (6.3)
Not yet thinking about recovery	91 (12.3)	64 (6.3)
Working on recovery	510 (69)	784 (76.6)
Living beyond disability	84 (11.4)	101 (9.9)

[a]GCSE: General Certificate of Secondary Education.

[b]AS: Advanced Subsidiary.

[c]NVQ: National Vocational Qualification.

[d]NHS: National Health Service.

[e]N/A: not applicable; indicates a question that participants were not asked; in particular, only NEON trial participants were asked about their current use of mental health services.

[f]IAPT: Improving Access to Psychological Therapies.

Patterns of Narrative Requests and Feedback

Table 4 shows summary statistics on the participants, narratives, narrative requests, narrative ratings, intervention use length, and narrative request routes.

A histogram of the lengths of intervention use is shown in Figure S1 in Multimedia Appendix 5. In the NEON trial, 12.4% (92/739) of the participants used the intervention only once, whereas in the NEON-O trial, 19.45% (199/1023) of the participants used the intervention only once. Lengths of >400 days were merged to prevent participant identifiability. The lengths of intervention use for the first 30 days are shown in Figure S2 in Multimedia Appendix 5, with participants who only used the intervention once ("single-use participants") removed to show only nonzero time lengths.

Tables 5 and 6 show the number of narrative rating values that each question received for ratings from NEON trial participants.

Tables 7 and 8 show the number of narrative rating values that each question received for ratings from NEON-O trial participants.

Table 4. Number of narrative requests via the content-based filtering internal access route, collaborative filtering internal access route, and all other internal access routes.

	NEON[a]	NEON-O[b]
Participants, n (%)	739 (100)	1023 (100)
Participants who requested at least one narrative	365 (49.4)	562 (54.9)
Participants who requested and rated at least one narrative	284 (38.4)	409 (40)
Narratives at the start of the trial, n (%)	348 (100)	348 (100)
Narratives at the end of the trial, n (%)	657 (100)	657 (100)
Narratives given at least one rating	375 (57.1)	366 (55.7)
Narrative requests, n (%)	3762 (100)	3548 (100)
Narrative ratings, n (%)	2288 (100)	1896 (100)
Optional ratings	538 (23.5)	538 (28.4)
Ratings per narrative, median (IQR)	1 (0-4)	1 (0-3)
Ratings per rated narrative, median (IQR)	3 (2-6)	2 (1-5)
Length of intervention use, median (IQR)	20 days, 22 hours, and 17 minutes (0 days, 0 minutes, and 0 minutes-251 days, 0 minutes, and 0 minutes)	0 days, 0 hours, and 16 minutes (0 days, 0 hours, and 0 minutes-59 days, 19 hours, and 7 minutes)
Content-based filtering narrative request route, n (%)	554 (14.7)	554 (15.6)
Collaborative filtering narrative request route, n (%)	1113 (29.6)	763 (21.5)
Other narrative request route, n (%)	2095 (55.7)	2232 (62.9)

[a]NEON: Narrative Experiences Online.

[b]NEON-O: NEON for other (eg, nonpsychosis) mental health problems trial.

Table 5. Distribution of narrative rating values for the Narrative Experiences Online trial participants.

	−1, n (%)	0, n (%)	1, n (%)	2, n (%)
Hopefulness ratings (N=2288)	202 (8.83)	901 (39.38)	838 (36.63)	347 (15.17)

Table 6. Distribution of narrative rating values for the Narrative Experiences Online trial participants (N=538 ratings).

	0, n (%)	1, n (%)	2, n (%)	3, n (%)
Similarity to the narrator ratings	132 (24.5)	196 (36.4)	152 (28.3)	58 (10.8)
Similarity to the narrative ratings	103 (19.1)	144 (26.8)	173 (32.2)	118 (21.9)
Learning ratings	104 (19.3)	193 (35.9)	181 (33.6)	60 (11.2)
Empathy ratings	155 (28.8)	206 (38.3)	126 (23.4)	51 (9.5)

Table 7. Distribution of narrative rating values from Narrative Experiences Online for other (eg, nonpsychosis) mental health problems trial participants (N=1896 narrative ratings).

	−1, n (%)	0, n (%)	1, n (%)	2, n (%)
Hopefulness ratings (N=1896 narrative ratings)	206 (10.86)	845 (44.57)	649 (34.23)	196 (10.34)

Table 8. Distribution of narrative rating values from Narrative Experiences Online for other (eg, nonpsychosis) mental health problems trial participants (N=538 ratings).

	0, n (%)	1, n (%)	2, n (%)	3, n (%)
Similarity to the narrator ratings	195 (36.2)	211 (39.2)	104 (19.3)	28 (5.2)
Similarity to the narrative ratings	145 (27)	157 (29.2)	168 (31.2)	68 (12.6)
Learning ratings	134 (24.9)	242 (45)	144 (26.8)	18 (3.3)
Empathy ratings	236 (43.9)	193 (35.9)	90 (16.7)	19 (3.5)

Objective 2: Evaluate the NarraGive Recommender System

Overview

The best results (per metric per trial) are italicized. Where 2 values are equal, neither was better than the other.

For rating sets, better means that all 6 values (across both trials) were better than the 2 corresponding values for the other 2 rating sets, with *N/A* if no rating set was better.

For algorithms, we identified the filtering approach that was better (if any), comparing the content-based and collaborative subsystems of NarraGive per rating set across both trials. Specifically, if the kNN value was better than both SVD and SVD++ values, then we identified content-based filtering as better. If both SVD and SVD++ values were better than the kNN value, then we identified collaborative filtering as better. If neither the kNN nor SVD and SVD++ was better than the other, then the value was calculated per trial.

For trials, better means that each of the 9 values was better than the corresponding value in the other trial, with *N/A* if neither trial was better.

Prediction Accuracy

Tables 9 and 10 show the NMAE of the kNN, SVD, and SVD++ algorithms when trained and tested on the hopefulness, similarity to the narrator, and similarity to the narrative ratings using NEON and NEON-O trial data, respectively.

For NMAE, better means lower.

Hopefulness was the better rating set, collaborative filtering was the better approach, and NEON-O was the better trial.

Table 9. Normalized mean average error (NMAE; using Narrative Experiences Online [NEON] trial data).

NMAE (NEON trial)	Hopefulness, median (IQR)	Similarity to the narrator, median (IQR)	Similarity to the narrative, median (IQR)
kNN[a]	0.686 (0.670-0.703)	1.070 (1.059-1.077)	1.150 (1.140-1.153)
SVD[b]	0.650 (0.638-0.664)	1.043 (1.035-1.047)	1.098 (1.076-1.121)
SVD++	*0.646 (0.639-0.654)* [c]	1.044 (1.038-1.049)	1.099 (1.080-1.120)

[a]kNN: k-nearest neighbor.

[b]SVD: singular value decomposition.

[c]Best result is italicized (per metric per trial).

Table 10. Normalized mean average error (NMAE; using Narrative Experiences Online for other [eg, nonpsychosis] mental health problems trial [NEON-O] data).

NMAE (NEON-O trial)	Hopefulness, median (IQR)	Similarity to the narrator, median (IQR)	Similarity to the narrative, median (IQR)
kNN[a]	0.685 (0.677-0.697)	0.998 (0.992-1.006)	1.099 (1.097-1.106)
SVD[b]	0.650 (0.641-0.659)	0.978 (0.972-0.986)	1.076 (1.066-1.093)
SVD++	*0.644 (0.635-0.655)* [c]	0.983 (0.977-0.989)	1.079 (1.065-1.099)

[a]kNN: k-nearest neighbor.

[b]SVD: singular value decomposition.

[c]Best result is italicized (per metric per trial).

Usage prediction

Tables 11 and 12 show the precision of the kNN, SVD, and SVD++ algorithms when trained and tested on the hopefulness, similarity to the narrator, and similarity to the narrative ratings using NEON and NEON-O trial data, respectively.

For precision, better means higher.

Hopefulness was the better rating set, there was no better filtering approach, and NEON was the better trial.

Table 11. Precision (using Narrative Experiences Online [NEON] trial data).

Precision (NEON trial)	Hopefulness, median (IQR)	Similarity to the narrator, median (IQR)	Similarity to the narrative, median (IQR)
kNN[a]	0.255 (0.250-0.258)	0.054 (0.052-0.055)	0.057 (0.056-0.060)
SVD[b]	*0.256 (0.250-0.259)* [c]	0.053 (0.052-0.055)	0.057 (0.055-0.060)
SVD++	*0.256 (0.250-0.259)*	0.053 (0.051-0.055)	0.057 (0.055-0.060)

[a]kNN: k-nearest neighbor.

[b]SVD: singular value decomposition.

[c]Best result is italicized (per metric per trial).

Table 12. Precision (using Narrative Experiences Online for other [eg, nonpsychosis] mental health problems trial [NEON-O] trial data).

Precision (NEON-O trial)	Hopefulness, median (IQR)	Similarity to the narrator, median (IQR)	Similarity to the narrative, median (IQR)
kNN[a]	*0.181 (0.172-0.192)* [b]	0.037 (0.034-0.040)	0.041 (0.038-0.043)
SVD[c]	0.180 (0.172-0.191)	0.037 (0.035-0.039)	0.041 (0.039-0.043)
SVD++	0.180 (0.172-0.192)	0.036 (0.033-0.039)	0.040 (0.038-0.043)

[a]kNN: k-nearest neighbor.

[b]Best result is italicized (per metric per trial).

[c]SVD: singular value decomposition.

Diversity

Tables 13 and 14 show the ILD of the kNN, SVD, and SVD++ algorithms when trained and tested on the hopefulness, similarity to the narrator, and similarity to the narrative ratings using NEON and NEON-O trial data, respectively.

For ILD, better means lower.

There was no better rating set, collaborative filtering was the better approach for the NEON trial, there was no better approach for the NEON-O trial, and NEON-O was the better trial.

Table 13. Intralist diversity (ILD; using Narrative Experiences Online [NEON] trial data).

ILD (NEON trial)	Hopefulness, median (IQR)	Similarity to the narrator, median (IQR)	Similarity to the narrative, median (IQR)
kNN[a]	0.542 (0.542-0.542)	0.540 (0.540-0.540)	0.538 (0.538-0.538)
SVD[b]	0.531 (0.530-0.532)	0.539 (0.538-0.539)	0.538 (0.537-0.538)
SVD++	*0.530 (0.530-0.531)* [c]	0.539 (0.538-0.539)	0.538 (0.538-0.539)

[a]kNN: k-nearest neighbor.

[b]SVD: singular value decomposition.

[c]Best result is italicized (per metric per trial).

Table 14. Intralist diversity (ILD; using Narrative Experiences Online for other [eg, nonpsychosis] mental health problems trial [NEON-O] trial data).

ILD (NEON-O trial)	Hopefulness, median (IQR)	Similarity to the narrator, median (IQR)	Similarity to the narrative, median (IQR)
kNN[a]	*0.497 (0.497-0.497)* [b]	0.499 (0.499-0.499)	0.500 (0.500-0.500)
SVD[c]	0.498 (0.498-0.499)	0.499 (0.499-0.499)	0.499 (0.499-0.500)
SVD++	0.498 (0.498-0.499)	0.499 (0.499-0.499)	0.499 (0.499-0.499)

[a]kNN: k-nearest neighbor.

[b]Best result is italicized (per metric per trial).

[c]SVD: singular value decomposition.

Coverage

Tables 15 and 16 show the ISC of the kNN, SVD, and SVD++ algorithms when trained and tested on the hopefulness, similarity to the narrator, and similarity to the narrative ratings using NEON and NEON-O trial data, respectively.

For ISC, better means higher.

There was no better rating set, content-based filtering was the better approach for the NEON trial, there was no better approach for the NEON-O trial, and NEON-O was the better trial.

Table 15. Item space coverage (ISC; using Narrative Experiences Online [NEON] trial data).

ISC (NEON trial)	Hopefulness, median (IQR)	Similarity to the narrator, median (IQR)	Similarity to the narrative, median (IQR)
kNN[a]	*0.816 (0.816-0.816)* [b]	0.811 (0.811-0.811)	0.800 (0.800-0.800)
SVD[c]	0.716 (0.711-0.721)	0.761 (0.759-0.764)	0.763 (0.763-0.764)
SVD++	0.709 (0.709-0.711)	0.757 (0.755-0.761)	0.771 (0.769-0.772)

[a]kNN: k-nearest neighbor.

[b]Best results is italicized (per metric per trial).

[c]SVD: singular value decomposition.

Table 16. Item space coverage (ISC; using Narrative Experiences Online for other [eg, nonpsychosis] mental health problems trial [NEON-O] trial data).

ISC (NEON-O trial)	Hopefulness, median (IQR)	Similarity to the narrator, median (IQR)	Similarity to the narrative, median (IQR)
kNN[a]	*0.891 (0.891-0.891)* [b]	0.852 (0.852-0.852)	0.847 (0.847-0.847)
SVD[c]	0.848 (0.844-0.852)	0.840 (0.838-0.842)	0.847 (0.846-0.847)
SVD++	0.850 (0.849-0.851)	0.842 (0.841-0.843)	0.848 (0.844-0.852)

[a]kNN: k-nearest neighbor.

[b]Best results is italicized (per metric per trial).

[c]SVD: singular value decomposition.

Unfairness Across Participants

Tables 17 and 18 show the unfairness, based on gender, of the kNN, SVD, and SVD++ algorithms when trained and tested on the hopefulness, similarity to the narrator, and similarity to the narrative using NEON and NEON-O trial data, respectively.

For unfairness across participants based on gender, better means lower.

Hopefulness was the better rating set, collaborative filtering was the better approach, and there was no better trial.

Tables 19 and 20 show the unfairness, based on ethnicity, of the kNN, SVD, and SVD++ algorithms when trained and tested on the hopefulness, similarity to the narrator, and similarity to the narrative ratings using NEON and NEON-O trial data, respectively.

For unfairness across participants based on ethnicity, better means lower.

Hopefulness was the better rating set, collaborative filtering was the better approach, and there was no better trial.

Table 17. Unfairness across participants based on gender (using Narrative Experiences Online [NEON] trial data).

Unfairness (gender; NEON trial)	Hopefulness, median (IQR)	Similarity to the narrator, median (IQR)	Similarity to the narrative, median (IQR)
kNN[a]	0.429 (0.407-0.455)	0.715 (0.687-0.738)	0.728 (0.715-0.738)
SVD[b]	0.375 (0.347-0.422)	0.664 (0.637-0.695)	0.683 (0.669-0.688)
SVD++	*0.371 (0.342-0.412)* [c]	0.673 (0.643-0.706)	0.695 (0.680-0.700)

[a]kNN: k-nearest neighbor.

[b]SVD: singular value decomposition.

[c]Best results are italicized (per metric per trial).

Table 18. Unfairness across participants based on gender (using Narrative Experiences Online for other [eg, nonpsychosis] mental health problems trial [NEON-O] trial data).

Unfairness (gender; NEON-O trial)	Hopefulness, median (IQR)	Similarity to the narrator, median (IQR)	Similarity to the narrative, median (IQR)
kNN[a]	0.387 (0.368-0.410)	0.727 (0.685-0.769)	0.765 (0.757-0.781)
SVD[b]	0.327 (0.312-0.344)	0.640 (0.634-0.652)	0.671 (0.659-0.696)
SVD++	*0.317 (0.303-0.338)* [c]	0.647 (0.639-0.662)	0.685 (0.678-0.705)

[a]kNN: k-nearest neighbor.

[b]SVD: singular value decomposition.

[c]Best results are italicized (per metric per trial).

Table 19. Unfairness across participants based on ethnicity (using Narrative Experiences Online [NEON] trial data).

Unfairness (ethnicity; NEON trial)	Hopefulness, median (IQR)	Similarity to the narrator, median (IQR)	Similarity to the narrative, median (IQR)
kNN[a]	0.395 (0.370-0.417)	0.769 (0.757-0.783)	0.795 (0.776-0.808)
SVD[b]	0.345 (0.330-0.370)	0.732 (0.712-0.742)	0.727 (0.709-0.745)
SVD++	*0.338 (0.324-0.361)* [c]	0.744 (0.722-0.755)	0.739 (0.716-0.763)

[a]kNN: k-nearest neighbor.

[b]SVD: singular value decomposition.

[c]Best results are italicized (per metric per trial).

Table 20. Unfairness across participants based on ethnicity (using Narrative Experiences Online for other [eg, nonpsychosis] mental health problems trial [NEON-O] trial data).

Unfairness (ethnicity; NEON-O trial)	Hopefulness, median (IQR)	Similarity to the narrator, median (IQR)	Similarity to the narrative, median (IQR)
kNN[a]	0.399 (0.378-0.421)	0.717 (0.687-0.754)	0.751 (0.724-0.787)
SVD[b]	*0.343 (0.328-0.350)* [c]	0.652 (0.642-0.657)	0.667 (0.661-0.690)
SVD++	0.345 (0.331-0.353)	0.658 (0.649-0.664)	0.688 (0.680-0.708)

[a]kNN: k-nearest neighbor.

[b]SVD: singular value decomposition.

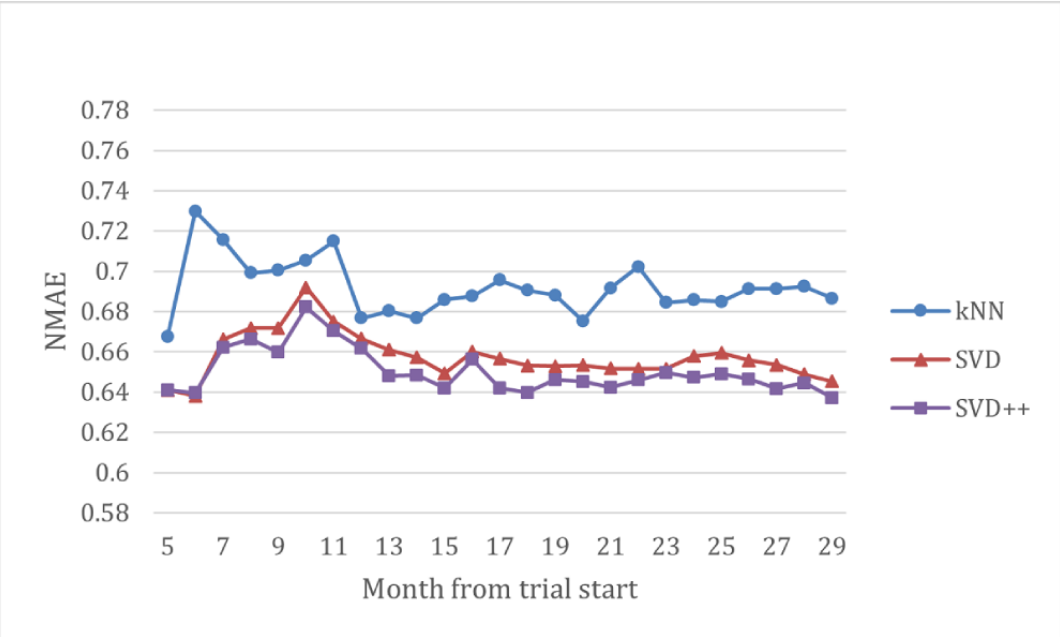
[c]Best results are italicized (per metric per trial).

MAE Over Time

Multimedia Appendix 6 shows how the median NMAE values changed over time (with an interval of 1 month) for the kNN, SVD, and SVD++ algorithms using "Hopefulness" ratings from NEON trial participants.

Figure 1 shows that the 2 collaborative filtering algorithms were more accurate than the content-based filtering algorithm. As the number of ratings increases (and the IQR decreases), the NMAE stabilizes, which happens for all 3 algorithms at approximately 2000 ratings.

Figure 1. Comparison of the accuracy of the k-nearest neighbor (kNN), singular value decomposition (SVD), and SVD++ algorithms over time. NMAE: normalized mean absolute error.



Other Results

Further analysis of the coverage metric showed that certain narratives were not routinely recommended by NarraGive, as described in Multimedia Appendix 7.

Discussion

Principal Findings

Overview

For the NEON trial, the content-based filtering algorithm performed better for coverage; the collaborative filtering algorithms performed better for accuracy, diversity, and unfairness across both gender and ethnicity; and neither algorithm performed better for precision. For the NEON-O trial, the content-based filtering algorithm did not perform better on any metric; the collaborative filtering algorithms performed better on accuracy and unfairness across both gender and ethnicity; and neither algorithm performed better for precision, diversity, or coverage. These findings provide preliminary evidence to inform future implementations.

Table 21 shows, for each metric, the filtering method that was better overall (per trial), the filtering method for the best result (per trial), and the trial in which all 9 values were better than those of the other trial. *N/A* indicates that neither trial nor filtering method was better overall. The table also indicates whether the metric was feature based or prediction based.

These results suggest that clinical population may be associated with recommender system performance. The content-based filtering algorithm had the best performance on feature-based metrics, suggesting that collaborative filtering methods may be associated with producing more accurate predictions whereas content-based filtering methods may be associated with recommending a wider range of items.

A low number of ratings for an item (or having only low ratings) can substantially influence how (or *if*) that item is recommended to other participants, as demonstrated by the 3 unrecommendable narratives in NarraGive.

Table 21. Summary of NarraGive evaluation metrics.

Metric	Prediction based or feature based	Better filtering method		Filtering method of best value		Better trial
		NEON[a]	NEON-O[b]	NEON	NEON-O	
NMAE[c]	Prediction	Collaborative	Collaborative	Collaborative	Collaborative	NEON-O
Precision	Prediction	N/A[d]	N/A	Collaborative	Content	NEON
ILD[e]	Feature	Collaborative	N/A	Collaborative	Content	NEON-O
ISC[f]	Feature	Content	N/A	Content	Content	NEON-O
Unfairness (gender)	Prediction	Collaborative	Collaborative	Collaborative	Collaborative	N/A
Unfairness (ethnicity)	Prediction	Collaborative	Collaborative	Collaborative	Collaborative	N/A

[a]NEON: Narrative Experiences Online.

[b]NEON-O: Narrative Experiences Online for other (eg, nonpsychosis) mental health problems trial.

[c]NMAE: normalized mean absolute error.

[d]N/A: not applicable.

[e]ILD: intralist diversity.

[f]ISC: item space coverage.

Unrecommendable Narratives

Each trial comprised between 1 and 3 ratings for each narrative. This preliminary evaluation only used data from NEON and NEON-O participants (to mirror the data used in the metrics). No rating given by a NEON or NEON-O participant was the highest, and only 1 rating had the optional questions answered (and none of these values were the highest value). The low number of ratings and the low scores given could contribute to the unrecommendableness of these narratives.

NarraGive, and recommender systems in general, often requires that there is a minimum amount of information about a participant or an item before being able to produce recommendations for or about them. In this case, due to SurPRISE's implementation of recommender system algorithms, the kNN, SVD, and SVD++ algorithms require at least one rating for an item for it to be recommended and at least one rating by a participant for them to be recommended a narrative (to the extent that the recommender system never sees these unrated items and nonrating participants as they are filtered out before being passed to the recommender system).

This means that newly added narratives cannot be recommended immediately—they need at least one participant to access the narrative through another method (such as browsing to it) and then rate it. Consequently, there are other unrecommendable narratives (where unrecommendable in this case means that the recommender system does not have *access* to it in the first place rather than having access to it but not producing it as a recommendation for any participant). This is known as the *cold start problem*.

Interpretation of Metrics

In the context of NarraGive, diversity is not necessarily better or worse. An earlier substudy of the NEON study showed that there is unlikely to be a universally hopeful narrative [29], so a skew in recommended narratives is not necessarily a flaw.

Similarly, having greater coverage (ie, being able to recommend a greater proportion of available narratives) may not be useful if some narratives are only hope inspiring for a very small subset of participants.

The 2 unfairness metrics (unfairness across participants based on ethnicity and gender) capture 2 types of unfairness but not all. A previous NEON study [29] identified 7 harm minimization strategies for the NEON intervention, which provided the basis for the unfairness metric (ie, that being recommended narratives with a predicted rating that is higher than the resulting rating may be harmful). However, other types of unfairness, such as those based on disability, were not explored.

Relationship to Prior Work

Recommender Systems for Nonnarrative Texts

Several book recommendation systems exist [59-61], but the focus has mostly been on novellike books rather than on health narratives or recovery narratives. In addition, many health recommender systems exist or have been proposed [62-72], but the focus has largely been on physical health and behavior changes rather than on providing desired content, such as enjoyable, useful, or hope-inspiring recovery narratives.

Dimensionality Reduction

The per-trial values for ILD were very similar. One explanation for this is a known phenomenon called the curse of dimensionality, where the increase in the number of dimensions (where the number of INCRESE characteristics represents the number of dimensions) causes a rapid increase in the "volume" that samples can occupy, which increases the data sparsity exponentially [73]. This increase in dimensionality produces effects such as the concentration of measure [74], where distance values converge and the difference between the furthest and the nearest point tends toward 0 [75], effectively making distance-based similarity comparisons meaningless. High-dimensionality problems can occur with as few as 10 dimensions [75], making the 77-item INCRESE measure susceptible to these issues. This may explain why the results for ILD are so similar—they are based on cosine distance. Other distance measures such as the Euclidean distance and correlation are also susceptible to this challenge [73].

One solution is to reduce the dimensionality of the characteristics before analysis through dimension reduction techniques [75] such as principal component analysis or matrix factorization, which retain as much of the original meaning of the data as possible while reducing the number of dimensions to a practical number.

Strengths and Limitations

There are several strengths to this study. First, the NEON and NEON-O trials produced a unique data set of participant ratings, comprising ratings from both mental health service users and non–service users. This data set was suitable for analysis over time and for comparison of content-based and collaborative filtering algorithms.

Second, the numerous narrative request routes helped prevent exposure bias, and requiring a rating for each narrative helped prevent selection bias.

There are also some limitations to this study. This analysis did not consider individual participants' rating patterns. Further analysis could add participant and item biases, which take into account items' and participants' average ratings to find the *deviation* from this average [76], or weight high-data participants (who have provided many ratings) as more informative than low-data participants.

The decision to include "no change" in hopefulness as an indicator of relevance was made to distinguish from actively hope-reducing narratives, but an alternative approach would be to only include those narratives rated as "a bit more hopeful" or "much more hopeful."

The unfairness across participants metric (overestimation of unfairness) was based on the assumption that overestimated narratives are more likely to be harmful, but it is possible that a participant could rate a narrative highly and still find it harmful and, similarly, rate a narrative as lower than predicted but not find it harmful. The unfairness metrics also did not cover all aspects of unfairness.

Finally, this study is the first evaluation of a recommender system application to lived experience narratives. This is a complex area involving both technical challenges such as the choice of algorithm and ethical challenges such as managing narratives with respect and not just as another form of data. This complexity means that there are no existing standards against which NarraGive can be currently judged, and hence, the comprehensive evaluation presented in this study is primarily intended to be formative for the field rather than evaluative of NarraGive.

Recommendations

There are 6 recommendations for researchers, intervention developers, recommender system developers, and health care professionals.

First, recommender systems with a focus on providing the greatest variety and widest range of content may benefit from using a content-based kNN algorithm, whereas recommender systems with a focus on predicting participants' ratings most accurately may benefit more from the SVD or SVD++ algorithm. Recommender systems with a focus on both should implement a hybrid model with suitably weighted filtering algorithms.

Second, health care professionals should be aware of the unrecommendability of some items and not rely on recommendations to cover the entire search space.

Third, researchers and intervention developers should carefully decide which feature of the recommender system (ie, variety or accuracy) is most important and optimize the recommender system for a specific feature. Depending on what aspect of a recommender system is most important, different methods exist for optimizing for a specific metric, such as for diversity [43] and unfairness [46].

Fourth, platforms containing a recommender system should include other item access mechanisms (such as being able to browse through items) to prevent feedback loops where participants can only rate items that already have many high ratings while unrated items remain unrated and unrecommended [28]. This helps reduce the number of inaccessible narratives (because if the recommender system is the only access route, any unrecommendable narratives will be entirely inaccessible to participants), and developers could include a "random" access route (in addition to the recommender system access route) that is weighted toward these inaccessible narratives.

Fifth, recommender system developers should actively encourage the rating of new items, such as by suggesting them to participants or having a random button that is weighted toward new narratives. Alternatively, the narratives could be given an initial set of ratings before being published.

Sixth, initial studies should be conducted on proposed recommender systems to find the number of ratings required for the accuracy to stabilize; for NarraGive, this was approximately 2000 ratings. Because clinical population may be associated with recommender system performance, initial studies should also be used to inform clinical population selection.

Implications for Future Work

Key future questions include whether a single or hybrid recommender system is optimal, a wider consideration of available algorithms and clarification of the rationale for selection, the rationale and timing of training and retraining the algorithm, and the identification of the most important metrics through which algorithmic performance should be evaluated.

For example, future studies investigating the use of recommender systems for recommending narratives could incorporate the similarity between participants and *narratives* by training filtering algorithms to recommend narratives with narrators that are either similar to or different from participants depending on (either implicit or explicit) participant preference.

Conclusions

Clinical population may be associated with recommender system performance. The collaborative filtering algorithms were more accurate and less unfair than the content-based filtering algorithm. Recommender systems are susceptible to a wide range of biases, and it is important to mitigate these by providing enough data for the recommender system to start with (to prevent overfitting), ensuring that there are other ways of accessing items besides through the recommender system (to prevent a feedback loop between accessed items and recommended items), and encouraging participants to provide feedback on every item they interact with (to prevent participants from only providing feedback when they have strong opinions).

Acknowledgments

This study was funded by the National Institute for Health and Care Research (NIHR; personal experience as a recovery resource in psychosis: Narrative Experiences Online Programme [RP-PG-0615-20016]). MS and SRE acknowledge the support of the NIHR Nottingham Biomedical Research Centre (NIHR203310). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. FN acknowledges support from an NIHR Advanced Fellowship (NIHR302218) and an Anne McLaren Fellowship.

Data Availability

Data will be available upon reasonable request, supervised by the study sponsor. Only anonymous and pseudonymous elements of the data sets used or analyzed during the study will be available. Research data will be available from the study sponsor until the end of the retention period. After the retention period, availability through the study sponsor or chief investigator may be at their discretion. Contact the study sponsor through Research@nottshc.nhs.uk citing Integrated Research Application System ID249015.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Definitions of the categories within the "Browse Stories" page of the Narrative Experiences Online intervention and screenshots of the web application.

[PDF File (Adobe PDF File), 190 KB - mental_v11i1e45754_app1.pdf]

Multimedia Appendix 2

Available options in the Personal Profile in the Narrative Experiences Online intervention.

[XLSX File (Microsoft Excel File), 11 KB - mental_v11i1e45754_app2.xlsx]

Multimedia Appendix 3

Equations used for the normalized mean absolute error, precision, intralist diversity, item space coverage, and unfairness across participants.

[PDF File (Adobe PDF File), 273 KB - mental_v11i1e45754_app3.pdf]

Multimedia Appendix 4

Rationale for not using 5 of the evaluation categories.

[PDF File (Adobe PDF File), 191 KB - mental_v11i1e45754_app4.pdf]

Multimedia Appendix 5

Graphs showing the length of use of the Narrative Experiences Online (NEON) intervention comparing NEON and NEON for other (eg, nonpsychosis) mental health problems (NEON-O) trial participants.

[PDF File (Adobe PDF File), 218 KB - mental_v11i1e45754_app5.pdf]

Multimedia Appendix 6

Graphs showing the normalized mean absolute error over time for the k-nearest neighbor, singular value decomposition (SVD), and SVD++ algorithms.

[PDF File (Adobe PDF File), 204 KB - mental_v11i1e45754_app6.pdf]

Multimedia Appendix 7

Nonrecommendable narratives.

[PDF File (Adobe PDF File), 175 KB - mental_v11i1e45754_app7.pdf]

References

1. Ricci F, Rokach L, Shapira B, Kantor PB. Recommender Systems Handbook. Cham, Switzerland: Springer; 2011.
2. About the Music Genome Project®. Pandora Media. URL: https://www.pandora.com/about/mgp [accessed 2022-11-14]
3. Salmi S, Mérelle S, Gilissen R, Brinkman WP. Content-based recommender support system for counselors in a suicide prevention chat helpline: design and evaluation study. J Med Internet Res 2021 Jan 07;23(1):e21690 [FREE Full text] [doi: 10.2196/21690] [Medline: 33410755]
4. Gannod GC, Abbott KM, Van Haitsma K, Martindale N, Heppner A. A machine learning recommender system to tailor preference assessments to enhance person-centered care among nursing home residents. Gerontologist 2019 Jan 09;59(1):167-176 [FREE Full text] [doi: 10.1093/geront/gny056] [Medline: 29790930]
5. Guo L, Jin B, Yao C, Yang H, Huang D, Wang F. Which doctor to trust: a recommender system for identifying the right doctors. J Med Internet Res 2016 Jul 07;18(7):e186 [FREE Full text] [doi: 10.2196/jmir.6015] [Medline: 27390219]
6. Given LM. The SAGE Encyclopedia of Qualitative Research Methods. Thousand Oaks, CA: Sage Publications; 2008.
7. Wise M, Han JY, Shaw B, McTavish F, Gustafson DH. Effects of using online narrative and didactic information on healthcare participation for breast cancer patients. Patient Educ Couns 2008 Mar;70(3):348-356 [FREE Full text] [doi: 10.1016/j.pec.2007.11.009] [Medline: 18201859]
8. Houston TK, Cherrington A, Coley HL, Robinson KM, Trobaugh JA, Williams JH, et al. The art and science of patient storytelling-harnessing narrative communication for behavioral interventions: the ACCE project. J Health Commun 2011 Aug;16(7):686-697. [doi: 10.1080/10810730.2011.551997] [Medline: 21541875]
9. Gucciardi E, Reynolds E, Karam G, Beanlands H, Sidani S, Espin S. Group-based storytelling in disease self-management among people with diabetes. Chronic Illn 2021 Sep;17(3):306-320 [FREE Full text] [doi: 10.1177/1742395319859395] [Medline: 31266352]
10. Andreae SJ, Andreae LJ, Cherrington A, Richman J, Safford M. Peer coach delivered storytelling program for diabetes medication adherence: intervention development and process outcomes. Contemp Clin Trials Commun 2020 Dec;20:100653 [FREE Full text] [doi: 10.1016/j.conctc.2020.100653] [Medline: 33024882]
11. Greenhalgh T, Collard A, Begum N. Sharing stories: complex intervention for diabetes education in minority ethnic groups who do not speak English. BMJ 2005 Mar 19;330(7492):628 [FREE Full text] [doi: 10.1136/bmj.330.7492.628] [Medline: 15774990]
12. Gidman J. Listening to stories: valuing knowledge from patient experience. Nurse Educ Pract 2013 May;13(3):192-196. [doi: 10.1016/j.nepr.2012.09.006] [Medline: 23084838]
13. Rodrigo C, Tedla N, Thomas S, Polly P, Herbert C, Velan G, et al. Using narratives to teach students enrolled in science and medical science bachelor's degree programs. Med Sci Educ 2019 Jun;29(2):357-361 [FREE Full text] [doi: 10.1007/s40670-019-00698-1] [Medline: 34457492]

14. Stacey G, Hardy P. Challenging the shock of reality through digital storytelling. Nurse Educ Pract 2011 Mar;11(2):159-164. [doi: 10.1016/j.nepr.2010.08.003] [Medline: 20829115]
15. Goodman R, Newman D. Testing a digital storytelling intervention to reduce stress in adolescent females. Storytell Self Soc 2014;10(2):177-193. [doi: 10.13110/storselfsoci.10.2.0177]
16. Dawson L, Mullan B, Touyz S, Rhodes P. Are recovery stories helpful for women with eating disorders? A pilot study and commentary on future research. J Eat Disord 2018;6:21 [FREE Full text] [doi: 10.1186/s40337-018-0206-2] [Medline: 30128153]
17. Williams A, Fossey E, Farhall J, Foley F, Thomas N. Recovery after psychosis: qualitative study of service user experiences of lived experience videos on a recovery-oriented website. JMIR Ment Health 2018 May 08;5(2):e37 [FREE Full text] [doi: 10.2196/mental.9934] [Medline: 29739737]
18. Bove A, Tryon R. The power of storytelling: the experiences of incarcerated women sharing their stories. Int J Offender Ther Comp Criminol 2018 Nov;62(15):4814-4833. [doi: 10.1177/0306624X18785100] [Medline: 29978737]
19. Holm AK, Lepp M, Ringsberg KC. Dementia: involving patients in storytelling--a caring intervention. A pilot study. J Clin Nurs 2005 Feb;14(2):256-263. [doi: 10.1111/j.1365-2702.2004.01042.x] [Medline: 15669935]
20. Due-Christensen M, Zoffmann V, Hommel E, Lau M. Can sharing experiences in groups reduce the burden of living with diabetes, regardless of glycaemic control? Diabet Med 2012 Feb;29(2):251-256. [doi: 10.1111/j.1464-5491.2011.03521.x] [Medline: 22061095]
21. Gunnarsson AB, Wagman P, Håkansson C, Hedin K. The Tree Theme Method® (TTM), an occupational therapy intervention for treating depression and anxiety: study protocol of a randomized controlled trial. BMC Psychol 2015 Nov 09;3:40 [FREE Full text] [doi: 10.1186/s40359-015-0097-9] [Medline: 26552426]
22. Coleman L, Ramm J, Cooke R. The effectiveness of an innovative intervention aimed at reducing binge drinking among young people: Results from a pilot study. Drugs Educ Prev Policy 2010 Jun 30;17(4):413-430. [doi: 10.3109/09687630802572599]
23. Abma TA. Learning by telling: story telling workshops as an organizational learning intervention. Manag Learn 2016 Aug 18;34(2):221-240. [doi: 10.1177/1350507603034002004]
24. Yeo C, Hare-Duke L, Rennick-Egglestone S, Bradstreet S, Callard F, Hui A, et al. The VOICES typology of curatorial decisions in narrative collections of the lived experiences of mental health service use, recovery, or madness: qualitative study. JMIR Ment Health 2020 Sep 18;7(9):e16290 [FREE Full text] [doi: 10.2196/16290] [Medline: 32945771]
25. Yeo C, Rennick-Egglestone S, Armstrong V, Borg M, Franklin D, Klevan T, et al. Uses and misuses of recorded mental health lived experience narratives in healthcare and community settings: systematic review. Schizophr Bull 2022 Jan 21;48(1):134-144 [FREE Full text] [doi: 10.1093/schbul/sbab097] [Medline: 34423840]
26. Yeo C, Rennick-Egglestone S, Armstrong V, Borg M, Charles A, Duke LH, et al. The influence of curator goals on collections of lived experience narratives: a qualitative study. J Recovery Ment Health 2021;4(2):16-28 [FREE Full text] [Medline: 34988285]
27. Rennick-Egglestone S, Elliott R, Smuk M, Robinson C, Bailey S, Smith R, et al. Impact of receiving recorded mental health recovery narratives on quality of life in people experiencing psychosis, people experiencing other mental health problems and for informal carers: Narrative Experiences Online (NEON) study protocol for three randomised controlled trials. Trials 2020 Jul 20;21(1):661 [FREE Full text] [doi: 10.1186/s13063-020-04428-6] [Medline: 32690105]
28. Rennick-Egglestone S, Elliott R, Newby C, Robinson C, Slade M. Impact of receiving recorded mental health recovery narratives on quality of life in people experiencing non-psychosis mental health problems (NEON-O Trial): updated randomised controlled trial protocol. Trials 2022 Jan 29;23(1):90 [FREE Full text] [doi: 10.1186/s13063-022-06027-z] [Medline: 35093141]
29. Slade M, Rennick-Egglestone S, Llewellyn-Beardsley J, Yeo C, Roe J, Bailey S, et al. Recorded mental health recovery narratives as a resource for people affected by mental health problems: development of the Narrative Experiences Online (NEON) intervention. JMIR Form Res 2021 May 27;5(5):e24417 [FREE Full text] [doi: 10.2196/24417] [Medline: 34042595]
30. Kotera Y, Rennick-Egglestone S, Ng F, Llewellyn-Beardsley J, Ali Y, Newby C, et al. Assessing diversity and inclusivity is the next frontier in mental health recovery narrative research and practice. JMIR Ment Health 2023 Apr 17;10:e44601 [FREE Full text] [doi: 10.2196/44601] [Medline: 37067882]
31. Llewellyn-Beardsley J, Barbic S, Rennick-Egglestone S, Ng F, Roe J, Hui A, et al. INCRESE: development of an inventory to characterize recorded mental health recovery narratives. J Recovery Ment Health 2020;3(2):25-44 [FREE Full text] [Medline: 34988284]
32. Çano E, Morisio M. Hybrid recommender systems: a systematic literature review. Intell Data Anal 2017 Nov 15;21(6):1487-1524. [doi: 10.3233/ida-163209]
33. Chen J, Dong H, Wang X, Feng F, Wang M, He X. Bias and debias in recommender system: a survey and future directions. ACM Trans Inf Syst 2023 Feb 07;41(3):1-39. [doi: 10.1145/3564284]
34. Rennick-Egglestone S, Morgan K, Llewellyn-Beardsley J, Ramsay A, McGranahan R, Gillard S, et al. Mental health recovery narratives and their impact on recipients: systematic review and narrative synthesis. Can J Psychiatry 2019 Oct 02;64(10):669-679 [FREE Full text] [doi: 10.1177/0706743719846108] [Medline: 31046432]

35. Rennick-Egglestone S, Ramsay A, McGranahan R, Llewellyn-Beardsley J, Hui A, Pollock K, et al. The impact of mental health recovery narratives on recipients experiencing mental health problems: qualitative analysis and change model. PLoS One 2019;14(12):e0226201 [FREE Full text] [doi: 10.1371/journal.pone.0226201] [Medline: 31834902]
36. Ng F, Charles A, Pollock K, Rennick-Egglestone S, Cuijpers P, Gillard S, et al. The mechanisms and processes of connection: developing a causal chain model capturing impacts of receiving recorded mental health recovery narratives. BMC Psychiatry 2019 Dec 21;19(1):413 [FREE Full text] [doi: 10.1186/s12888-019-2405-z] [Medline: 31864315]
37. Shi L, Li S, Ding X, Bu Z. Selection bias mitigation in recommender system using uninteresting items based on temporal visibility. Expert Syst Appl 2023 Mar;213:118932. [doi: 10.1016/j.eswa.2022.118932]
38. Hug N. Surprise: a Python library for recommender systems. J Open Source Softw 2020 Aug;5(52):2174. [doi: 10.21105/joss.02174]
39. Understanding matrix factorization for recommendation (part 1) - preliminary insights on PCA. Nicolas Hug. 2017. URL: https://nicolas-hug.com/blog/matrix_facto_1 [accessed 2022-11-14]
40. Slade M, Rennick-Egglestone S, Elliott RA, Newby C, Robinson C, Gavan SP, NEON study group. Effectiveness and cost-effectiveness of online recorded recovery narratives in improving quality of life for people with non-psychotic mental health problems: a pragmatic randomized controlled trial. World Psychiatry 2024 Feb;23(1):101-112 [FREE Full text] [doi: 10.1002/wps.21176] [Medline: 38214639]
41. Zangerle E, Bauer C. Evaluating recommender systems: survey and framework. ACM Comput Surv 2022 Dec 23;55(8):1-38. [doi: 10.1145/3556536]
42. Priebe S, Huxley P, Knight S, Evans S. Application and results of the Manchester Short Assessment of quality of life (MANSA). Int J Soc Psychiatry 1999 Mar 01;45(1):7-12 [FREE Full text] [doi: 10.1177/002076409904500102] [Medline: 10443245]
43. Priebe S, Kelley L, Omer S, Golden E, Walsh S, Khanom H, et al. The effectiveness of a patient-centred assessment with a solution-focused approach (DIALOG+) for patients with psychosis: a pragmatic cluster-randomised controlled trial in community care. Psychother Psychosom 2015;84(5):304-313 [FREE Full text] [doi: 10.1159/000430991] [Medline: 26278784]
44. Rennick-Egglestone S. Principles for the production and dissemination of recruitment material for three clinical trials of an online intervention. Trials 2021 Jul 09;22(1):441 [FREE Full text] [doi: 10.1186/s13063-021-05412-4] [Medline: 34243794]
45. Handbook on statistical disclosure control. European Union. 2010. URL: https://ec.europa.eu/eurostat/cros/system/files/SDC_Handbook.pdf [accessed 2022-11-14]
46. Meng Z, McCreadie R, Macdonald C, Ounis I. Exploring data splitting strategies for the evaluation of recommendation models. In: Proceedings of the 14th ACM Conference on Recommender Systems. 2020 Presented at: RecSys '20; September 22-26, 2020; Virtual Event p. 681-686 URL: https://dl.acm.org/doi/10.1145/3383313.3418479 [doi: 10.1145/3383313.3418479]
47. Hug N. Surprise: a Python library for recommender systems. J Open Source Softw 2023;5(52):2174 [FREE Full text] [doi: 10.21105/joss.02174]
48. NumPy community. NumPy reference, release 1.21.0. NumPy. 2021. URL: https://numpy.org/doc/1.21/numpy-ref.pdf [accessed 2024-03-16]
49. Herlocker JL, Konstan JA, Terveen LG, Riedl JT. Evaluating collaborative filtering recommender systems. ACM Trans Inf Syst 2004 Jan;22(1):5-53. [doi: 10.1145/963770.963772]
50. Silveira T, Zhang M, Lin X, Liu Y, Ma S. How good your recommender system is? A survey on evaluations in recommendation. Int J Mach Learn Cybern 2017 Dec 14;10(5):813-831. [doi: 10.1007/s13042-017-0762-9]
51. Ziegler CN, McNee SM, Konstan JA, Lausen G. Improving recommendation lists through topic diversification. In: Proceedings of the 14th international conference on World Wide Web. 2005 Presented at: WWW '05; May 10-14, 2005; Chiba, Japan p. 22-32 URL: https://dl.acm.org/doi/10.1145/1060745.1060754 [doi: 10.1145/1060745.1060754]
52. Vargas S, Castells P. Rank and relevance in novelty and diversity metrics for recommender systems. In: Proceedings of the fifth ACM conference on Recommender systems. 2011 Presented at: RecSys '11; October 23-27, 2011; Chicago, IL p. 109-116 URL: https://dl.acm.org/doi/10.1145/2043932.2043955 [doi: 10.1145/2043932.2043955]
53. Ge M, Delgado-Battenfeld C, Jannach D. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In: Proceedings of the 4th ACM Conference on Recommender Systems. 2010 Presented at: RecSys '10; September 26-30, 2010; Barcelona, Spain p. 257-260 URL: https://dl.acm.org/doi/10.1145/1864708.1864761 [doi: 10.1145/1864708.1864761]
54. Yao S, Huang B. Beyond parity: fairness objectives for collaborative filtering. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS '17; December 4-9, 2017; Long Beach, CA p. 2925-2934 URL: https://dl.acm.org/doi/pdf/10.5555/3294996.3295052
55. Sattler FA, Zeyen J. Intersecting identities, minority stress, and mental health problems in different sexual and ethnic groups. Stigma Health 2021 Nov;6(4):457-466. [doi: 10.1037/sah0000340]
56. Heidari S, Babor TF, De Castro P, Tort S, Curno M. Sex and gender equity in research: rationale for the SAGER guidelines and recommended use. Res Integr Peer Rev 2016;1:2 [FREE Full text] [doi: 10.1186/s41073-016-0007-6] [Medline: 29451543]

57. Diaz F, Mitra B, Ekstrand MD, Biega AJ, Carterette BA. Evaluating stochastic rankings with expected exposure. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020 Presented at: CIKM '20; October 19-23, 2020; Virtual Event p. 275-284 URL: https://dl.acm.org/doi/10.1145/3340531.3411962 [doi: 10.1145/3340531.3411962]
58. Rennick-Egglestone S, Newby C, Robinson C, Yeo C, Ng F, Elliott RA, et al. Differences between online trial participants who have used statutory mental health services and those who have not: analysis of baseline data from 2 pragmatic trials of a digital health intervention. J Med Internet Res 2023 Jun 27;25:e44687 [FREE Full text] [doi: 10.2196/44687] [Medline: 37368471]
59. Rana C, Jain S. Building a book recommender system using time based content filtering. WSEAS Trans Comput 2012;11(2):27-33 [FREE Full text]
60. Pathak D, Matharia S, Murthy CN. NOVA: hybrid book recommendation engine. In: Proceedings of the 3rd IEEE International Advance Computing Conference. 2013 Presented at: IACC '13; February 22-23, 2013; Ghaziabad, India p. 22-23 URL: https://ieeexplore.ieee.org/document/6514359 [doi: 10.1109/iadcc.2013.6514359]
61. Ali Z, Khusro S, Ullah I. A hybrid book recommender system based on table of contents (ToC) and association rule mining. In: Proceedings of the 10th International Conference on Informatics and Systems. 2016 Presented at: INFOS '16; May 9-11, 2016; Giza, Egypt p. 68-74 URL: https://dl.acm.org/doi/10.1145/2908446.2908481 [doi: 10.1145/2908446.2908481]
62. Granda Morales LF, Valdiviezo-Diaz P, Reátegui R, Barba-Guaman L. Drug recommendation system for diabetes using a collaborative filtering and clustering approach: development and performance evaluation. J Med Internet Res 2022 Jul 15;24(7):e37233 [FREE Full text] [doi: 10.2196/37233] [Medline: 35838763]
63. Sanchez Bocanegra CL, Sevillano Ramos JL, Rizo C, Civit A, Fernandez-Luque L. HealthRecSys: a semantic content-based recommender system to complement health videos. BMC Med Inform Decis Mak 2017 May 15;17(1):63 [FREE Full text] [doi: 10.1186/s12911-017-0431-7] [Medline: 28506225]
64. Ali F, Islam SM, Kwak D, Khan P, Ullah N, Yoo SJ, et al. Type-2 fuzzy ontology–aided recommendation systems for IoT–based healthcare. Comput Commun 2018 Apr;119:138-155. [doi: 10.1016/j.comcom.2017.10.005]
65. Chen J, Li K, Rong H, Bilal K, Yang N, Li K. A disease diagnosis and treatment recommendation system based on big data mining and cloud computing. Inf Sci 2018 Apr;435:124-149. [doi: 10.1016/j.ins.2018.01.001]
66. Deepa N, Pandiaraja P. Hybrid context aware recommendation system for E-Health care by Merkle hash tree from cloud using evolutionary algorithm. Soft Comput 2019 Sep 09;24(10):7149-7161. [doi: 10.1007/s00500-019-04322-7]
67. Folino F, Pizzuti C. A comorbidity-based recommendation engine for disease prediction. In: Proceedings of the 23rd International Symposium on Computer-Based Medical Systems. 2010 Presented at: CBMS '10; October 12-15, 2010; Bentley, WA p. 6-12 URL: https://ieeexplore.ieee.org/document/6042664 [doi: 10.1109/cbms.2010.6042664]
68. Hoens TR, Blanton M, Chawla NV. Reliable medical recommendation systems with patient privacy. In: Proceedings of the 1st ACM International Health Informatics Symposium. 2010 Presented at: IHI '10; November 11-12, 2010; Arlington, TX p. 173-182 URL: https://dl.acm.org/doi/10.1145/1882992.1883018 [doi: 10.1145/1882992.1883018]
69. Hussein AS, Omar WM, Li X, Ati M. Efficient chronic disease diagnosis prediction and recommendation system. In: Proceedings of the 2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences. 2012 Presented at: IECBES '12; December 17-19, 2012; Langkawi, Malaysia p. 209-214 URL: https://ieeexplore.ieee.org/document/6498117 [doi: 10.1109/iecbes.2012.6498117]
70. Iwendi C, Khan S, Anajemba JH, Bashir AK, Noor F. Realizing an efficient IoMT-assisted patient diet recommendation system through machine learning model. IEEE Access 2020;8:28462-28474. [doi: 10.1109/access.2020.2968537]
71. Kim J, Lee D, Chung KY. Item recommendation based on context-aware model for personalized u-healthcare service. Multimed Tools Appl 2011 Nov 22;71(2):855-872. [doi: 10.1007/s11042-011-0920-0]
72. Kim JH, Lee JH, Park JS, Lee YH, Rim KW. Design of diet recommendation system for healthcare service based on user information. In: Proceedings of the 4th International Conference on Computer Sciences and Convergence Information Technology. 2009 Presented at: ICCIT '09; November 24-26, 2009; Seoul, Korea p. 516-518 URL: https://ieeexplore.ieee.org/document/5367898 [doi: 10.1109/iccit.2009.293]
73. Altman N, Krzywinski M. The curse(s) of dimensionality. Nat Methods 2018 Jun 31;15(6):399-400. [doi: 10.1038/s41592-018-0019-x] [Medline: 29855577]
74. Pestov V. On the geometry of similarity search: dimensionality curse and concentration of measure. Inf Process Lett 2000 Jan;73(1-2):47-51. [doi: 10.1016/s0020-0190(99)00156-8]
75. Beyer K, Goldstein J, Ramakrishnan R, Shaft U. When is “nearest neighbor” meaningful? In: Proceedings of the 7th International Conference on Database Theory. 1999 Presented at: ICDT '99; January 10-12, 1999; Jerusalem, Israel p. 217-235 URL: https://link.springer.com/chapter/10.1007/3-540-49257-7_15
76. Johnson CC. Logistic matrix factorization for implicit feedback data. Stanford University. 2014. URL: https://web.stanford.edu/~rezab/nips2014workshop/submits/logmat.pdf [accessed 2024-03-11]

Abbreviations

FEVR: framework for evaluating recommender systems

ILD: intralist diversity
INCRESE: Inventory of Characteristics of Recovery Stories
kNN: k-nearest neighbor
MAE: mean absolute error
NEON: Narrative Experiences Online
NEON-O: Narrative Experiences Online for other (eg, nonpsychosis) mental health problems
NMAE: normalized mean absolute error
RMSE: root-mean-square error
SurPRISE: Simple Python Recommendation System Engine
SVD: singular value decomposition

Edited by J Torous; submitted 16.01.23; peer-reviewed by N Josifovski, S Salmi, G Pu; comments to author 31.05.23; revised version received 11.09.23; accepted 15.02.24; published 29.03.24.

Please cite as:
Slade E, Rennick-Egglestone S, Ng F, Kotera Y, Llewellyn-Beardsley J, Newby C, Glover T, Keppens J, Slade M
The Implementation of Recommender Systems for Mental Health Recovery Narratives: Evaluation of Use and Performance
JMIR Ment Health 2024;11:e45754
URL: https://mental.jmir.org/2024/1/e45754
doi:10.2196/45754
PMID:38551630

The Opportunities and Risks of Large Language Models in Mental Health

Hannah R Lawrence[1], PhD; Renee A Schneider[2], PhD; Susan B Rubin[1], PhD; Maja J Matarić[2], PhD; Daniel J McDuff[2], PhD; Megan Jones Bell[2], PsyD

1

2

Corresponding Author:
Megan Jones Bell, PsyD

Abstract

Global rates of mental health concerns are rising, and there is increasing realization that existing models of mental health care will not adequately expand to meet the demand. With the emergence of large language models (LLMs) has come great optimism regarding their promise to create novel, large-scale solutions to support mental health. Despite their nascence, LLMs have already been applied to mental health–related tasks. In this paper, we summarize the extant literature on efforts to use LLMs to provide mental health education, assessment, and intervention and highlight key opportunities for positive impact in each area. We then highlight risks associated with LLMs' application to mental health and encourage the adoption of strategies to mitigate these risks. The urgent need for mental health support must be balanced with responsible development, testing, and deployment of mental health LLMs. It is especially critical to ensure that mental health LLMs are fine-tuned for mental health, enhance mental health equity, and adhere to ethical standards and that people, including those with lived experience with mental health concerns, are involved in all stages from development through deployment. Prioritizing these efforts will minimize potential harms to mental health and maximize the likelihood that LLMs will positively impact mental health globally.

(JMIR Ment Health 2024;11:e59479) doi:10.2196/59479

KEYWORDS

artificial intelligence; AI; generative AI; large language models; mental health; mental health education; language model; mental health care; health equity; ethical; development; deployment

Introduction

Globally, half of all individuals will experience a mental health disorder in their lifetimes [1], and at any given point, 1 in 8 people are experiencing a mental health concern [2]. Despite greater attention provided in the recent years to mental health, the rate of mental health concerns has increased [2,3], and access to mental health care has not expanded to adequately meet the demand [4]. In the United States alone, the average time between the onset of mental health symptoms and treatment is 11 years [5], and nearly half of the global population lives in regions with a shortage of mental health professionals [2].

To overcome inadequate access to effective and equitable mental health care, large-scale solutions are needed. The emergence of large language models (LLMs) brings hope regarding their application to mental health and their potential to provide such solutions due to their relevance to mental health education, assessment, and intervention. LLMs are artificial intelligence models trained using extensive data sets to predict language sequences [6]. By leveraging huge neural architectures, LLMs can organize complex and abstract concepts. This enables them to identify, translate, predict, and generate new content. LLMs can be fine-tuned for specific domains (eg, mental health) and enable interactions in natural language, as do many mental health assessments and interventions, highlighting the enormous potential they have to revolutionize mental health care. In this paper, we first summarize the research done to date applying LLMs to mental health. Then, we highlight key opportunities and risks associated with mental health LLMs and put forth suggested risk mitigation strategies. Finally, we make recommendations for the responsible use of LLMs in the mental health domain.

Applications of LLMs to Mental Health

Overview

Initial tests of LLMs' capabilities across mental health education, assessment, and intervention are promising. When considering this literature base, which we review next, it is important to first distinguish between general-purpose, consumer LLMs (eg, ChatGPT [OpenAI] and Gemini [Google]) and domain-specific LLMs (eg, Med-LM [Google]). General-purpose LLMs are trained on large corpora of text and are designed to perform a wide range of tasks. Domain-specific LLMs, on the other hand, typically build upon general-purpose LLMs through various strategies of fine-tuning with curated data to complete tasks within an area of focus. Given that general-purpose LLMs are largely trained with unrestricted text, they risk generating inaccurate, biased, stigmatizing, and harmful information about mental health. Developers of domain-specific LLMs can mitigate some of this risk by incorporating strategies during fine-tuning and evaluation such as using high-quality evidence-based information and attribution techniques [7], but it remains difficult to remove all possible risk from LLM-generated content. Given these important distinctions, in the paper that follows we clarify when findings are specific to general-purpose versus domain-specific LLMs where possible.

Education

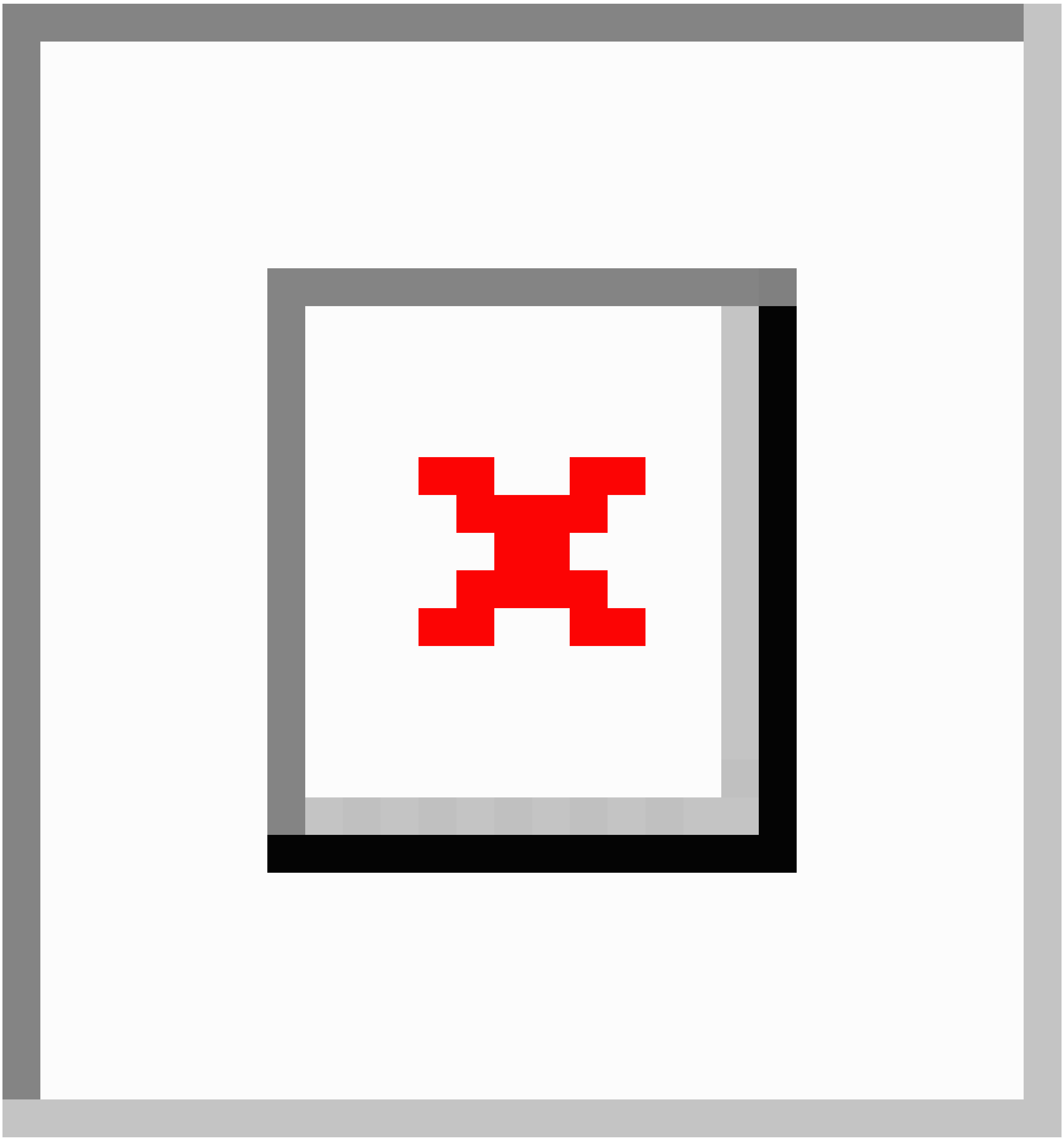
One area of opportunity for LLMs in the mental health domain is to provide education about mental health (see Figure 1) [8]. Although lagging behind the success of LLMs in the medical domain [9], there is evidence that LLMs are capable of generating accurate, helpful, and immediate mental health information. The psychological support with LLM (Psy-LLM), for example, is a domain-specific LLM designed to answer mental health questions [10]. Psy-LLM was pretrained with a data set of psychology articles, question-answer pairs from psychologists, and by crawling social media platforms. The model achieved moderate levels of helpfulness, fluency, relevance to the question asked, and logic based on human ratings of Psy-LLM responses.

The abilities of general-purpose LLMs to answer questions about mental health has also been evaluated. Sezgin et al [11] compared Google Search, GPT-4 (using ChatGPT), and LaMDA (using Bard [Google DeepMind]) responses to questions about postpartum depression relative to responses from an American College of Obstetricians and Gynecologists (ACOG) frequently asked questions document. Board-certified human physicians rated ChatGPT responses as more in line with ACOG responses than Bard or Google Search responses, and on average, ChatGPT responses were rated at near ceiling for clinical accuracy, scoring a 3.93 out of a possible 4. Importantly, however, general-purpose LLMs differ in their policies regarding the generation of medical or mental health advice. Bard's accuracy ratings were impacted by Bard's policy to advise consulting a health care provider when asked questions about mental health. This practice protects individuals from potential harm, though such responses received lower ratings of quality in this study.

LLM-generated answers to mental health questions may not be comparable to human-generated answers, however. It is critical for LLMs to meet or exceed human performance in order for LLMs to be trusted and to ease the demand for human providers. In the case of Psy-LLM and ChatGPT, there is evidence that responses to mental health and substance use questions fall short of human-generated responses in dimensions such as accuracy, quality, and alignment with evidence-based practice (EBP) [10,12].

Another way that LLMs may serve to educate is to support provider training. Barish et al [13] used ChatGPT to generate content and associated learning objectives for an online learning platform for behavioral health professionals. Researchers compared the time providers needed to write their own content versus the time needed to edit ChatGPT-generated content, finding that using ChatGPT improved provider efficiency by 37.5%. LLMs can also be leveraged to train providers to optimize interactions with their patients. As two examples, Chan and Li [14] developed a chatbot trained to mimic a patient capable of describing their mental health symptoms in colloquial terms, and Sharma et al [15] used artificial intelligence to coach peer support providers to increase empathetic responding. These approaches illustrate ways that LLMs can support provider training and potentially enhance provider efficacy without providers becoming reliant on LLMs for in the moment critical thinking or decision-making.

Figure 1. Potential opportunities for LLMs in mental health education. CBT: cognitive behavioral therapy; EST: empirically supported treatment; LLM: large language model.



Assessment

A second function of LLMs within the domain of mental health is to assess mental health symptoms, identify diagnoses, and track changes in mental well-being (see Figure 2). LLMs can at times predict mental health symptoms and diagnoses accurately. Ji et al [16] initially developed two domain-specific models, MentalBERT and MentalRoBERTa, pretrained on mental health information. Compared with existing models pretrained in different domains, specifically clinical notes and biomedicine, MentalBERT and MentalRoBERTa were generally better able to detect depression and suicidal ideation from social media posts (notably, these results were achieved with Bidirectional Encoder Representations From Transformers [BERT]-based models that represent early-genderation LLMs, with newer models and architectures demonstrating potential for even more advanced capabilities). LLMs such as Mental-Alpaca, a mental health domain–specific LLM, Med-PaLM 2, a medical domain–specific LLM, and ChatGPT, which is general-purpose, have also been shown to screen for possible depressive symptoms and suicide risk, with varying degrees of accuracy [17-20].

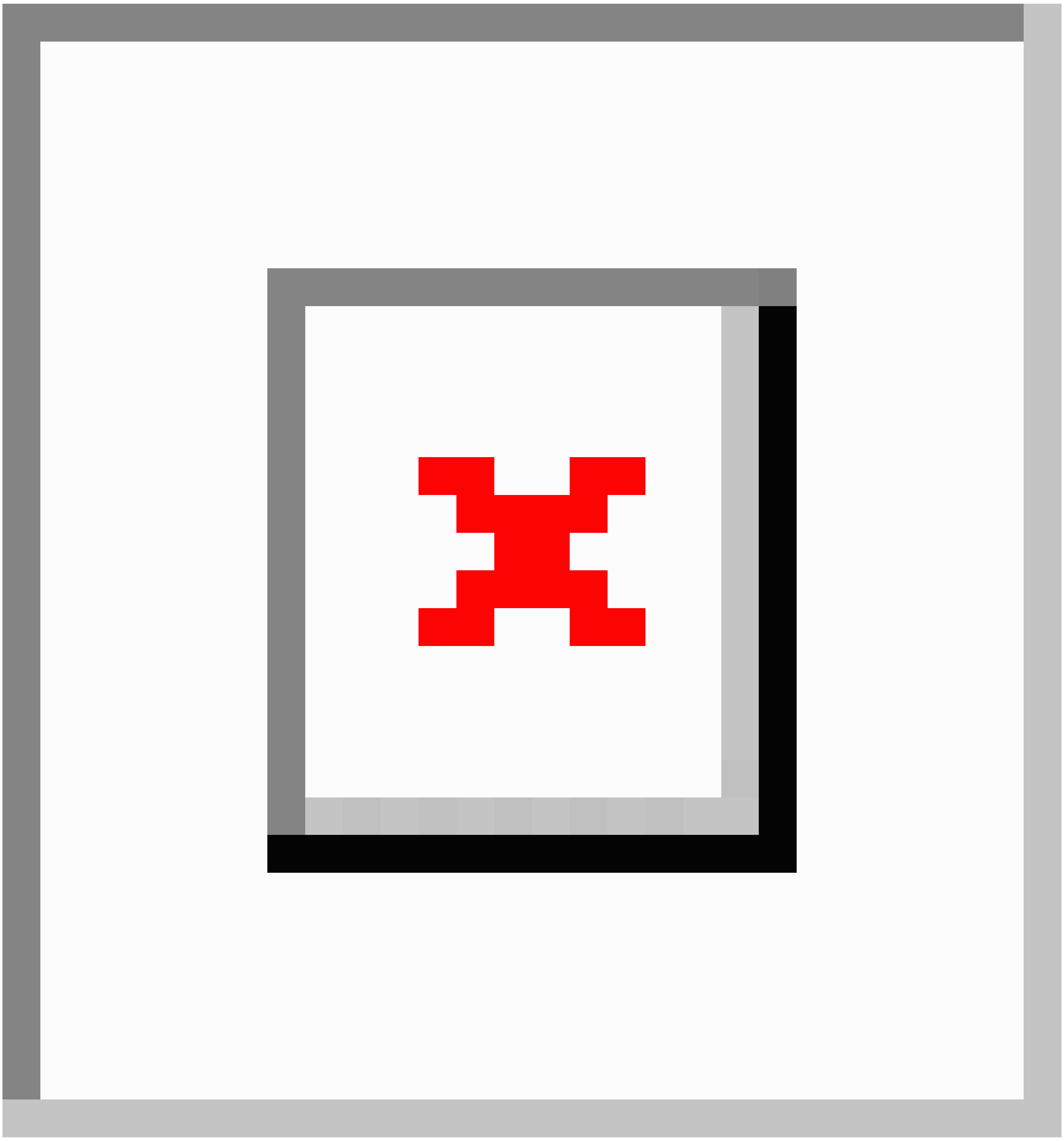
When it comes to predicting mental health diagnoses specifically, there is evidence that Med-PaLM 2 can do so accurately. When presented with a series of case studies from the American Psychiatric Association book of *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders, Fifth*

Edition) case examples [21], Med-PaLM 2 predicted the correct diagnosis 77.5% of the time, and performance increased to 92.5% when asked to specify the correct diagnostic category (eg, depressive disorder vs major depressive disorder) [20]. Similarly, when PaLM 2 was fine-tuned with medical domain data and optimized for differential diagnosis, the model was able to generate more appropriate and comprehensive lists of diagnoses than specialist medical doctors in response to challenging case studies, some of which involved psychiatric diagnoses [22].

LLM-predicted assessments do not, however, always match those of human mental health clinicians, suggesting that more work is needed before LLMs can engage in assessment without human oversight. In one study [23], four iterations of a case vignette [24] were presented to ChatGPT. Each vignette varied in levels of perceived burdensomeness and thwarted belongingness—two primary risk factors for suicide [25,26]. ChatGPT appropriately determined that the risk for suicidal ideation and suicide attempts was highest for the vignette with both high perceived burdensomeness and high thwarted belongingness, but it predicted lower suicide risk overall than did mental health professionals who reviewed the same vignettes. Med-PaLM 2 also at times does not achieve human clinician-level performance. The model predicted more severe posttraumatic stress disorder symptoms than human clinicians from clinical interview data, classified possible cases of posttraumatic stress disorder with high specificity (0.98) but low sensitivity (0.30), and the model only correctly predicted whether a case example had a comorbid diagnosis or diagnostic modifier 20% of the time [20].

In all the efforts described thus far, LLMs had been provided with information about symptoms and tasked with determining whether those symptoms indicated a possible mental health concern or diagnosis. LLMs also may be leveraged to ask the questions needed to screen for a mental health concern or to predict a mental health diagnosis. Chan and Li [14] developed a chatbot trained to engage in mental health assessment with patients. Compared with human psychiatrists, the chatbot displayed more empathy and asked more thorough questions about some symptoms (eg, sleep), but was less likely to rule out associated conditions.

Figure 2. Potential opportunities for LLMs in mental health assessment. LLM: large language model.

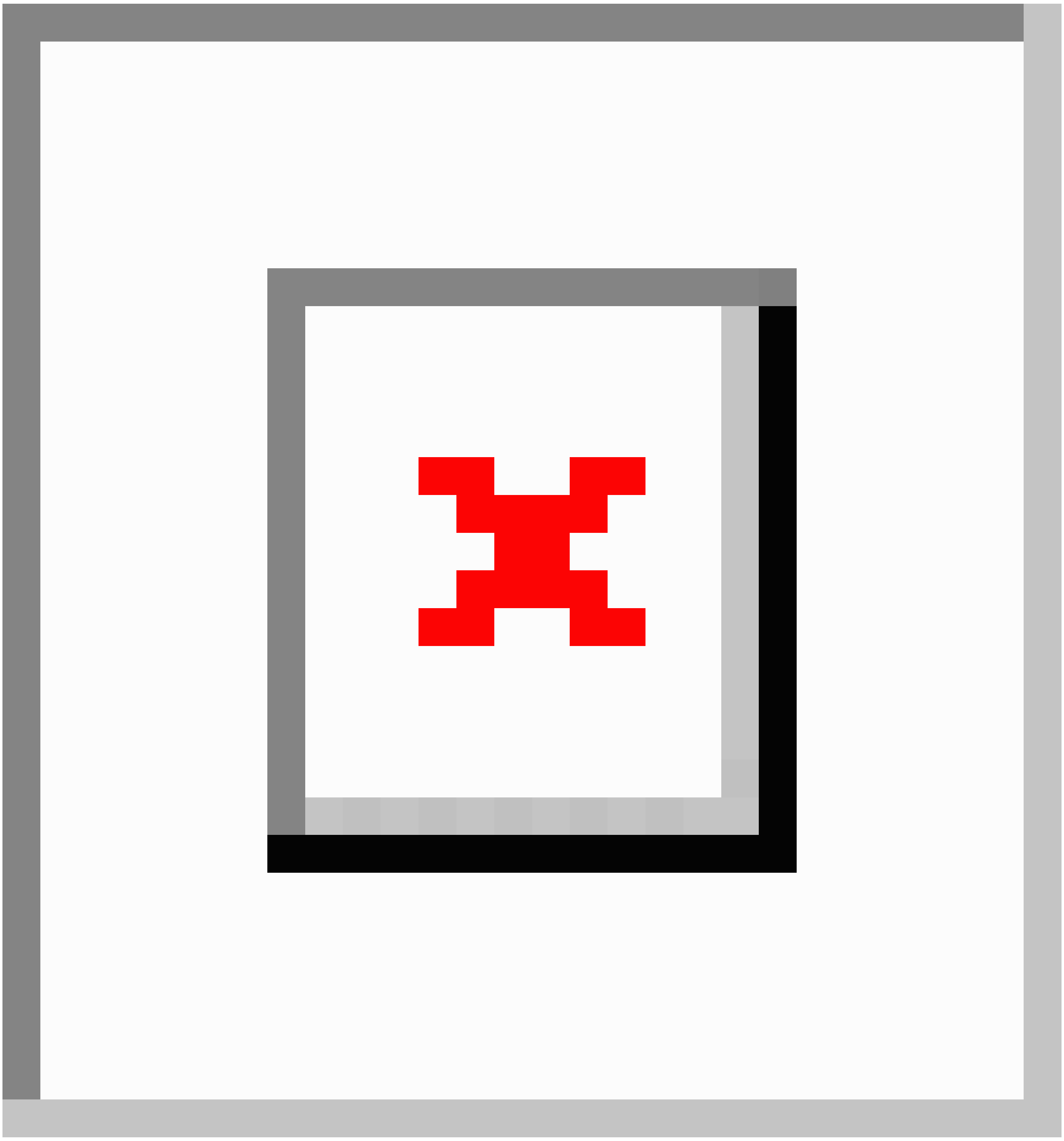


Intervention

A third opportunity for LLMs in the mental health domain is to implement mental health interventions (see Figure 3). To date, such efforts have largely focused on chatbots. Prominent chatbots, some of which are LLM-based, include Woebot [27], Wysa [28], Tess [29], Replika [30], Ellie [31], and Sibly [32]. Many of these chatbots were trained in empirically supported treatments such as cognitive behavioral therapy, dialectical behavior therapy, and motivational interviewing. There is initial evidence that such chatbots may be effective in reducing depressive and anxiety symptoms, as well as stress [33-36]. Additionally, research finds that chatbots can be trained to express empathy [37-39], provide nonjudgmental responses [40], and maintain therapeutic conversations [14] and that individuals can establish therapeutic rapport with chatbots [41].

Caution is warranted when using chatbots to deliver mental health interventions. To date, chatbots are not effective in treating all types of mental health distress [36] and at times have difficulty personalizing interventions [38], forget information (eg, that they had talked with someone previously) [37], and provide nontherapeutic and iatrogenic advice including encouraging substance use, dieting, and weight loss [40,42,43]. Also concerning is that chatbots do not consistently or adequately respond to suicide risk, at times being dismissive and neglecting to provide crisis resources or referals to human providers [38,44].

Figure 3. Potential opportunities for LLMs in mental health intervention. CBT: cognitive behavioral therapy; EBP: evidence-based practice; LLM: large language model.



Risks Associated With Mental Health LLMs

Overview

To maximize the positive impact of LLMs on mental health, LLM development, testing, and deployment must be done ethically and responsibly (see Textbox 1). This requires identification and evaluation of risks, taking preemptive steps to mitigate risks, and establishing plans to monitor for ongoing or new and unexpected risks [45,46]. It is also important to recognize that the risks associated with the use of LLMs for mental health support may differ across education, assessment, and intervention (see Table 1). Here, we highlight primary risks that largely cut across uses of LLMs for mental health-related tasks and identify potential steps that can be taken to mitigate these risks.

Textbox 1. Recommendations for responsible use of LLMs to support mental health.

- LLMs should only engage in mental health tasks when trained and shown to perform well.
- Mental health LLMs should advance mental health equity.
- Privacy or confidentiality should be paramount when LLMs operate to support mental health.
- Informed consent should be obtained when people engage with mental health LLMs.
- Mental health LLMs should respond appropriately to mental health risk.
- Mental health LLMs should only operate within the bounds of their competence.
- Mental health LLMs should be transparent and capable of explanation.
- Humans should provide oversight and feedback to mental health LLMs.

Table . Potential risks to people when LLMs[a] engage in mental health education, assessment, and intervention.[b]

	Mental health education	Mental health assessment	Mental health intervention
Perpetuate inequalities, disparities, and stigma	Medium	Higher	Higher
Unethical provision of mental health services			
Practice beyond the boundaries of competence	Lower	Higher	Higher
Neglect to obtain informed consent	Lower	Higher	Higher
Fail to preserve confidentiality or privacy	Lower	Higher	Higher
Build and maintain inappropriate levels of trust	Lower	Medium	Higher
Lack reliability	Lower	Higher	Higher
Generate inaccurate or iatrogenic output	Medium	Higher	Higher
Lack transparency or explainability	Lower	Medium	Medium
Neglect to involve humans	Lower	Medium	Higher

[a]LLM: large language model.

[b]This table aims to represent the potential for negative impacts on individuals should LLMs perform problematically across mental health education, assessment, and intervention. As depicted here, there may be additional risks to consider when LLMs engage in direct provision of mental health assessment and, perhaps especially, mental health intervention, relative to mental health education. As such, greater caution is warranted when considering use of LLMs for mental health interventions, and rigorous testing is needed before deployment of these LLMs. Risk estimates provided here are not meant to represent the risk associated with every possible LLM use case nor to minimize the negative impacts that are possible (eg, if LLMs were to perpetuate stigma when engaging in mental health education).

Perpetuating Inequalities, Disparities, and Stigma

There exists the risk that LLMs perpetuate inequities and stigma, further widening mental health disparities [47]. Mental health concerns are highly stigmatized [48], and there are disparities in who is at risk for mental health concerns, in who is diagnosed with mental health disorders, and with which mental health disorders people are diagnosed [49-51]. There are also inequities in who receives mental health care [52,53]. Much of the publicly available information and discourse about mental health contains inaccurate and stigmatizing information about mental health, and the existing research literature on mental health largely represents the perspectives of people who are White, are educated, are of high socioeconomic status, and speak English [54]. Far less information is available about the etiology of mental health concerns and effective assessments and interventions for populations that have been pushed to the margins. Training LLMs on existing data without appropriate safeguards and thoughtful human supervision and evaluation can, therefore, lead to problematic generation of biased content and disparate model performance for different groups [45,55-57] (of note, however, there is some evidence that clinicians perceive less bias in LLM-generated responses [58] relative to clinician-generated responses, suggesting that LLMs may have the potential to reduce bias compared to human clinicians).

LLMs should disseminate accurate, destigmatizing information about mental health and be trained to identify and combat stigma and discrimination. To do so, models need to be fine-tuned and evaluated for the mental health domain. Training models with data representative of the diverse populations being served is helpful, but new types of bias, such as semantic biases, may arise in LLMs [59]. Opportunities to train models to identify and exclude toxic and discriminatory language should be explored, both during the training of the underlying foundation models and during the domain-specific fine-tuning (see Keeling [59] for a discussion of the trade-offs of data filtration in this context) [45]. If LLMs perform differently for different groups or generate problematic or stigmatizing language during testing, additional model fine-tuning is required prior to deployment. Individuals developing LLMs should be transparent about the limitations of the training data, the approaches to data filtration and fine-tuning, and the populations for whom LLM performance has not been sufficiently demonstrated.

There is also hope that LLMs can be scaled to increase people's access to mental health information, assessment, and treatment. LLMs have the potential to support delivery of mental health interventions in regions where access to mental health providers is limited and where significant barriers (eg, cost) exist. They can additionally help to personalize treatments to better fit people's unique preferences, interests, identities, and language, hopefully improving treatment outcomes. LLMs may support increased access through more direct provision of mental health services, or LLMs can aid the expansion of the mental health workforce, training novice providers and community members in EBP at scale. There will undoubtedly be challenges in implementing and scaling LLMs globally. Revising and testing implementation frameworks for this new and evolving context and engagement in thoughtful public health and industry partnerships could all increase the likelihood that when mental health LLMs are scaled globally, implementation is sustained and best supports the populations most in need.

Failing to Provide Mental Health Services Ethically

A second risk is that LLMs will engage in unethical practices. When human mental health providers behave unethically, harm is done to patients and public trust is eroded [60]. LLMs will similarly do harm if they are not designed and implemented in consideration of and are not consistent with relevant ethical principles and standards when operating in the domain of mental health. Core ethical principles in the health care context include beneficence, nonmaleficence, justice, and autonomy [61]. Next, we highlight additional standards of ethical professional conduct that should apply when LLMs engage in mental health service provision (see the American Psychological Association Ethical Principles of Psychologists and Code of Conduct for parallel ethical principles and standards).

LLMs should operate within the boundaries of their competence and only engage in mental health tasks they have rigorously been proven to accomplish well. LLM developers should clearly communicate the limits and relevant evaluation results of LLMs, education should be provided to individuals about when it is and is not appropriate to use LLMs, and LLMs should withhold output when they are not competent in a task. LLM competence should be assessed and maintained over time. When competence is lacking in a certain domain, the LLM should no longer be deployed until the needed competence is gained (eg, via retraining and fine-tuning models with human validation).

Individuals should provide informed consent when interacting with mental health LLMs. They should be fully informed about the nature of mental health services they will receive and what role LLMs will have in that service. Information presented to individuals to help make decisions about consent should be understandable and include the possible risks and benefits of engaging with LLMs. Individuals should have the ability to choose not to consent to the use of LLMs in the direct provision of their mental health care, as well as the ability to withdraw their consent and opt out of the use of LLMs even if consent was initially given. As LLMs become further integrated into health care contexts, care should be taken to ensure that clients' decisions to opt out of LLM involvement or to confine LLM involvement to less direct (eg, administrative) tasks do not limit their access to mental health care.

Confidentiality should be protected when individuals interact with LLMs to support their mental health. Individuals should be clearly informed about expectations for confidentiality. This should include information about the limits of confidentiality (eg, in the case of imminent risk for suicide), the foreseeable uses of information generated through engagement with LLMs, where and how their data are stored, and whether it is possible to delete their data. Policies related to data security should be strict and in line with relevant mental health data protection regulations [34]. Solutions such as developing on-device storage that does not require transmission of personal data [62] or systems with robust cloud-based encryption, pursuing LLMs that support compliance with relevant data protection laws (eg, Health Insurance Portability and Accountability Act [HIPAA]), and responsibly aggregating and deidentifying mental health data to fine-tune and test models all help to protect confidentiality.

Human mental health providers establish trusting relationships with those with whom they work and are obligated to ensure that the nature of the trusting provider-patient relationship does not lead to exploitation or harm. Appropriate trust is built through effective mental health assessment and treatment and, perhaps even more crucially, ethical practice. Trust should be evaluated through feedback from individuals engaged with LLMs. If and when trust is broken, this should be acknowledged and work should be done to repair trust. On the other hand, people may trust LLMs more than is warranted because of LLMs' ability to produce humanlike natural language and to be trained to express emotion and empathy (this may especially be the case for individuals experiencing mental health concerns such as anxiety [63]) [64]. Unearned trust can have consequences, leading people to disclose personal information or trust content generated by LLMs even when it is not accurate. Education should be provided about the limits of LLMs and individuals should be cautioned against blanket trust in these models.

Insufficient Reliability

A third risk is that LLMs will not generate reliable or consistent output. When prompted to complete the same task or provide an answer to the same question multiple times, LLMs at times produce different responses [46,65]. Varied and creative output is a benefit of LLMs; however, the underlying response should be consistent even when articulated in different ways. Take for example an LLM repeatedly presented with a client's description of depressive symptoms. The LLM should reliably reach the conclusion that the client meets the criteria for major depressive disorder even if this diagnostic conclusion is communicated to the client using different phrasing. Issues of low reliability of LLMs can erode trust and increase the possibility of harm, including leading some individuals to be misdiagnosed or to pursue treatments that are not best suited to their mental health concern.

LLM reliability should be measured and enhanced. Prompting approaches may help to improve LLM reliability. Self-consistency [66] and ensemble refinement [9] are strategies that sample multiple model answers to arrive at a more consistent response, improving model reliability [9]. Grounding models in data other than linguistic descriptions of symptoms (eg, objective behavioral or physiological signals) is another way of reducing variability in LLM performance, as words alone may not fully capture all of the necessary information to complete a given mental health task [67]. Finally, LLMs should not be deployed until they exceed prespecified thresholds of adequate reliability.

Inaccuracy

LLMs risk producing inaccurate information about mental health [46,68]. If LLMs are trained on data that contain inaccurate or outdated information, iatrogenic treatment options, or biased representations of mental health, that information can be reproduced by LLMs [45]. An additional consideration is that accuracy of LLM outputs has multiple dimensions and is not as simple to evaluate as answers to multiple-choice questions. Accuracy can be a function of how factual an answer is, how specific it is, or how devoid of irrelevant information it is. Generating inaccurate mental health information may be more damaging than no information, especially when it may be difficult for an individual to detect inaccuracies or inconsistencies (eg, about a complex mental health diagnosis).

Standards for accuracy should be defined a priori and should be high. When thresholds for LLM accuracy are not met, the risk of harm is too high and LLMs should not generate output. The accuracy of LLMs depends on the quality of data the model is trained and fine-tuned on [47,69,70]. LLMs should be adapted to the domain of mental health; models fine-tuned on mental health data perform better than models trained on non-domain-specific data [42] or general medical domains [16]. When data are limited, it is recommended that smaller but more variable data sets be prioritized over a larger single data set [19]). Training data should be highly curated, be grounded in authoritative and trusted sources, be specific to evidence-based health care, and represent diverse populations [46,58]. In mental health, the nature of consensus is continuing to evolve, and the amount of data available is continuing to increase, which should be taken into account when considering whether to further fine-tune models. Strategies such as implementing a Retrieval Augmented Generation system, in which LLMs are given access to an external database of up-to-date, quality-verified information to incorporate in the generation process, may help to improve accuracy and enable links to sources while also maintaining access to updated information. Accuracy of LLMs should be monitored over time to ensure that model accuracy improves and does not deteriorate with new information [45].

Measuring the accuracy of mental health LLMs is complex. It is not sufficient for models to merely outperform previous models. Rather, performance of LLMs should be compared with the performance of human clinicians, both of which should be compared against gold-standard, evidence-based care. When LLMs are tasked with mental health evaluation, their ability to predict scores on reliable and valid mental health assessments should be tested, and LLMs should meet human clinician performance in diagnostic accuracy. When LLMs are tasked with aiding mental health intervention delivery, their ability to detect, support, and engage in EBP is critical. Additional criteria to consider when evaluating the accuracy of LLMs include the level of agreement between human clinicians and LLMs, metrics of effect size rather than only statistical significance, and the balance of sensitivity and specificity in making diagnostic predictions.

LLMs should communicate confidence in the accuracy of generated output and limit or withhold output when confidence is lacking [58]. As an example, Med-PaLM 2's accuracy improved when results were weighted based on confidence scores and when a cutoff threshold was set for confidence [20]. Communicating confidence in generated output and withholding output when confidence is low both help to enhance transparency and trust in LLMs' ability to perform on mental health tasks and to limit potential harms associated with generating inaccurate information.

Prompt fine-tuning can boost LLM accuracy [9,19,58]. When applied to mental health, instruction fine-tuning improved performance of Mental-Alpaca relative to zero-shot and few-shot prompting and allowed Mental-Alpaca to reach a performance level across multiple mental health tasks (eg, identifying stress and classifying individuals as depressed or not based on Reddit posts) similar to that of Mental-RoBERTa, a task-specific model [19]. Prompting to concentrate on the emotional clues in text was also shown to improve ChatGPT performance on a variety of mental health-related tasks [71]. Conversely, however, instruction prompt fine-tuning can also increase inaccurate or inappropriate content [55]; thus, LLMs should continue to be evaluated for accuracy at all stages of prompt tuning.

Lack of Transparency and Explainability

LLMs risk generating output without being able to explain how they came to the decisions they did or without being able to identify the source of information used to generate the output [72]. There remains much that is not known about how LLMs generate reasoning for their responses and how sensitive these reasons are to context and prompting. It should be apparent when information is generated using LLMs, how LLMs were developed and tested, and whether LLMs are general-purpose or fine-tuned for the domain of mental health [46,58,68]. Additional steps to enhance transparency include explicitly telling individuals to exercise caution when interpreting or acting on LLM output and being clear about the bounds of LLMs' competence [39].

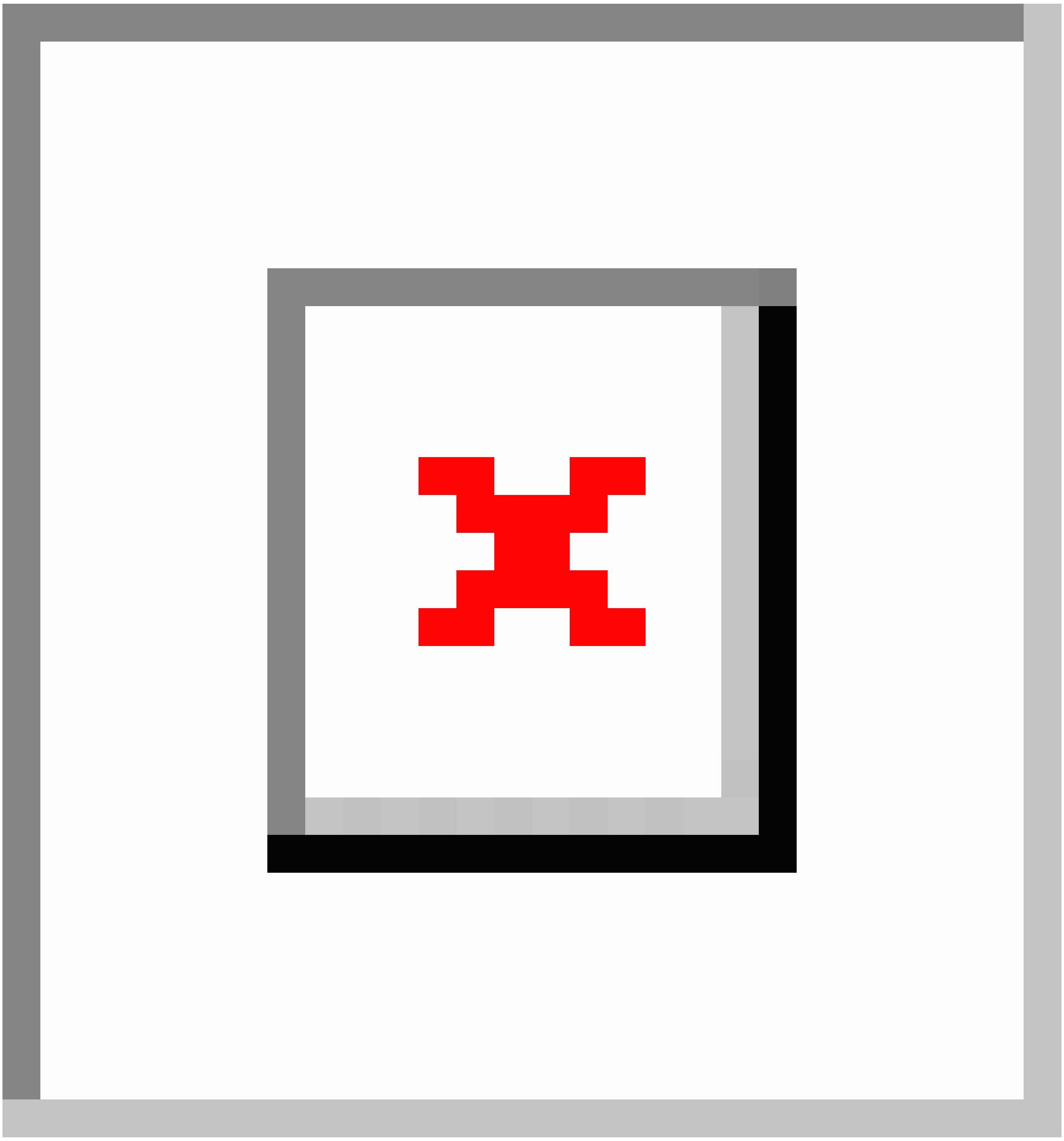
Explainability, one aspect of transparency, was identified as a key priority by individuals engaged in mental health LLMs [39]. If asked to explain why they decided on a mental health diagnostic prediction or intervention, LLMs should explain what information was used to come to that decision. ChatGPT has been shown to be able to explain why an individual was classified as experiencing stress or depressive symptoms [71], and Med-PaLM 2 communicated why it predicted a particular symptom score and diagnosis [20]. Although LLMs are capable of producing plausible explanations through techniques such as chain-of-thought reasoning [73], more research is needed to ensure that explanations are internally consistent. Explainability is perhaps especially beneficial in the domain of mental health, as part of mental health assessment and intervention is communicating results of an evaluation or justification for an intervention to patients.

Neglecting to Involve Humans

There are risks associated with LLMs providing anonymous mental health services. Unlike mental health apps, where content can be highly curated, the content generated by LLMs is unpredictable. This makes interacting with LLMs more engaging, more appealing, and perhaps also more humanlike. However, it also increases the risk that LLMs may produce harmful or nontherapeutic content when tasked with independently providing mental health services. Legal and regulatory frameworks are needed to protect individuals' safety and mental health when interacting with LLMs, as well as to clarify clinician liability when using LLMs to support their work or to clarify the liability of individuals and companies who develop these LLMs. There are ongoing discussions regarding the regulation of LLMs in medicine [74-76] that can inform how LLMs can support mental health while limiting the potential for harm and liability.

Humans should be actively involved in all stages of mental health LLM development, testing, and deployment. For mental health LLMs to be effective, rigorous, and ongoing, human supervision and input are needed (see Figure 4) [46]. Reinforcement learning through human feedback can improve model accuracy and uncover problematic LLM responses [14,42]. This feedback should be obtained from individuals who reflect the diverse populations the LLM aims to help, including members of the public, patients, and human clinicians [9,14,34,58,68,77,78]. Their input should be leveraged to identify and correct biases, to ensure generated content is inclusive, culturally appropriate, and accurate, and to reduce the likelihood of harm. Particularly important is prioritizing the perspectives of individuals at heightened risk for mental health concerns (eg, sexual and gender minorities) and individuals with lived experience with mental health concerns. These individuals should play a central role in co-defining the role LLMs will play in mental health care and in co-designing tools that leverage LLMs. Practically, use cases should focus on opportunities to support and augment provider care. As just one example, LLMs may have a role in suggesting language used in clinical notes, but clinicians should have the final say in whether they adopt those suggestions or not.

Figure 4. Examples of human involvement across all stages of LLM development through deployment and evaluation. LLM: large language model.



Conclusions

The need for mental health services is pressing, and the potential of LLMs to expand access to information about mental health and to mental health care is great. LLMs are advancing rapidly and have been applied across mental health education, assessment, and intervention. Especially promising is the potential for LLMs to provide mental health education and assessment—tasks that are well aligned with LLM strengths. LLMs have made exceptional progress in related tasks such as answering medical questions and assessing medical conditions, reaching and in some cases exceeding the performance of human clinicians. Greater caution is warranted when applying LLMs to mental health intervention, but there is also cause for optimism that LLMs could eventually help to support or augment human provision of mental health treatments. Additional research is needed in testing LLMs' ability to deliver or train providers in empirically supported treatments, to responsibly adapt approaches for youth and marginalized populations, to build appropriate rapport, and to detect risk for high-acuity mental health concerns for progress to be made in these areas.

Critical to effectively engaging in mental health care tasks is fine-tuning LLMs specifically for the domain of mental health and the prioritization of equity, safety, EBP, and confidentiality. No widely used, general-purpose LLM has been fine-tuned for mental health, trained on evidence-based mental health content, or sufficiently tested on mental health-related tasks. When LLMs are developed specifically for mental health, tested to ensure adherence with EBP, and aligned with the goals of people with lived experience with mental health concerns and those who have expertise in mental health care, there is great hope that they will expand access to evidence-based mental health information and services. Investing in developing, testing, and deploying mental health LLMs responsibly has the potential to finally reverse rising global mental health rates and to improve the mental health of the millions of people in need of mental health support.

Acknowledgments

We acknowledge and thank Michael Howell, MD, MPH; Bakul Patel, MSEE, MBA; Matthew Thompson, DPhil, MPH; Joseph Dooley, MPA; and David Steiner, MD, PhD, for reviewing and providing helpful feedback on this paper.

Conflicts of Interest

RAS, MJM, DJM, and MJB are employees of Google and receive monetary compensation from Google and equity in Google's parent company, Alphabet. HRL and SBR are employees of, and receive compensation from, Magnit and are contracted for work at Google. In addition, MJB is a shareholder in Meeno Technologies, Inc and The Orange Dot (Headspace Health). RAS is a shareholder in Lyra Health and Trek Health, and she consults with Understood.

References

1. McGrath JJ, Al-Hamzawi A, Alonso J, et al. Age of onset and cumulative risk of mental disorders: a cross-national analysis of population surveys from 29 countries. Lancet Psychiatry 2023 Sep;10(9):668-681. [doi: 10.1016/S2215-0366(23)00193-1] [Medline: 37531964]
2. World mental health report: transforming mental health for all. : World Health Organization; 2022 URL: https://www.who.int/publications/i/item/9789240049338 [accessed 2024-07-18]
3. 2022 National Healthcare Quality and Disparities Report. : Agency for Healthcare Research and Quality; 2022 URL: https://www.ahrq.gov/research/findings/nhqrdr/nhqdr22/index.html [accessed 2024-07-18]
4. Wainberg ML, Scorza P, Shultz JM, et al. Challenges and opportunities in global mental health: a research-to-practice perspective. Curr Psychiatry Rep 2017 May;19(5):28. [doi: 10.1007/s11920-017-0780-z] [Medline: 28425023]
5. Mental health by the numbers. National Alliance on Mental Illness. 2023. URL: https://www.nami.org/about-mental-illness/mental-health-by-the-numbers/ [accessed 2024-07-29]
6. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
7. Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: a survey. arXiv. Preprint posted online on Dec 18, 2023. [doi: 10.48550/arXiv.2312.10997]
8. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
9. Singhal K, Tu T, Gottweis J, et al. Towards expert-level medical question answering with large language models. arXiv. Preprint posted online on May 16, 2023. [doi: 10.48550/arXiv.2305.09617]
10. Lai T, Shi Y, Du Z, et al. Psy-LLM: scaling up global mental health psychological services with AI-based large language models. arXiv. Preprint posted online on Sep 1, 2023. [doi: 10.48550/arXiv.2307.11991]
11. Sezgin E, Chekeni F, Lee J, Keim S. Clinical accuracy of large language models and Google search responses to postpartum depression questions: cross-sectional study. J Med Internet Res 2023 Sep 11;25:e49240. [doi: 10.2196/49240] [Medline: 37695668]
12. Spallek S, Birrell L, Kershaw S, Devine EK, Thornton L. Can we use ChatGPT for mental health and substance use education? Examining its quality and potential harms. JMIR Med Educ 2023 Nov 30;9:e51243. [doi: 10.2196/51243] [Medline: 38032714]
13. Barish G, Marlotte L, Drayton M, Mogil C, Lester P. Automatically enriching content for a behavioral health learning management system: a first look. Presented at: The 9th World Congress on Electrical Engineering and Computer Systems and Science; Aug 3-5, 2023; London, United Kingdom. [doi: 10.11159/cist23.125]
14. Chan C, Li F. Developing a natural language-based AI-chatbot for social work training: an illustrative case study. China J Soc Work 2023 May 4;16(2):121-136. [doi: 10.1080/17525098.2023.2176901]
15. Sharma A, Lin IW, Miner AS, Atkins DC, Althoff T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nat Mach Intell 2023;5(1):46-57. [doi: 10.1038/s42256-022-00593-2]
16. Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E. MentalBERT: publicly available pretrained language models for mental healthcare. arXiv. Preprint posted online on Oct 29, 2021. [doi: 10.48550/arXiv.2110.15621]
17. Lamichhane B. Evaluation of ChatGPT for NLP-based mental health applications. arXiv. Preprint posted online on Mar 28, 2023. [doi: 10.48550/arXiv.2303.15727]

18. Amin MM, Cambria E, Schuller BW. Will affective computing emerge from foundation models and general AI? A first evaluation on ChatGPT. arXiv. Preprint posted online on Mar 3, 2023. [doi: 10.48550/arXiv.2303.03186]
19. Xu X, Yao B, Dong Y, et al. Mental-LLM: leveraging large language models for mental health prediction via online text data. arXiv. Preprint posted online on Aug 16, 2023. [doi: 10.48550/arXiv.2307.14385]
20. Galatzer-Levy IR, McDuff D, Natarajan V, Karthikesalingam A, Malgaroli M. The capability of large language models to measure psychiatric functioning. arXiv. Preprint posted online on Aug 3, 2023. [doi: 10.48550/arXiv.2308.01834]
21. Barnhill JW. DSM-5-TR® clinical cases. Psychiatry online. URL: https://dsm.psychiatryonline.org/doi/book/10.1176/appi.books.9781615375295 [accessed 2024-07-18]
22. McDuff D, Schaekermann M, Tu T, et al. Towards accurate differential diagnosis with large language models. arXiv. Preprint posted online on Nov 30, 2023. [doi: 10.48550/arXiv.2312.00164]
23. Elyoseph Z, Levkovich I. Beyond human expertise: the promise and limitations of ChatGPT in suicide risk assessment. Front Psychiatry 2023 Aug;14:1213141. [doi: 10.3389/fpsyt.2023.1213141] [Medline: 37593450]
24. Levi-Belz Y, Gamliel E. The effect of perceived burdensomeness and thwarted belongingness on therapists' assessment of patients' suicide risk. Psychother Res 2016 Jul;26(4):436-445. [doi: 10.1080/10503307.2015.1013161] [Medline: 25751580]
25. Joiner T. Why People Die by Suicide: Harvard University Press; 2007.
26. Van Orden KA, Witte TK, Cukrowicz KC, Braithwaite SR, Selby EA, Joiner TE. The interpersonal theory of suicide. Psychol Rev 2010 Apr;117(2):575-600. [doi: 10.1037/a0018697] [Medline: 20438238]
27. Darcy A, Beaudette A, Chiauzzi E, et al. Anatomy of a Woebot® (WB001): agent guided CBT for women with postpartum depression. Expert Rev Med Devices 2022 Apr;19(4):287-301 Retracted inExpert Rev Med Devices. 2023;20(11):989. [doi: 10.1080/17434440.2023.2267389] [Medline: 37801290]
28. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth 2018 Nov 23;6(11):e12106. [doi: 10.2196/12106] [Medline: 30470676]
29. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. JMIR Ment Health 2018 Dec 13;5(4):e64. [doi: 10.2196/mental.9782] [Medline: 30545815]
30. Murphy M, Templin J. Our story. Replika. 2021. URL: https://replika.ai/about/story [accessed 2023-08-20]
31. Kim H, Yang H, Shin D, Lee JH. Design principles and architecture of a second language learning chatbot. Lang Learn Technol 2022;26:1-18 [FREE Full text]
32. Wilbourne P, Dexter G, Shoup D. Research driven: Sibly and the transformation of mental health and wellness. Presented at: Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare; May 21-24, 2018; New York, NY p. 389-391. [doi: 10.1145/3240925.3240932]
33. Denecke K, Abd-Alrazaq A, Househ M. Artificial intelligence for chatbots in mental health: opportunities and challenges. In: Househ M, Borycki E, Kushniruk A, editors. Multiple Perspectives on Artificial Intelligence in Healthcare: Opportunities and Challenges: Springer International Publishing; 2021:115-128. [doi: 10.1007/978-3-030-67303-1]
34. Omarov B, Zhumanov Z, Gumar A, Kuntunova L. Artificial intelligence enabled mobile chatbot psychologist using AIML and cognitive behavioral therapy. IJACSA 2023;14(6). [doi: 10.14569/IJACSA.2023.0140616]
35. Pham KT, Nabizadeh A, Selek S. Artificial intelligence and chatbots in psychiatry. Psychiatr Q 2022 Mar;93(1):249-253. [doi: 10.1007/s11126-022-09973-8] [Medline: 35212940]
36. Abd-Alrazaq AA, Rababeh A, Alajlani M, Bewick BM, Househ M. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. J Med Internet Res 2020 Jul;22(7):e16021. [doi: 10.2196/16021] [Medline: 32673216]
37. Brocki L, Dyer GC, Gładka A, Chung NC. Deep learning mental health dialogue system. Presented at: 2023 IEEE International Conference on Big Data and Smart Computing (BigComp); Feb 13-16, 2023; Jeju, Korea p. 395-398.
38. Martinengo L, Lum E, Car J. Evaluation of chatbot-delivered interventions for self-management of depression: content analysis. J Affect Disord 2022 Dec;319:598-607. [doi: 10.1016/j.jad.2022.09.028] [Medline: 36150405]
39. You Y, Tsai CH, Li Y, Ma F, Heron C, Gui X. Beyond self-diagnosis: how a chatbot-based symptom checker should respond. ACM Trans Comput-Hum Interact 2023 Aug 31;30(4):1-44. [doi: 10.1145/3589959]
40. Ma Z, Mei Y, Su Z. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. AMIA Annu Symp Proc 2024 Jan 11;2023:1105-1114. [Medline: 38222348]
41. Lee J, Lee JG, Lee D. Influence of rapport and social presence with an AI psychotherapy chatbot on users' self-disclosure. SSRN. Preprint posted online on Mar 22, 2022. [doi: 10.2139/ssrn.4063508]
42. Das A, Selek S, Warner AR, et al. Conversational bots for psychotherapy: a study of generative transformer models using domain-specific dialogues. In: Demner-Fushman D, Cohen KB, Ananiadou S, Tsujii J, editors. Proceedings of the 21st Workshop on Biomedical Language Processing: Association for Computational Linguistics; 2022:285-297. [doi: 10.18653/v1/2022.bionlp-1.27]
43. Demner-Fushman D, Ananiadou S, Cohen KB, editors. The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks: Association for Computational Linguistics; 2023.

44. Heston TF. Evaluating risk progression in mental health chatbots using escalating prompts. medRxiv. Preprint posted online on Sep 12, 2023. [doi: 10.1101/2023.09.10.23295321]
45. Weidinger L, Mellor J, Rauh M, et al. Ethical and social risks of harm from language models. arXiv. Preprint posted online on Dec 8, 2021. [doi: 10.48550/arXiv.2112.04359]
46. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. eBioMedicine 2023 Apr;90:104512. [doi: 10.1016/j.ebiom.2023.104512] [Medline: 36924620]
47. Koutsouleris N, Hauser TU, Skvortsova V, De Choudhury M. From promise to practice: towards the realisation of AI-informed mental health care. Lancet Digital Health 2022 Nov;4(11):e829-e840. [doi: 10.1016/S2589-7500(22)00153-4]
48. Sickel AE, Seacat JD, Nabors NA. Mental health stigma update: a review of consequences. Adv Ment Health 2014 Dec;12(3):202-215. [doi: 10.1080/18374905.2014.11081898]
49. Alegría M, Green JG, McLaughlin KA, Loder S. Disparities in child and adolescent mental health and mental health services in the U.S. William T. Grant Foundation. 2015 Mar. URL: https://wtgrantfoundation.org/wp-content/uploads/2015/09/Disparities-in-Child-and-Adolescent-Mental-Health.pdf [accessed 2024-07-18]
50. Primm AB, Vasquez MJT, Mays RA. The role of public health in addressing racial and ethnic disparities in mental health and mental illness. Prev Chronic Dis 2010 Jan;7(1):A20. [Medline: 20040235]
51. Schwartz RC, Blankenship DM. Racial disparities in psychotic disorder diagnosis: a review of empirical literature. World J Psychiatry 2014 Dec;4(4):133. [doi: 10.5498/wjp.v4.i4.133] [Medline: 25540728]
52. McGuire TG, Miranda J. New evidence regarding racial and ethnic disparities in mental health: policy implications. Health Aff (Millwood) 2008 Mar;27(2):393-403. [doi: 10.1377/hlthaff.27.2.393] [Medline: 18332495]
53. Snowden LR, Cheung FK. Use of inpatient mental health services by members of ethnic minority groups. Am Psychol 1990 Mar;45(3):347-355. [doi: 10.1037//0003-066x.45.3.347] [Medline: 2310083]
54. Henrich J, Heine SJ, Norenzayan A. Beyond WEIRD: towards a broad-based behavioral science. Behav Brain Sci 2010 Jun;33(2-3):111-135. [doi: 10.1017/S0140525X10000725]
55. Lin I, Njoo L, Field A, et al. Gendered mental health stigma in masked language models. arXiv. Preprint posted online on Oct 27, 2022. [doi: 10.48550/arXiv.2210.15144]
56. Liu Y, et al. Trustworthy LLMs: a survey and guideline for evaluating large language models' alignment. arXiv. Preprint posted online on Aug 10, 2023. [doi: 10.48550/arXiv.2308.05374]
57. Straw I, Callison-Burch C. Artificial intelligence in mental health and the biases of language based models. PLoS One 2020 Dec;15(12):e0240376. [doi: 10.1371/journal.pone.0240376] [Medline: 33332380]
58. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
59. Keeling G. Algorithmic bias, generalist models, and clinical medicine. arXiv. Preprint posted online on May 6, 2023. [doi: 10.48550/arXiv.2305.04008]
60. Koocher GP, Keith-Spiegel P. Ethics in Psychology and the Mental Health Professions: Standards and Cases: Oxford University Press; 2008.
61. Varkey B. Principles of clinical ethics and their application to practice. Med Princ Pract 2021 Feb;30(1):17-28. [doi: 10.1159/000509119] [Medline: 32498071]
62. Rajagopal A, Nirmala V, Andrew J, Arun M. Novel AI to avert the mental health crisis in COVID-19: novel application of GPT2 in cognitive behaviour therapy. Research Square. Preprint posted online on Apr 1, 2021. [doi: 10.21203/rs.3.rs-382748/v1]
63. Gratch J, Lucas G. Rapport between humans and socially interactive agents. In: Lugrin B, Pelachaud C, Traum D, editors. The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition: Association for Computing Machinery; 2021:433-462. [doi: 10.1145/3477322.3477335]
64. McDuff D, Czerwinski M. Designing emotionally sentient agents. Commun ACM 2018 Nov 20;61(12):74-83. [doi: 10.1145/3186591]
65. Lundin RM, Berk M, Østergaard SD. ChatGPT on ECT: can large language models support psychoeducation? J ECT 2023 Sep 1;39(3):130-133. [doi: 10.1097/YCT.0000000000000941] [Medline: 37310145]
66. Wang B, Min S, Deng X, et al. Towards understanding chain-of-thought prompting: an empirical study of what matters. arXiv. Preprint posted online on Dec 20, 2022. [doi: 10.48550/arXiv.2212.10001]
67. Glenberg AM, Havas D, Becker R, Rinck M. Grounding language in bodily states: the case for emotion. In: Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking: Cambridge University Press; 2005:115-128. [doi: 10.1017/CBO9780511499968]
68. Zhong Y, Chen YJ, Zhou Y, Lyu YAH, Yin JJ, Gao YJ. The artificial intelligence large language models and neuropsychiatry practice and research ethic. Asian J Psychiatr 2023 Jun;84:103577. [doi: 10.1016/j.ajp.2023.103577] [Medline: 37019020]
69. Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model AI chatbots require approval as medical devices. Nat Med 2023 Oct;29(10):2396-2398. [doi: 10.1038/s41591-023-02412-6] [Medline: 37391665]
70. Singh OP. Artificial intelligence in the era of ChatGPT - opportunities and challenges in mental health care. Indian J Psychiatry 2023 Mar;65(3):297-298. [doi: 10.4103/indianjpsychiatry.indianjpsychiatry_112_23] [Medline: 37204980]

71. Yang K, Ji S, Zhang T, Xie Q, Kuang Z, Ananiadou S. Towards interpretable mental health analysis with large language models. arXiv. Preprint posted online on Oct 11, 2023. [doi: 10.48550/arXiv.2304.03347]
72. Balasubramaniam N, Kauppinen M, Rannisto A, Hiekkanen K, Kujala S. Transparency and explainability of AI systems: from ethical guidelines to requirements. Inf Softw Technol 2023 Jul;159:107197. [doi: 10.1016/j.infsof.2023.107197]
73. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Koyejo S, Mohamed S, Agarwal A, editors. Advances in Neural Information Processing Systems: Curran Associates, Inc; 2022, Vol. 35:24824-24837.
74. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 6;6(1):120. [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
75. Ong JCL, Chang SYH, William W, et al. Ethical and regulatory challenges of large language models in medicine. Lancet Digit Health 2024 Jun;6(6):e428-e432. [doi: 10.1016/S2589-7500(24)00061-X] [Medline: 38658283]
76. Minssen T, Vayena E, Cohen IG. The challenges for regulating medical use of ChatGPT and other large language models. JAMA 2023 Jul 25;330(4):315-316. [doi: 10.1001/jama.2023.9651] [Medline: 37410482]
77. van Heerden AC, Pozuelo JR, Kohrt BA. Global mental health services and the impact of artificial intelligence-powered large language models. JAMA Psychiatry 2023 Jul 1;80(7):662-664. [doi: 10.1001/jamapsychiatry.2023.1253] [Medline: 37195694]
78. Cabrera J, Loyola MS, Magaña I, Rojas R. Ethical dilemmas, mental health, artificial intelligence, and LLM-based chatbots. In: Bioinformatics and Biomedical Engineering: Springer Nature Switzerland; 2023:313-326. [doi: 10.1007/978-3-031-34960-7_2]

Abbreviations

ACOG: American College of Obstetricians and Gynecologists
BERT: Bidirectional Encoder Representations From Transformers
***DSM-5*:** *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*
EBP: evidence-based practice
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
Psy-LLM: psychological support with large language model

Edited by J Torous; submitted 13.04.24; peer-reviewed by BMM Kious, JC Ling Ong, TF Heston; revised version received 31.05.24; accepted 01.06.24; published 29.07.24.

Please cite as:
Lawrence HR, Schneider RA, Rubin SB, Matarić MJ, McDuff DJ, Jones Bell M
The Opportunities and Risks of Large Language Models in Mental Health
JMIR Ment Health 2024;11:e59479
URL: https://mental.jmir.org/2024/1/e59479
doi:10.2196/59479

Using Biosensor Devices and Ecological Momentary Assessment to Measure Emotion Regulation Processes: Pilot Observational Study With Dialectical Behavior Therapy

Shireen L Rizvi[1], PhD; Allison K Ruork[2], PhD; Qingqing Yin[3], MS; April Yeager[1], PsyM; Madison E Taylor[4], MA; Evan M Kleiman[3], PhD

1
2
3
4

Corresponding Author:
Shireen L Rizvi, PhD

Abstract

Background: Novel technologies, such as ecological momentary assessment (EMA) and wearable biosensor wristwatches, are increasingly being used to assess outcomes and mechanisms of change in psychological treatments. However, there is still a dearth of information on the feasibility and acceptability of these technologies and whether they can be reliably used to measure variables of interest.

Objective: Our objectives were to assess the feasibility and acceptability of incorporating these technologies into dialectical behavior therapy and conduct a pilot evaluation of whether these technologies can be used to assess emotion regulation processes and associated problems over the course of treatment.

Methods: A total of 20 adults with borderline personality disorder were enrolled in a 6-month course of dialectical behavior therapy. For 1 week out of every treatment month, participants were asked to complete EMA 6 times a day and to wear a biosensor watch. Each EMA assessment included measures of several negative affect and suicidal thinking, among other items. We used multilevel correlations to assess the contemporaneous association between electrodermal activity and 11 negative emotional states reported via EMA. A multilevel regression was conducted in which changes in composite ratings of suicidal thinking were regressed onto changes in negative affect.

Results: On average, participants completed 54.39% (SD 33.1%) of all EMA (range 4.7% - 92.4%). They also wore the device for an average of 9.52 (SD 6.47) hours per day and for 92.6% of all days. Importantly, no associations were found between emotional state and electrodermal activity, whether examining a composite of all high-arousal negative emotions or individual emotional states (within-person *r* ranged from −0.026 to −0.109). Smaller changes in negative affect composite scores were associated with greater suicidal thinking ratings at the subsequent timepoint, beyond the effect of suicidal thinking at the initial timepoint.

Conclusions: Results indicated moderate overall compliance with EMA and wearing the watch; however, there was no concurrence between EMA and wristwatch data on emotions. This pilot study raises questions about the reliability and validity of these technologies incorporated into treatment studies to evaluate emotion regulation mechanisms.

(JMIR Ment Health 2024;11:e60035) doi:10.2196/60035

KEYWORDS

wearable device; ecological momentary assessment; emotion regulation; psychotherapy mechanisms; dialectical behavior therapy; wearable; wristwatch; novel technology; psychological; treatment; pilot study; adult; personality disorder; mental health; mobile phone; EMA; observational study

Introduction

There is growing research interest in evaluating proposed mechanisms of change within evidence-based psychological interventions. This work is necessary to understand how and when treatments are effective and to improve and refine these interventions. With such research comes increased reliance on novel technologies to assess variables of interest. These technologies include ecological momentary assessment (EMA) and wearable biosensor wristwatches, which are increasingly used to assess outcomes and mechanisms of change in psychological interventions [1,2]. There is also interest in using these technologies to measure dynamic and mechanistic psychotherapy processes, which could replace assessment approaches that have relied on static and infrequent measures [3].

There are several reasons for the enthusiasm for EMA and wearable sensors in clinical research. The use of these more objective measures (in the case of wearable sensors) could reduce subjective bias in reporting. Using both EMA with multiple prompts per day and wearable sensors would lead to data being obtained in the individual's natural environment (as opposed to therapy sessions) which would aid in the understanding of contextual factors and symptom variability that can be assessed at the person level [1,4]. Finally, some research has indicated that EMA is more sensitive to change than self-report measures of depression and anxiety [5,6]. Together, these reasons make a strong rationale for incorporating these technologies into treatment research to assess treatment effects at a more granular level. However, despite these reasons for using technologies to address questions regarding mechanisms and processes of treatment, there has been limited attention paid to the overall feasibility and acceptability of using such methods. Such research is critically important because if studies find that compliance rates are low or the technologies are not otherwise feasible to incorporate into treatment research, the benefits of their use are limited.

Two recent studies with nonclinical samples have looked at adherence to intensive data collection methods (EMA, smartwatches, and chest patches) [7,8]. In a study of 45 healthy adults, King et al [8] found generally high compliance rates to EMA prompts (78.92%) and no indication that compliance dropped over time. However, the study lasted only 10 days so implications for longer-term studies are unclear. Ponnada et al [7] report on the preliminary results of a study with healthy young adults in which participants are asked to complete 4-day EMA bursts every 2 weeks. This study used "microinteraction EMA" which restricted the EMA items to simple questions answered directly on the smartwatch with a single tap. Results from a subsample of participants who had completed at least six months of the study (n=81) indicated a compliance rate of 67% to EMA prompts. This is a promising compliance rate for a long-term study; however, the use of healthy young adults and single-item questions may not translate to clinical samples in treatment mechanism studies.

In terms of research with clinical samples, very few studies have incorporated experience sampling methodology into treatment studies. In a study of 55 adults with depression, Eddington et al [9] used a program that automatically called participants 8 times a day for 1 week at baseline and 1 week posttreatment. During those calls, participants were asked 31 questions using prerecorded voice prompts that they answered by using their phone keypad. Among the treatment completers (n=29), compliance rates significantly dropped from 72.23% of call completion during the baseline week to 63.66% of calls during the posttreatment week. In a study of a 1-week internet-based intervention for individuals with social anxiety, Daniel et al [10] had participants complete 5 weeks of EMA (7 surveys a day). Similarly, they found EMA compliance to drop as the EMA period went on, suggesting that the participant burden was high and reduced the incentive to complete. Both of these studies have implications for incorporating technologies into longer-term treatment models.

Another important point for the incorporation of novel methodologies is that it is challenging for research on feasibility and acceptability to keep pace with technological advances. Often, by the time research is published, the technology that was used is outdated which also has implications for the generalizability of findings. For example, an early study [11] incorporating EMA into outpatient psychotherapy used a standalone iPod for participants to carry. EMA compliance rates were not provided, making it difficult to determine how acceptable this technology was for the participants. Furthermore, there remains a lack of consensus in the field regarding design standards and how to capture variables of interest while incorporating such technology into treatment [12]. For example, although interest in capturing emotional processes via EMA is high, there currently exists no standard for operationalizing emotion regulation using EMA [13]. In this study, we operationalized emotion regulation as the presence of high negative affect followed by a subsequent reduction.

In terms of wearable biosensor devices, there is a relatively large body of laboratory-based literature that has examined electrodermal activity (EDA; how conductive is electricity to one's skin, also called skin conductance level) as an index of physiological arousal that corresponds to the experience of high arousal emotions [14,15]. However, there has been far less exploration of the extent to which ambulatory EDA measures correspond to momentary reports of emotion. If EDA and momentary reports of negative emotion do not correspond, then there is likely limited use in using EDA to detect the changes in emotion associated with emotion regulation. The aims of this study were (1) to evaluate the feasibility and acceptability of incorporating EMA and wearable devices into standard dialectical behavior therapy (DBT) and (2) to conduct an exploratory analysis to evaluate whether these technologies can be used to assess emotion regulation processes and their relation to suicide thinking.

Methods

Ethical Considerations

This study was reviewed and approved by the Rutgers University institutional review board (Pro2019001864). All participants provided informed consent as part of the recruitment process and agreed for their deidentified data to be used for research purposes. Participants were compensated US $50 - 60 in gift cards for each assessment, excluding the baseline assessment for which there was no compensation.

Participants

Participants were 20 individuals with borderline personality disorder (BPD) who underwent 6 months of standard DBT in a university-based training clinic. The overall clinic procedures are registered (see ClinicalTrials.org, NCT03123198); however, this substudy of 20 sequentially admitted participants is not. Inclusion criteria required that study participants (1) be 18 years of age or older, (2) agree to take part in research assessments and video recording of therapy and assessments, (3) agree to pay sliding scale fee for treatment sessions, (4) maintain residence within commuting distance of the clinic (<45 minutes), (5) agree to discontinue other forms of therapy (excluding psychotropic medication management), (6) meet diagnostic criteria for BPD, and (7) have an iOS or Android smartphone compatible with our EMA software (MetricWire; ie, from the past 5 years). Potential participants were excluded if they (1) required mental health services not covered by DBT (eg, schizophrenia and life-threatening anorexia nervosa); (2) were non–English-speaking; (3) presented indication of intellectual disability; or (4) were unable to understand consent forms. All participants provided written informed consent for inclusion in the study.

Measures

Of relevance to this study are variables collected via EMA and wearable devices.

Ecological Momentary Assessment

At each EMA prompt, we assessed ratings of negative affect, suicide ideation, as well as variables unrelated to the study (eg, urges to engage in other harmful behaviors like drug use). Participants were asked to rate on a scale from 0 to 5 how much they felt each of the emotions of anxious, sad, agitated, irritated, shame, guilt, self-hate, angry, hopeless, lonely, and burdensome. We created an overall negative affect composite that combined these 11 variables for each prompt (within-person reliability α=.89 and between-person α=.95) and a high-arousal negative affect composite that combined agitated, angry, anxious, and irritated (within-person reliability α=.78 and between-person α=.95). Additionally, at each prompt, participants were asked to answer 2 questions on a 0 - 5 scale—“Right now, how strong is your desire to kill yourself” and “Right now, how strong is your desire to stay alive” (reverse coded). These 2 items were combined to create a suicide ideation (SI) composite score (within-person reliability α=.53 and between-person α=.67).

Wearable Device Assessment

The first 11 participants were provided the Empatica Embrace and the last 9 participants the Phillips Healthband. Both devices included an accelerometer that can be used to derive information on movement and sleep. The Embrace measured EDA at 4 Hz with electrodes facing the top of the first. The Healthband measured blood volume pressure using a photoplethysmograph at 32 Hz in order to derive heart rate (HR). Both EDA and HR can be used as markers of physiological arousal possibly associated with high arousal negative emotion (and the regulation of this emotion). However, raw HR data are not available for the Healthband, meaning that we could not use these data to explore physiological correlates of emotion regulation. Thus, we focus only on EDA from the Embrace for this study.

Procedure

Recruitment, Screening, and Assessment

Participants self-referred to the clinic and were briefly screened via telephone to determine initial eligibility. Interested and eligible clients were scheduled for an assessment to provide informed consent, confirm eligibility, and complete diagnostic interviews and self-report measures. These meetings were conducted by a graduate student or postdoctoral fellow in clinical psychology, under the supervision of the first author. Given the timing of the study within the COVID-19 pandemic, 85% (n=17) of the intake assessments were conducted via Health Insurance Portability and Accountability Act (HIPAA)–compliant Zoom (Zoom Video Communications). If the client was eligible for the study and interested in participation, they were oriented to the EMA and wearable device procedures. Participants then began DBT, completing standard assessments again at midtreatment, posttreatment, and 3-month follow-up.

EMA and Wearable Device

Participants were asked to complete EMA prompts and wear the wristwatch for 1 week of every treatment month, yielding up to 6 weeks of data for each participant. For EMA, participants were prompted 6 times per day, with the first and last prompts based on user-identified sleep and wake times and the remainder sent randomly within prespecified windows. A total of 5 of these surveys were shorter assessments of momentary effect and related factors. The final survey was a longer nightly survey, which contained the items in the random survey plus other items reflecting on the day. Participants were compensated US $0.25 for each of the 5 daily momentary surveys completed and US $0.50 for completing each nightly survey. Participants received an additional US $1.00 per day each for completing 4 or more surveys and wearing the physiological monitor for 6 or more consecutive hours each day. As a bonus, participants were compensated US $5.00 for each week they wore the biosensor for at least 5 days (for 6 consecutive hours each day).

DBT Treatment

Treatment providers were clinical psychology graduate students or postdoctoral fellows who were supervised by the first author and completed fundamental coursework in DBT. Clients completed 6 months of comprehensive DBT including weekly individual therapy, weekly group skills training, and intersession skills coaching per the treatment manuals [16,17]. Clients who missed 4 consecutive individual therapy appointments or group skills training sessions were considered treatment dropouts. Fees for services were assigned on a sliding scale determined by household income ranging from US $10 to US $100 per week. Due to the onset of the COVID-19 pandemic, the vast majority of the treatment was delivered via telehealth.

Data Analysis

To assess feasibility, descriptive statistics were used to summarize compliance rates with EMA and the wearable device over time. To determine whether these technologies can be used to assess emotion regulation processes, we conceptualized the process as what happens within the same day after a high negative affect instance. Thus, we examined all EMA data where we had 2 consecutive data points within 24 hours of each other and where the first data point (T1) was >0.5 SD above the participant mean on a negative affect composite variable. To explore how psychophysiological data corresponded to emotion, we focused on the EDA data that were collected with the

Embrace watch. To examine the correspondence between EDA and emotion, we first removed all the data that were likely recorded when the device was not being worn; specifically, we removed any data where the device detected temperature that was unlikely to be skin temperature (ie, <30 °C). We then preprocessed the EDA data using the *signal* R package (R Core Team) [18] by (1) upsampling it to 8 Hz and (2) applying a Butterworth filter to "flatten" the signal (ie, reduce possible noise). After preprocessing, we averaged all EDA data that occurred in the 60 seconds before each EMA prompt. This pre-EMA measure of EDA was used to examine correlations with each affected state, as well as an overall negative affect composite (all 11 emotions) and a high-arousal negative affect composite (agitated, angry, anxious, and irritated). We calculated in the *psych* R package [19] the average between-person correlations (ie, the average of each person's mean EDA and mean EMA) and within-person (ie, the average of each person's within-person correlation matrix across momentary observations) as our index of correspondence. Given the high number of correlations being examined here, we also provide a Holm-corrected *P* value that adjusts for potential type I error by using a "step-down" approach where the lowest *P* value is divided by the total number of analyses, the next lowest is divided by the number of analyses minus 1, and so on. Finally, we were interested in examining whether there were changes in suicidal ideation during instances of high negative affect and subsequent change. To evaluate these relationships, a multilevel regression was conducted using the *lme4* package in R [20], where changes in a composite rating of SI were regressed onto changes in negative affect composite score while controlling time in treatment.

Results

Participant Characteristics

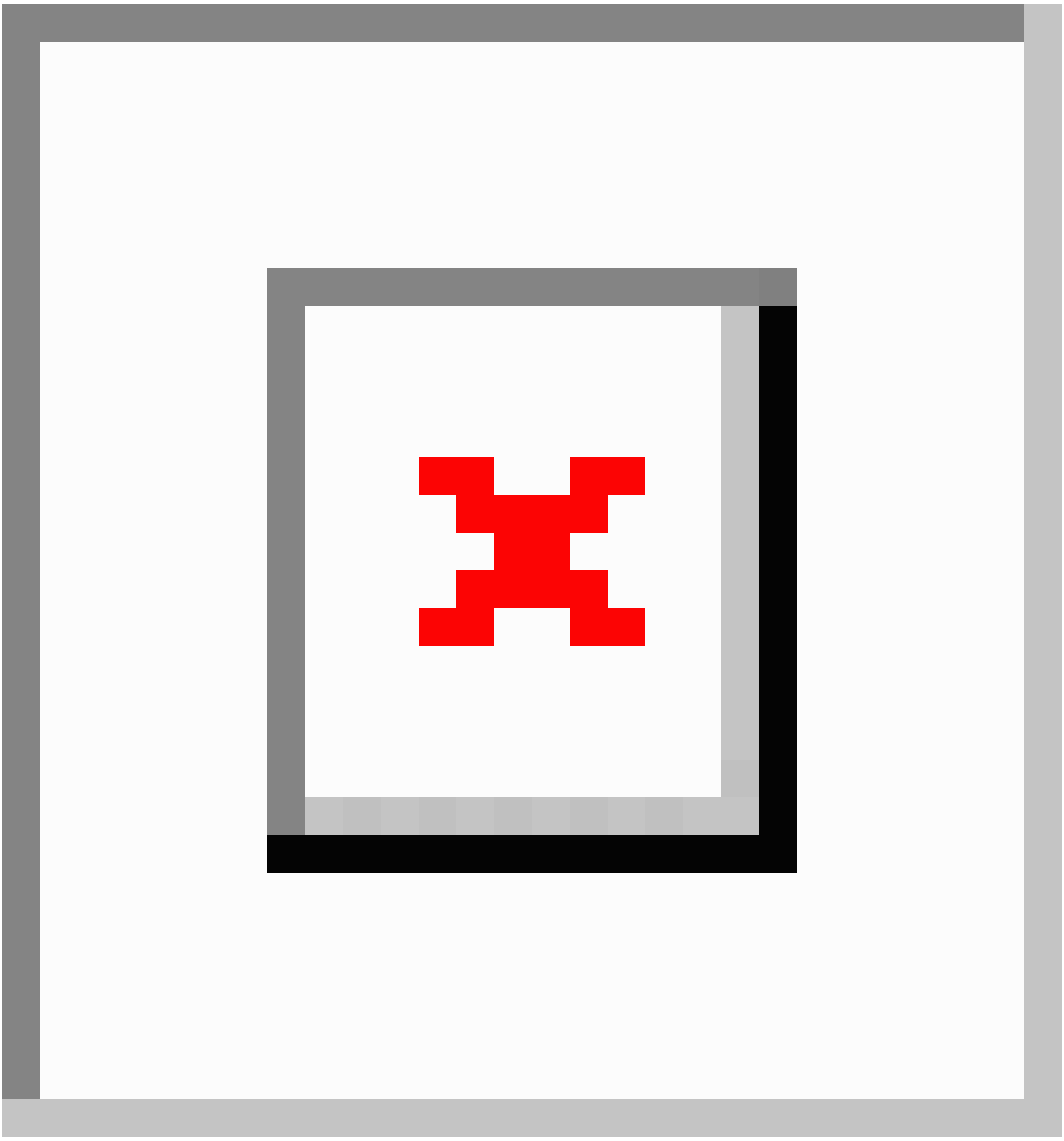
Participants' average age was 28.45 (SD 10.93; range 19-65) years. A total of 16 participants (80%) identified as female. Participants identified their race as Caucasian (n=16, 80%), Asian (n=3, 15%), or Black or African American (n=1, 5%), and 20% (n=4) identified as Hispanic. Most participants reported some college (n=9, 45%) or college (n=3, 15%) as their highest level of education, with others reporting completing at least some graduate school (n=5, 25%) and high school (n=3, 15%). Most participants (n=17, 85%) reported being prescribed psychotropic medications. On average, participants reported 2.85 (SD 1.81) current and 3.55 (SD 2.35) lifetime comorbid psychiatric disorders as measured by the Structured Clinical Interview for DSM-5 Disorders (SCID-5) [21]. In terms of treatment compliance, participants attended an average of 24.55 (SD 5.38; range 8 - 36) individual therapy sessions and 20.85 (SD 4.49; range 9 - 25) group skills training sessions. One (5%) participant dropped out after 2 months of treatment.

Compliance With Technologies

On average, participants completed 54.39% (SD 33.1%) of all EMA surveys. Compliance ranged from 4.7% to 92.4%. Participants responded to at least 1 survey on 76.33% of days, and 2 or more on 58.3% of days. EMA compliance decreased over the course of the study; average compliance was >50% for the first 4 data collection weeks and then dropped below 50% in the fifth week of data collection (see Figure 1).

Participants wore 1 of the 2 biosensors—Empatica Embrace and Philips Healthband. Participants (n=11) wore the Embrace on average for 9.52 hours per day during the weeks they were asked to wear the device (SD 6.47 hours) and wore the device for at least some amount of time on 92.6% of all days. Participants (n=9) wore the Healthband for 10.58 hours per day during the weeks they were asked to wear the device (SD 10.51 hours) and wore the device for at least some of the time on 74.6% of days.

Figure 1. EMA compliance changes over the course of study involvement. EMA: ecological momentary assessment.



Assessing Emotion Regulation Processes Via EMA

For these analyses, an opportunity for emotion regulation processes to occur was operationalized to be (1) a higher rating (>0.5 SD above the participant mean) on an overall negative affect composite variable that (2) had a subsequent data point within 24 hours (*k*=230; n=12). A total of 8 participants were not included in the analyses due to a lack of observations that fit these parameters. There was a high degree of variability in the negative affect composite over the course of the study at both the within- and between-person level (intraclass correlation=0.51). In addition, there was a decrease in negative affect (ie, change >0.5 SD) from T1 to T2 in 55.35% (*k*=128) of these emotion regulation opportunities, an increase in 35.36% (*k*=82) of the opportunities, and a change within ±0.5 SD in 8.29% (*k*=20) of the opportunities. Table 1 shows the results of the correlations between EMA and EDA. As can be seen from the table, none of the correlations were significant after the Holm correction. Table 2 shows the relationship between emotion regulation processes and suicide ideation. Smaller changes in negative affect composite scores were associated with greater SI ratings at time 2, beyond the effect of SI ratings at time 1.

Table . Correlations between EMA[a] and EDA[b] data.

Variable	Within person			Between person		
	r	*P* value	Holm *p*	*r*	*P* value	Holm *p*
Composites						
Negative affect	−0.090	.05	0.514	−0.324	.36	1.00
High arousal negative affect[c]	−0.082	.07	0.668	−0.284	.43	1.00
Individual states						
Agitated	−0.057	.22	0.876	−0.406	.25	1.00
Angry	−0.074	.11	0.760	−0.435	.21	1.00
Anxious	−0.058	.21	1.00	−0.114	.75	0.753
Burden	−0.044	.34	1.00	−0.296	.41	1.00
Guilt	−0.026	.57	1.00	−0.459	.18	1.00
Hopeless	−0.019	.68	0.683	−0.353	.32	1.00
Lonely	−0.075	.10	0.817	−0.115	.75	1.00
Sad	−0.097	.03	0.379	−0.347	.33	1.00
Self-hate	−0.109	.02	0.207	−0.261	.47	1.00
Shame	−0.060	.19	1.00	−0.344	.33	1.00

[a]EMA: ecological momentary assessment.

[b]EDA: electrodermal activity.

[c]High arousal composite includes agitated, angry, anxious, and irritated.

Table . Results of regression analysis.

Predictors	Dependent variable: suicidal thinking at T2	
	β (95% CI)	*P* value
Intercept	0.599 (0.302 to 0.896)	<.001
Suicidal thinking at T1	0.636 (0.572 to 0.701)	<.001
Change in negative affect from T1 to T2	−0.097 (−0.107 to 0.086)	<.001
Random effects		
σ^2	1.21	—[a]
$\tau_{00\ \mathrm{ID}}$	0.18	—
Marginal R^2/conditional R^2	0.568/0.625	—

[a]Not applicable.

Discussion

The main findings of this study were that compliance with technologies was moderate and dropped over the course of the study and that no associations were found between data obtained from EMA and data obtained via the wearable device. To our knowledge, this pilot study was the first to incorporate both biosensor wearable devices and EMA to measure a key target mechanism and emotion regulation in real time during a psychological intervention (ie, DBT). Although a small sample, there were a number of important findings and their implications lead to concern for future research in this domain.

In terms of feasibility and acceptability, we found a moderate rate of EMA compliance (54%) with at least 1 survey completed on 76.33% of days. This compliance rate is lower than rates found in other EMA studies conducted [22]. This may be related to the long duration of the study, especially since compliance dropped over the course of 6 months. To effectively study emotion regulation in real time and as a function of treatment, it is preferred to have several completed EMA prompts within the same day and for the completed EMA prompts to be relatively stable over time. Our finding that 2 or more prompts were completed on just 58.3% of days indicates that future studies will need to place greater emphasis on increasing the number of surveys completed per day (to examine T1-T2 changes) within clinical samples. Indeed, 8 of the 20 (40%) participants had to be excluded from EMA analyses of emotion regulation processes because they did not provide data that met our measurement parameters. Similarly, compliance with the wearable biosensor device was generally high but waned over time. Our finding that compliance with both technologies waned over the course of the study renders observing treatment effects more difficult. Determining how to keep compliance up over time is likely crucial for evaluating treatment changes.

Our findings that EMA and EDA data were not correlated are noteworthy. Because EMA requires the participant to report their own experiences, it is often considered more "subjective" while psychophysiological data are often seen as more "objective." However, it is important to recognize that the field does not have a ground truth (gold standard) that states which form of measurement is "accurate." This lack of consensus coupled with our results makes it difficult to determine which technology needs to improve to increase validity. It is also possible that this lack of correspondence provides counterevidence for the classical view of emotions which suggests emotions have natural and physical essences that may be better captured using perceiver-independent tools (eg, autonomic nervous system activation). Rather, emotions could be multidimensional and different assessment approaches offer unique information that does not necessarily correlate [23].

There are some limitations to the study. First, the sample had only 20 participants and data collection occurred primarily during the pandemic. It is unclear how representative these individuals are of clients in DBT generally. Second, we focused on a few variables of interest in our EMA surveys. Other variables may have proven more reliable or consistent with EDA. The lack of consensus on how to define emotion regulation via EMA leaves researchers to determine appropriate variables themselves. Third, our sample included only individuals diagnosed with BPD. Although this is the target sample for DBT, this population may not represent other therapy populations and studies with different samples may yield different compliance rates. Fourth, because raw HR data were not available, we were unable to include it in these more granular analyses, this is unfortunate given the evidence linking BPD and emotional ability with HR variability [24]. While not without problems (eg, poor data quality for diverse skin) [25], future studies would benefit from including HR.

Although new technologies are often quickly embraced and used in psychological research, our study suggests that researchers should be cautious about using these technologies to measure emotion regulation processes in real time. It is likely that solutions to the problem require effort in engineering (eg, making the devices easier to use) and psychosocial (eg, designing protocols to maximize wear time) domains, as well as advances in broader emotion research. Until these solutions are identified and implemented, continuing to use these technologies, such as EDA, in psychotherapy studies may prove premature and unlikely to yield accurate pictures of treatment mechanisms and processes.

Acknowledgments

This study was funded by the Rutgers University Busch Biomedical Grant Program. SLR is a consultant and trainer for Behavioral Tech, LLC, and receives royalties from Guilford Press. EMK receives compensation from Wiley Publishers. The authors thank all the participants in the study, as well as the therapists and assessors at the Dialectical Behavior Therapy Clinic at Rutgers University (DBT-RU).

Conflicts of Interest

None declared.

References

1. Ebner-Priemer UW, Trull TJ. Ecological momentary assessment of mood disorders and mood dysregulation. Psychol Assess 2009 Dec;21(4):463-475. [doi: 10.1037/a0017075] [Medline: 19947781]
2. Rosenthal MZ, Fang CM, Chapman AL. Ambulatory measurement of emotional dysfunction in borderline personality disorder. Curr Opin Psychol 2015 Jun;3:75-79. [doi: 10.1016/j.copsyc.2015.02.008]
3. Hofmann SG, Curtiss JE, Hayes SC. Beyond linear mediation: toward a dynamic network approach to study treatment processes. Clin Psychol Rev 2020 Mar;76:101824. [doi: 10.1016/j.cpr.2020.101824] [Medline: 32035297]
4. Myin-Germeys I, Kasanova Z, Vaessen T, et al. Experience sampling methodology in mental health research: new insights and technical developments. World Psychiatry 2018 Jun;17(2):123-132. [doi: 10.1002/wps.20513] [Medline: 29856567]
5. Forbes EE, Stepp SD, Dahl RE, et al. Real-world affect and social context as predictors of treatment response in child and adolescent depression and anxiety: an ecological momentary assessment study. J Child Adolesc Psychopharmacol 2012 Feb;22(1):37-47. [doi: 10.1089/cap.2011.0085] [Medline: 22339611]
6. Moore RC, Depp CA, Wetherell JL, Lenze EJ. Ecological momentary assessment versus standard assessment instruments for measuring mindfulness, depressed mood, and anxiety among older adults. J Psychiatr Res 2016 Apr;75:116-123. [doi: 10.1016/j.jpsychires.2016.01.011] [Medline: 26851494]
7. Ponnada A, Wang S, Chu D, Do B, Dunton G, Intille S. Intensive longitudinal data collection using microinteraction ecological momentary assessment: pilot and preliminary results. JMIR Form Res 2022 Feb 9;6(2):e32772. [doi: 10.2196/32772] [Medline: 35138253]

8. King ZD, Yu H, Vaessen T, Myin-Germeys I, Sano A. Investigating receptivity and affect using machine learning: ecological momentary assessment and wearable sensing study. JMIR Mhealth Uhealth 2024 Feb 7;12:e46347. [doi: 10.2196/46347] [Medline: 38324358]
9. Eddington KM, Burgin CJ, Silvia PJ, Fallah N, Majestic C, Kwapil TR. The effects of psychotherapy for major depressive disorder on daily mood and functioning: a longitudinal experience sampling study. Cognit Ther Res 2017 Apr;41(2):266-277. [doi: 10.1007/s10608-016-9816-7] [Medline: 28579660]
10. Daniel KE, Daros AR, Beltzer ML, Boukhechba M, Barnes LE, Teachman BA. How anxious are you right now? Using ecological momentary assessment to evaluate the effects of cognitive bias modification for social threat interpretations. Cogn Ther Res 2020 Jun;44(3):538-556. [doi: 10.1007/s10608-020-10088-2]
11. Husen K, Rafaeli E, Rubel JA, Bar-Kalifa E, Lutz W. Daily affect dynamics predict early response in CBT: feasibility and predictive validity of EMA for outpatient psychotherapy. J Affect Disord 2016 Dec;206:305-314. [doi: 10.1016/j.jad.2016.08.025]
12. Lutz W, Schwartz B, Delgadillo J. Measurement-based and data-informed psychological therapy. Annu Rev Clin Psychol 2022 May 9;18:71-98. [doi: 10.1146/annurev-clinpsy-071720-014821] [Medline: 34910567]
13. Bettis AH, Burke TA, Nesi J, Liu RT. Digital technologies for emotion-regulation assessment and intervention: a conceptual review. Clin Psychol Sci 2022 Jan;10(1):3-26. [doi: 10.1177/21677026211011982] [Medline: 35174006]
14. Caruelle D, Gustafsson A, Shams P, Lervik-Olsen L. The use of electrodermal activity (EDA) measurement to understand consumer emotions – a literature review and a call for action. J Bus Res 2019 Nov;104:146-160. [doi: 10.1016/j.jbusres.2019.06.041]
15. Horvers A, Tombeng N, Bosse T, Lazonder AW, Molenaar I. Detecting emotions through electrodermal activity in learning contexts: a systematic review. Sensors (Basel) 2021 Nov 26;21(23):7869. [doi: 10.3390/s21237869] [Medline: 34883870]
16. Linehan M. Skills Training Manual for Treating Borderline Personality Disorder: The Guilford Press; 1993.
17. Linehan M. DBT Skills Training Manual, 2nd edition: The Guilford Press; 2014.
18. Mersmann O, Ligges U, Krey S. Signal: signal processing. R-Forge. 2013. URL: http://r-forge.r-project.org/projects/signal [accessed 2023-05-28]
19. Revelle W. Psych: procedures for personality and psychological research (2.2.5.). The Comprehensive R Archive Network. 2022. URL: https://CRAN.R-project.org/package=psych [accessed 2023-05-28]
20. Bates D, Maechler M, Bolker B. lme4. R-Forge. 2004. URL: http://lme4.r-forge.r-project.org [accessed 2023-05-28]
21. First MB, Williams JBW, Karg RS, Spitzer RL. SCID-5-CV: Structured Clinical Interview for DSM-5 Disorders: Clinician Version: American Psychiatric Association Publishing; 2016.
22. Wrzus C, Neubauer AB. Ecological momentary assessment: a meta-analysis on designs, samples, and compliance across research fields. Assessment 2023 Apr;30(3):825-846. [doi: 10.1177/10731911211067538] [Medline: 35016567]
23. Barrett LF. Navigating the science of emotion. In: Meeiselman HL, editor. Emotion Measurement: Woodhead Publishing; 2016:31-63 URL: https://www.affective-science.org/pubs/2016/barrett-navigating-2016.pdf [accessed 2023-05-28] [doi: 10.1016/B978-0-08-100508-8.00002-3]
24. Koenig J, Kemp AH, Feeling NR, Thayer JF, Kaess M. Resting state vagal tone in borderline personality disorder: a meta-analysis. Prog Neuropsychopharmacol Biol Psychiatry 2016 Jan 4;64:18-26. [doi: 10.1016/j.pnpbp.2015.07.002] [Medline: 26169575]
25. Bent B, Goldstein BA, Kibbe WA, Dunn JP. Investigating sources of inaccuracy in wearable optical heart rate sensors. NPJ Digit Med 2020;3:18. [doi: 10.1038/s41746-020-0226-6] [Medline: 32047863]

Abbreviations

BPD: borderline personality disorder
DBT: dialectical behavior therapy
EDA: electrodermal activity
EMA: ecological momentary assessment
HIPAA: Health Insurance Portability and Accountability Act
HR: heart rate
SCID-5: Structured Clinical Interview for DSM-5 Disorders
SI: suicide ideation

Edited by J Torous; submitted 17.05.24; peer-reviewed by A Daros, P Washington; revised version received 12.07.24; accepted 20.07.24; published 09.10.24.

Please cite as:
Rizvi SL, Ruork AK, Yin Q, Yeager A, Taylor ME, Kleiman EM
Using Biosensor Devices and Ecological Momentary Assessment to Measure Emotion Regulation Processes: Pilot Observational Study With Dialectical Behavior Therapy
JMIR Ment Health 2024;11:e60035
URL: https://mental.jmir.org/2024/1/e60035
doi:10.2196/60035

Original Paper

A Comparison of ChatGPT and Fine-Tuned Open Pre-Trained Transformers (OPT) Against Widely Used Sentiment Analysis Tools: Sentiment Analysis of COVID-19 Survey Data

Juan Antonio Lossio-Ventura[1], PhD; Rachel Weger[2], BA; Angela Y Lee[3], MA; Emily P Guinee[1], BA; Joyce Chung[1], MD; Lauren Atlas[4], PhD; Eleni Linos[5], MD; Francisco Pereira[1], PhD

[1]National Institute of Mental Health, National Institutes of Health, Bethesda, MD, United States

[2]School of Medicine, University of Pittsburgh, Pittsburgh, PA, United States

[3]Department of Communication, Stanford University, Stanford, CA, United States

[4]National Center For Complementary and Alternative Medicine, National Institutes of Health, Bethesda, MD, United States

[5]School of Medicine, Stanford University, Stanford, CA, United States

Corresponding Author:
Juan Antonio Lossio-Ventura, PhD
National Institute of Mental Health
National Institutes of Health
3D41
10 Center Dr
Bethesda, MD, 20814
United States
Phone: 1 3018272632
Email: juan.lossio@nih.gov

Abstract

Background: Health care providers and health-related researchers face significant challenges when applying sentiment analysis tools to health-related free-text survey data. Most state-of-the-art applications were developed in domains such as social media, and their performance in the health care context remains relatively unknown. Moreover, existing studies indicate that these tools often lack accuracy and produce inconsistent results.

Objective: This study aims to address the lack of comparative analysis on sentiment analysis tools applied to health-related free-text survey data in the context of COVID-19. The objective was to automatically predict sentence sentiment for 2 independent COVID-19 survey data sets from the National Institutes of Health and Stanford University.

Methods: Gold standard labels were created for a subset of each data set using a panel of human raters. We compared 8 state-of-the-art sentiment analysis tools on both data sets to evaluate variability and disagreement across tools. In addition, few-shot learning was explored by fine-tuning Open Pre-Trained Transformers (OPT; a large language model [LLM] with publicly available weights) using a small annotated subset and zero-shot learning using ChatGPT (an LLM without available weights).

Results: The comparison of sentiment analysis tools revealed high variability and disagreement across the evaluated tools when applied to health-related survey data. OPT and ChatGPT demonstrated superior performance, outperforming all other sentiment analysis tools. Moreover, ChatGPT outperformed OPT, exhibited higher accuracy by 6% and higher *F*-measure by 4% to 7%.

Conclusions: This study demonstrates the effectiveness of LLMs, particularly the few-shot learning and zero-shot learning approaches, in the sentiment analysis of health-related survey data. These results have implications for saving human labor and improving efficiency in sentiment analysis tasks, contributing to advancements in the field of automated sentiment analysis.

(JMIR Ment Health 2024;11:e50150) doi:10.2196/50150

KEYWORDS

sentiment analysis; COVID-19 survey; large language model; few-shot learning; zero-shot learning; ChatGPT; COVID-19

Introduction

Background

Sentiment analysis is a field within natural language processing (NLP) that aims to extract sentiments and opinions from text related to specific entities and topics [1], such as people, organizations, events, and places [2]. Specifically, we consider the task of classifying texts as positive, neutral, or negative. Research in this area can occur at different levels of granularity, ranging from a single sentiment for an entire document or for each sentence within it to exploring various aspects associated with each entity, which can be associated with different sentiments [1,3].

Recently, we have witnessed an increase in the use of sentiment analysis to computationally evaluate the attitudes, perceptions, and emotions of social media users regarding the COVID-19 pandemic [4,5]. Most of these works study content from social media platforms such as Twitter, Reddit, and Facebook [6], as social media has been a main platform to express opinions related to COVID-19 in a public manner. Simultaneously, surveys, which refer to data collected from a group of people regarding their opinions, behavior, or knowledge through specifically designed questions, have also been used to investigate the impact of the COVID-19 pandemic. In particular, surveys conducted during the lockdown period in 2020 examined the effects on people's lives, behaviors, and mental health, among other topics [7-9]. Web-based surveys are often semistructured, that is, composed of closed-answer components (eg, different clinical questionnaires) and open-ended questions that allow a free-text answer. Sentiment analysis tools have been applied to the latter to help monitor the attitudes, sentiments, and perceptions of the participants during the pandemic to assist health decision-making [10].

The application of sentiment analysis tools on free-text data obtained from surveys poses challenges for health care providers and researchers in the health domain. This is partly attributed to the fact that most state-of-the-art applications are designed for different domains, such as social media, and there is limited knowledge regarding their performance in survey data. In addition, recent studies have applied the most well-known sentiment analysis tools, including TextBlob [11], VADER (Valence Aware Dictionary and Sentiment Reasoner) [12], and Stanza [13], to analyze health-related content on social media platforms [14-16] and, more recently, in the context of COVID-19 [6,17]. These studies highlighted the need for a more comprehensive evaluation of sentiment analysis tools, as the initial results exhibited a lack of accuracy and yielded inconsistent outcomes [15,16]. The main reason for this discrepancy was the disparity in data sets and the potential sensitivity of the tools to the composition of the data set [16]. Consequently, researchers trained new algorithms tailored to their specific data set.

Two COVID-19 survey data sets were used in this study, both collected by teams from the National Institutes of Health (NIH) and Stanford University. The collected data were used to assess the general topics experienced by the participants during the pandemic lockdown. Researchers from both institutions aimed to comprehend the general sentiment patterns over time and identify an overall sentiment for events during that period, such as vaccines and the 2020 presidential elections. In both data sets, it was often the case that a complete response contained multiple topics, with many sentences referring to distinct subjects. Thus, this study is focused on the analysis of sentiment at the sentence level. By assessing each sentence independently, subtle shifts in sentiment could be captured, which could potentially be neglected at the document level. Moreover, we thought that an analysis based on sentence level, rather than aspect-based level, was more appropriate, given that our focus was not on the granularity of the various aspects of an entity. For instance, when evaluating different features of an intensive care unit, aspects might encompass ventilators, rooms, staff, nurses, and others. Therefore, the decision to focus on sentence-level sentiment analysis is influenced by practical considerations, our research objectives, and the nature of the survey responses.

In this study, as the first contribution, we analyzed 2 independent survey data sets containing free-text data collected during the lockdown period of the COVID-19 pandemic, with accompanying ground-truth sentiment labels generated by human raters for hundreds of responses. The second contribution involves a comparison of 8 widely used state-of-the-art sentiment analysis tools, which have been frequently and recently used in the health domain [16], on COVID-19 surveys at the sentence level. We demonstrate that performance across tools varies and that there is a complex correlation structure between their predicted polarity scores. The third contribution of this paper is to investigate whether the polarity prediction performance can be improved through few-shot learning on a small labeled data set or zero-shot learning with ChatGPT [18].

Related Work

There are 2 main approaches to performing sentiment analysis: lexicon based and machine learning based. Initial lexicon methods are the simplest rule-based methods and seek to classify the sentiment of a sentence as a score function of the word polarities existing in a dictionary [19-23]. Lexicon-based techniques use mostly adjectives and adverbs to compute the overall sentiment score of a text, for instance, Linguistic Inquiry and Word Count (LIWC) [24], Affective Norms for English Words [25], and SentiWordNet [26]. Dictionaries of lexicons are created either manually or automatically [27,28]. First, a list is generated from a specific domain. Then synonyms and antonyms are added from other existing dictionaries such as WordNet [29]. More sophisticated lexicon-based methods focus on complex rules, such as regular expressions [30,31], instead of simply computing a sentiment score based on word polarities.

Machine learning–based techniques use statistical methods to compute sentiment polarity. The process involves training a classifier on a labeled data set, such as movie reviews or social media posts, and then using the model to predict the sentiment of new, unlabeled data. Obtaining labeled data to train the classifiers is a time-consuming task. Machine learning–based methods often face challenges when processing negative and intensifying statements and can have low performance when applied to different domains, as they rely mainly on the data set size. The rules proposed in the lexicon-based approaches have also been used to extract relevant features and used as input to machine learning algorithms (eg, naive Bayes, *k*-nearest neighbors, decision tree, and logistic regression) to predict the sentiment [32-37]. Other machine learning methods are based on deep neural networks (DNNs). DNNs have been successfully used for sentiment analysis, as described in detail by Birjali et al [3], Zhang et al [38], and Yadav and Vishwakarma [39], having achieved state-of-the-art performance on several benchmarks. DNN architectures used include recurrent neural networks [40,41], long short-term memory networks [42,43], and convolutional neural networks [44-46].

More recently, transformers [47] (deep learning architectures) and large language models (LLMs) have gained popularity due to their ability to perform NLP tasks, including sentiment analysis, with remarkable performance. These LLMs have been pretrained on large text corpora using transformers, such as Bidirectional Encoder Representations from Transformers (BERT) [48], Robustly optimized BERT approach (RoBERTa) [49], Embeddings from Language Models (ELMo) [50], Generative Pre-trained Transformers (GPT) [51], and Pathways Language Model (PaLM) [52]. LLMs in sentiment analysis can handle several data types and domains as well as identify patterns and relationships between the semantics of words and phrases that are indicative of sentiment. LLMs for sentiment analysis can also be fine-tuned to specific domains and applications, which usually lead to better results, as shown in previous studies [53-59]. Finally, ChatGPT (OpenAI) [18] has suddenly emerged to produce human-like responses to user inputs. The notable performance of LLMs has led to increased interest in few-shot and zero-shot learning methods using them. Few-shot learning algorithms enable a model to learn from only a few examples, whereas zero-shot learning algorithms can transfer knowledge from one task to another without additional labeled training examples. These approaches have demonstrated comparable or superior performance to prior state-of-the-art fine-tuning methods on various NLP tasks [60-62].

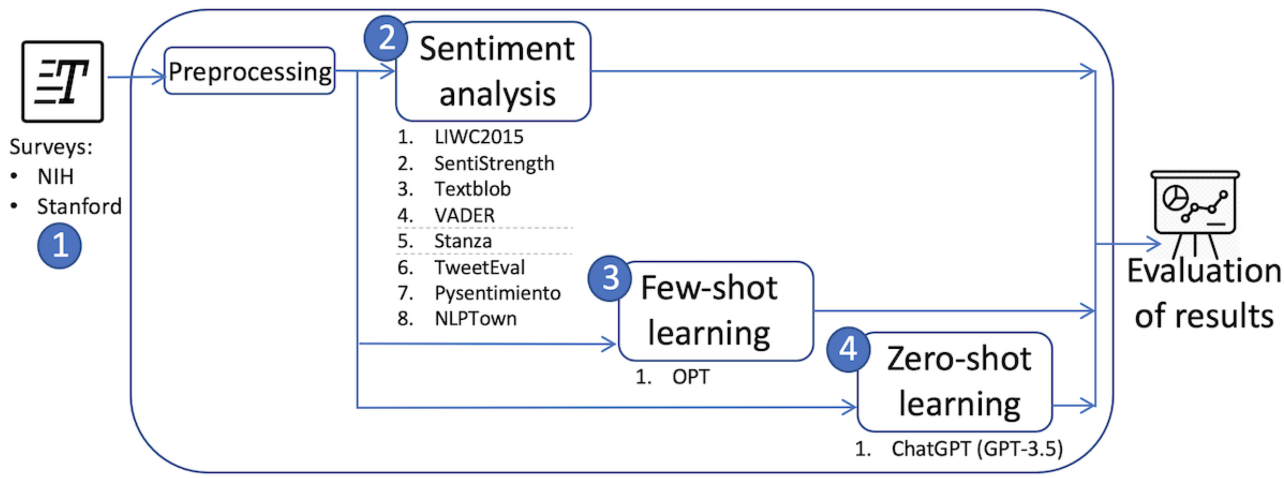
Sentiment analysis has become an increasingly popular technique in the health domain, as noted in the study by Rodríguez-Ibánez et al [63]. A recent study [64] also found that the main data source for studies on health is social media, such as Twitter and Facebook. This is attributed to advancements in mobile technology and their use as a source in health-related topics, such as finding treatments, sharing experiences and opinions, and addressing public health surveillance issues [65-67]. During the pandemic, we witnessed social media becoming the main forum to express opinions related to COVID-19, which helped authorities to understand and monitor sentiments toward topics related to the pandemic [68-73].

Various studies have proposed new sentiment analysis methods and compared existing tools (eg, TextBlob [74], VADER [12], and Stanza [13]) on topics related to COVID-19, mainly extracted from social media [6,16,17,75-78]. However, to the best of our knowledge, there are no studies that have compared several sentiment analysis tools on health-related surveys—a more structured type of text data than social media posts—that collected knowledge, beliefs, and habits during the COVID-19 pandemic [79-84]. The only study we are aware of that evaluates ChatGPT on various sentiment analysis tasks, comparing it with fine-tuned BERT, is the study by Wang at al [85]. The results demonstrated that ChatGPT exhibited promising zero-shot sentiment analysis ability, achieving performance on par with fine-tuned BERT and state-of-the-art models. However, it fell slightly behind domain-specific fully supervised state-of-the-art models.

Methods

This section presents the data sets used in this study along with our evaluation of sentence sentiment analysis methods, as illustrated in Figure 1. Specifically, we describe the (1) survey data sets, (2) state-of-the-art sentiment analysis tools, (3) few-shot learning with an LLM, and (4) zero-shot learning with ChatGPT.

Figure 1. Workflow of our study for evaluating sentence sentiment analysis using state-of-the-art sentiment analysis tools, few-shot learning with a large language model, and zero-shot learning with ChatGPT over health-related surveys. GPT: Generative Pre-trained Transformers; LIWC2015: Linguistic Inquiry and Word Count 2015; NIH: National Institutes of Health; OPT: Open Pre-Trained Transformers; VADER: Valence Aware Dictionary and Sentiment Reasoner.



Data

NIH Data Set

This data set was collected as part of a web-based survey assessing mental health during the pandemic, which started from April 2020 to May 2021. This was a sample of convenience, as participants were recruited from a pool of previous participants in the National Institute of Mental Health and National Center for Complementary and Alternative Medicine studies by advertising on social media and by flyers within the Washington metropolitan area. Participants who signed up completed various questionnaires at baseline, assessing demographics, clinical history, and psychological state [86]. The participants were then sent emails every 2 weeks for 6 months, inviting them to complete 3 of those questionnaires at that time. This latter survey consisted of 45 questions assessing various attitudes, behaviors, and impacts surrounding the pandemic and a single free-response question ("Is there anything else you would like to tell us that might be important that we did not ask about?"). There was a maximum of 13 potential survey (and free) responses per participant. Of the 3655 participants who enrolled in the study, 2497 (68.31%) responded at least once to the free-response item, yielding a total of 9738 item responses. These were composed of 26,411 sentences, which were the data used in this study. The semantic content of these responses (eg, main topics of concern over time) is available in the study by Weger et al [87].

Stanford Data Set

This data set was collected as part of a web-based survey conducted from March to September 2020 by a Stanford University team. The survey was conducted using a sample of convenience recruited through 3 social media platforms: Twitter, Facebook, and Nextdoor. They could participate by clicking on a survey link in the social media post upon seeing the recruitment materials. The survey comprised 21 questions including demographics and the impact of COVID-19 on individuals' lives [88]. In this study, we focus on the evaluation of 3 free-text responses to the following questions: (1) "Although this is a challenging time, can you tell us about any positive effects or 'silver linings' you have experienced during this crisis?" (2) "What are the reasons you are not self-isolating more?" and (3) "Have you experienced any difficulties due to the coronavirus crisis?." Of the 4582 participants recruited, 3349 (73.09%) responded to at least 1 of the 3 free-text questions, resulting in a total of 7182 item responses. These were composed of approximately 21,266 sentences, which were the data used in this study. The topics and sentiments in these responses are reported in the study by Lossio-Ventura et al [10]. Table 1 presents additional details regarding the NIH and Stanford data sets.

Table 1. Details of the National Institutes of Health (NIH) and Stanford data sets.

	NIH	Stanford
Start of the collection period	April 2020	March 2020
End of the collection period	May 2021	September 2020
Responders, n/N (%)	2497/3655 (68.31)	3349/4582 (73.09)
Response items, n	9738	7182
Sentences before processing, n	26,411	21,266
Sentences after processing, n/N (%)	26,188/26,411 (99.16)	21,035/21,266 (98.91)
Tokens after processing, n	462,518	299,735
Tokens per sentence, mean (SD)	17.66 (11.11)	14.25 (9.74)

Annotation

We created training and test sets for both the NIH and Stanford data sets. These sets were derived from the surveys after completing the preprocessing steps and were used for training, tuning, and the official evaluation.

Training Data Set

We randomly selected 260 sentences, with 130 sentences from each data set. Each subset of 130 sentences was annotated by a different annotator. The annotators were instructed to assign a polarity value of −1 (negative), 0 (neutral), or 1 (positive) to each sentence.

Test Data Set

A total of 1000 sentences were randomly chosen, with 500 sentences selected from each data set [89]. Each set was annotated by 3 separate and independent annotators: A.1, A.2, and A.3 for NIH and A.4, A.5, and A.6 for Stanford. The annotators were instructed to assess the polarity of each sentence on a scale of −1 (negative), 0 (neutral), or 1 (positive).

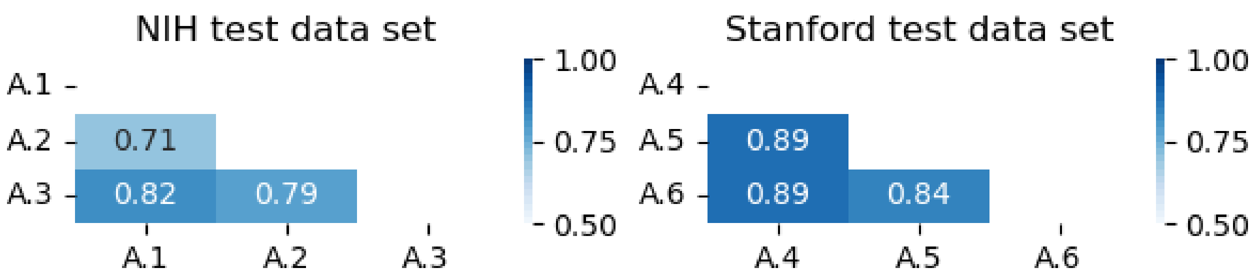
We used a 3-point scale to annotate the data. We then followed a 3-step procedure to determine the final labels, similar to that described in the studies by Nakov et al [90] and Rosenthal et al [91]. First, if all 3 annotators agreed on a label (full agreement), that label was accepted. Second, if 2 of the 3 agreed on a label (partial agreement), that label was also accepted. Third, if there was no agreement, the label was set as neutral (no agreement). Fleiss κ measure was calculated to assess the agreement between the 3 annotators of each test data set. The associated *P* values were computed to test if the agreement between annotators was substantially better than what would be expected by chance. Further details of the training and test data sets are provided in Table 2. Pearson correlation coefficients were also calculated to evaluate the degree of agreement between each pair of annotators, as shown in Figure 2.

Table 2. Details of the National Institutes of Health (NIH) and Stanford data sets.

	Training (n=130)		Test (n=500)	
	NIH	Stanford	NIH	Stanford
Sentences, n (%)	130 (100)	130 (100)	500 (100)	500 (100)
Negative sentences, n (%)	71 (54.6)	45 (34.6)	223 (44.6)	234 (46.8)
Neutral sentences, n (%)	51 (39.2)	41 (31.6)	232 (46.4)	117 (23.4)
Positive sentences, n (%)	8 (6.2)	44 (33.8)	45 (9)	149 (29.8)
Full agreement, n (%)	N/A[a]	N/A	340 (68)	385 (77)
Partial agreement, n (%)	N/A	N/A	159 (31.8)	112 (22.4)
No agreement, n (%)	N/A	N/A	1 (0.2)	3 (0.6)
Fleiss κ	N/A	N/A	0.6311	0.7572
P value	N/A	N/A	<.001	<.001

[a]N/A: not applicable.

Figure 2. Correlation of annotators on the National Institutes of Health (NIH) and Stanford test data sets. A.1, A.2, A.3 represent the 3 independent NIH annotators, while A.4, A.5, A.6 represent the Stanford annotators.



Preprocessing

The survey responses contained personal identifiable information and multiple sentences covering different themes, for example, 2020 presidential elections and COVID-19 vaccines. Therefore, preprocessing steps included splitting responses into sentences, replacing people's names, suppressing email addresses, and lemmatizing and converting text to lower case.

Sentiment Analysis Applications

We considered popular sentiment analysis applications available on the internet that use rules, machine learning, and fine-tuned LLMs.

Linguistic Inquiry and Word Count 2015

Linguistic Inquiry and Word Count 2015 (LIWC2015) [24,92,93] is a text analysis software that identifies and calculates the frequency of different categories of words in texts, such as pronouns, emotional words, cognitive words, and social words. LIWC2015 seeks to group words into categories that can be used to analyze psycholinguistic features in texts. Researchers in various fields, including psychology, sociology, and computer science, have used LIWC2015 to study a wide range of topics, such as personality, emotional expression, deception, and social interaction. LIWC2015 has also been used in various relevant studies on sentiment analysis. It provides with a summary variable "Tone" that combines positive and negative dimensions (*posemo* and *negemo*) into a single one. The higher the tone, the more positive it is. The tone ranges from 0 to 100. Numbers <50 indicate a more negative emotional tone. The default LIWC2015 Dictionary contains approximately 6400 words, word stems, and select emoticons.

SentiStrength

SentiStrength is a sentiment analysis tool that assigns scores to words and phrases based on their positive or negative sentiment [94-96]. It calculates an overall sentiment score for the text by combining these individual scores. This tool can provide dual-, binary-, trinary-, or single-scale results. In this study, a single scale ranging from −4 (extremely negative) to 4 (extremely positive) was chosen, with 0 indicating neutral sentiment. SentiStrength uses linguistic and lexicon-based methods. Linguistic methods involve rules and heuristics for identifying sentiment-bearing words and phrases, including cues such as repeated punctuation, emoticons, negations, and capital letters. The lexicon used consists of 2546 terms associated with polarity and intensity. Part of the lexicon was added from General Inquirer, including word roots such as "extrem*" to recognize variants. Training data sets included posts from various platforms such as BBC Forum, Twitter, YouTube, Digg.com, MySpace, and Runners World.

TextBlob

TextBlob is a Python library used in NLP tasks [11,74], such as part-of-speech tagging, sentiment analysis, and noun phrase extraction. TextBlob outputs a polarity score ranging from −1 to 1. A negative score signifies a negative sentiment, a positive score indicates a positive sentiment, and a score of 0 represents a neutral sentiment. TextBlob includes 2 analysis approaches: a rule-based model and a supervised machine learning naïve Bayes classifier model.

VADER

VADER [12,97] is a rule-based model designed for analyzing sentiment in social media text. It uses 5 rules based on grammatical and syntactical patterns to determine sentiment intensity. These rules involve punctuation, capitalization, degree modifiers, conjunctions such as "but," and trigram evaluation to identify negations that can affect polarity. VADER was developed and validated using a gold standard list of lexical features, including LIWC, General Inquirer, and Affective Norms for English Words. The model was trained on various data sets, including tweets, New York Times opinions, movie reviews, and Amazon product reviews.

Stanza

Stanza is an open-source Python library that provides several methods for performing NLP tasks [13,98], including part-of-speech tagging, named entity recognition, dependency parsing, and sentiment analysis. Stanza's sentiment analysis module assigns a positive, negative, or neutral sentiment score (0, 1, or 2, respectively) to each sentence in a given text. Stanza's sentiment analysis tool is based on a convolutional neural network model using the vectors trained by Mikolov et al [99] on 100 billion words from Google News as well as a combination of lexical and syntactic features. It was trained on large data sets including movie reviews and the Stanford Sentiment Treebank. Unlike other methods, Stanza includes preprocessing of its own (sentence splitter and tokenizer).

TweetEval

TweetEval is a benchmarking platform for Twitter-specific classification tasks [100]. TweetEval consists of 7 NLP tasks: irony detection, offensive language detection, emoji prediction, emotion recognition, hate speech detection, stance detection, and sentiment analysis. Using TweetEval, a common set of evaluation metrics and data set, researchers and practitioners can compare the performance of different models on the same tasks and identify the most effective models for different NLP applications. TweetEval provides a leaderboard for ranking the performance of different models on the sentiment analysis task. The leaderboard is based on the F_1-score. TweetEval returns 3 labels (positive, negative, and neutral) associated with a weight. TweetEval sentiment analysis is based on the RoBERTa model, an LLM based on BERT (trained on 58M tweets), and fine-tuned on the SemEval 2017 sentiment analysis data set (approximately 40,000 tweets) [91].

Pysentimiento

Pysentimiento is an open-source Python library that includes models for sentiment analysis and social NLP tasks, such as hate speech detection, irony detection, emotion analysis, named entity recognition, and part-of-speech tagging, in several languages such as English, Spanish, Portuguese, and Italian [101,102]. The English model for sentiment analysis is based on BERTweet [103], a RoBERTa model, trained on English tweets and also fine-tuned on the SemEval 2017 sentiment analysis data set [91]. Pysentimiento returns 3 polarity labels per text associated with a weight.

NLPTown

NLPTown [104] is a sentiment analysis application based on a BERT-base-multilingual-uncased model, fine-tuned for sentiment analysis on product reviews for 6 languages (English, Dutch, German, French, Spanish, and Italian), and predicts the sentiment of the review as the number of stars (1-5).

Few-Shot Learning With Open Pre-Trained Transformers Language Models

As mentioned previously, few-shot learning seeks to address the challenge of sentiment analysis when only a small amount of labeled data is available for training. In traditional supervised learning, models are trained on large data sets with many labeled examples. However, in some applications such as sentiment analysis, labeled survey data are scarce or expensive to obtain, making it difficult to train accurate models. In this study, we used the Open Pre-Trained Transformers (OPT) [105], a suite of decoder-only pre-trained transformers ranging from 125M to 175B parameters created by Meta AI. OPT has been used in several applications but has never been applied to sentiment analysis. This model has shown to perform similarly to the GPT-3 [60] on several NLP tasks. The OPT model was built using a data set of 180B tokens. This represents approximately 23% (180B/780B) of the amount of data set tokens used for the Pathways Language Model [52]. The largest OPT model has comparable number of parameters to GPT-3 (175B parameters) [60], although we used all models except for the latter given graphics processing unit limitations. The novelty of OPT is its availability as open source (albeit only for academic research).

Zero-Shot Learning With ChatGPT

Zero-shot learning refers to the use of a model to perform a task for which it has not been explicitly trained. Thus, zero-shot learning for sentiment analysis recognizes and classifies sentiment in text without being explicitly provided with examples of sentiment labels. Instead, the model is trained on related tasks, such as language modeling or machine translation, which enables it to understand the underlying structure of the language and the context in which it is used. In this study, we used ChatGPT (based on GPT-3.5), which has significantly improved the performance of several NLP tasks. GPT-3.5 is a model with 175B parameters created by OpenAI and trained on a vast amount of text data sourced from the internet using both reinforcement and supervised learning techniques. For this paper, we generated a polarity score for each sentence *x* by asking ChatGPT "What is the sentiment of the following sentence 'x.'"

Ethical Considerations

The NIH survey was approved by the Institutional Review Board of the NIH (reference number 20MN085), and all participants provided consent for the study. The Stanford survey was approved by Stanford's Institutional Review Board (reference number 55436), and all participants provided consent for the study.

All survey data and responses in both the NIH and the Stanford data sets were anonymized and associated with a unique ID. Participants from both studies were not compensated for participating in the surveys.

Results

Evaluation Metrics

To assess the overall performance of the sentiment analysis tools, we evaluated the accuracy, macro *F*-measure, macro precision, and macro recall. Macro evaluation metrics were recommended in the NLP competition SemEval-2017 Task 4 [91].

Preparation of Applications for Evaluation

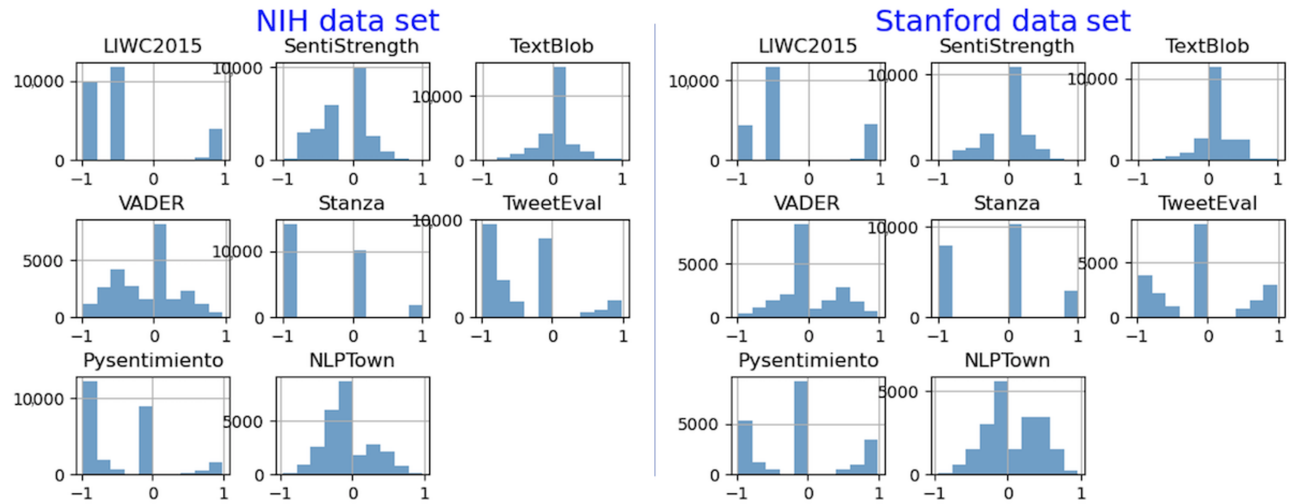
Harmonization of Applications' Outputs

The LIWC2015, Stanza, and SentiStrength applications produce outputs that are measured on distinct scales. LIWC2015 generates a continuous value ranging from 0 to 100; SentiStrength generates an integer score ranging from −4 to 4; and Stanza produces a discrete whole number score of 0, 1, or 2, which correspond to negative, neutral, and positive sentiments, respectively. Therefore, it is necessary to convert these scores to a common range of [−1, 1], as formally defined in equation 1.

$$score'(x) = 2 \times (score[x] - score[x]_{min}) / (score[x]_{max} - score[x]_{min}) - 1 \quad \textbf{(1)}$$

The distribution of sentiment scores across all tools is shown in Figure 3. We then classify all negative values as negative sentiment, all 0 values as neutral, and all positive values as positive sentiment. It is important to note that the VADER application uses a slightly different classification approach, considering a score ≤0.05 to be negative, a score between −0.05 and 0.05 to be neutral, and a score ≥0.05 to be positive.

Figure 3. Distribution of sentiment scores across all applications on the National Institutes of Health (NIH) and Stanford data sets. LIWC2015: Linguistic Inquiry and Word Count 2015; VADER: Valence Aware Dictionary and Sentiment Reasoner.



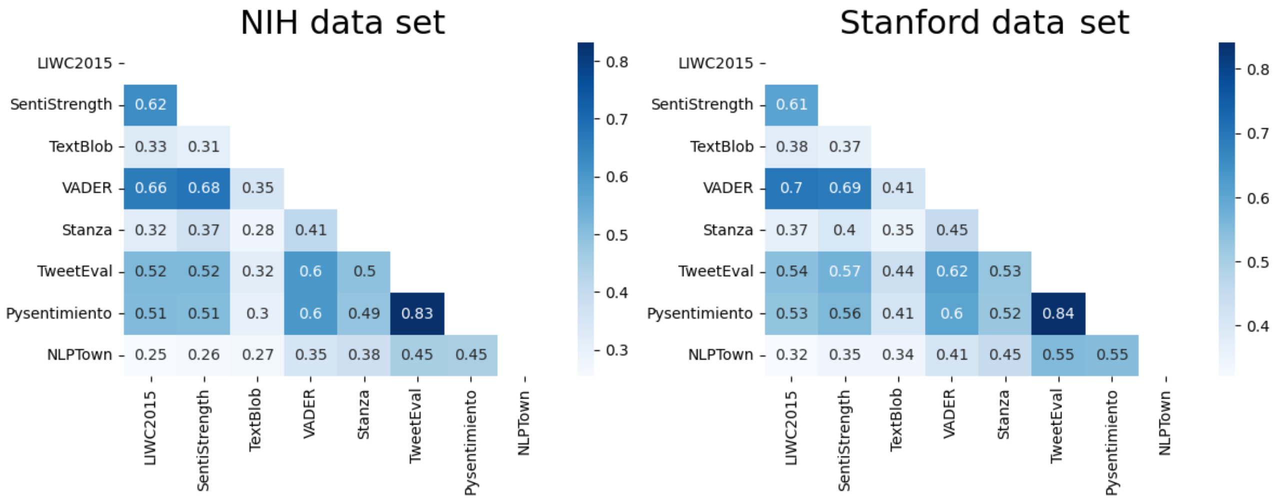
Fine-Tuning for Few-Shot Learning

We used few-shot learning using our small amount of training data to fine-tune the OPT models, rather than training them from scratch. For this experiment, the training data set was split into 85% (110/130) for feeding the model and 15% (20/130) for validation. Given the memory constraints, we considered only OPT 125M, 350M, 1.3B, and 2.7B. We performed a hyperparameter search to optimize the performance of the model on sentiment analysis. We considered learning rate=[3×10^{-4}, 1×10^{-4}, 3×10^{-5}, and 1×10^{-5}], batch size=[4, 8, 16, and 32], number of epochs from 1 to 7, and the AdamW optimizer. The models that performed the best were OPT-1.3B and OPT-2.7B, using a learning rate of 1×10^{-5}, a batch size of 32, and 5 epochs. These were the models used to obtain the test set results reported in next subsections.

Experiment 1: Correlation Between the Outputs of Applications

The objective was to evaluate the agreement level among various methods for predicting the sentiments of COVID-19 survey responses. Understanding the methods' agreement or divergence was crucial in determining the reliability and accuracy of predictions, allowing for accurate studies of the relationship between language use and mental health. The Pearson correlation coefficient was used to assess the reliability of the tools, as shown in Figure 4. Disagreement among the methods prompted us to evaluate few-shot learning to obtain high-quality predictions.

Figure 4. Pearson correlation matrix of score applications on the National Institutes of Health (NIH) and Stanford data sets. LIWC2015: Linguistic Inquiry and Word Count 2015; VADER: Valence Aware Dictionary and Sentiment Reasoner.



Experiment 2: Prediction of Sentiment Scores

Tables 3 and 4 show the performance results obtained by all applications, few-shot learning, and zero-shot learning techniques on the NIH and Stanford test data sets, respectively. Both test sets comprised 500 sentences each, as detailed in the *Data* section. The top 2 performance results are italicized. Of note, a perfect classifier that accurately categorizes all items obtains a value of 1, whereas a perverse classifier that misclassifies all items achieves a value of 0. However, a trivial classifier that assigns all sentences to the same category (positive, negative, or neutral) and a random classifier both have a value of 0.3333.

ChatGPT achieved a significant improvement in sentiment analysis compared with other models through zero-shot learning. On the NIH data set, ChatGPT outperformed few-shot learning (OPT-1.3B and OPT-2.7B) by 6% in accuracy and 7% in *F*-measure. Similarly, on the Stanford data set, ChatGPT showed better results than the OPT-1.3B and OPT-2.7B models, with 6% higher accuracy and 4% higher *F*-measure.

Moreover, to further evaluate the sentiment analysis tools, we used Bayesian analysis, as recommended by Benavoli et al [106], to assess the statistical significance of the performance of the methods. Specifically, we applied the Bayesian signed-rank test [107] to compare the accuracies achieved across multiple data sets. This test quantifies the likelihood of observing the signed ranks of accuracy differences under both the null hypothesis (indicating no significant difference) and alternative hypothesis (indicating a significant difference). The Bayesian signed-rank test is designed to compare performance over multiple data sets (≥2); therefore, we further partitioned the independent Stanford and NIH data sets. Each data set was partitioned into 3 subsets, based on the sentiment label assigned to them, resulting in positive, neutral, and negative subsets for each data set.

This division was influenced by insights from our prior analysis, which highlighted inherent distinctions among sentences associated with positive, neutral, and negative labels. For instance, positive sentences exhibited a preponderance of positive adjectives, whereas negative sentences featured more negative adjectives, and neutral sentences tended to emphasize facts that are characteristic of the neutral category. Therefore, we assumed a degree of independence across subsets within each data set. The heat map diagram in Figure 5 shows the results of our Bayesian analysis, with cells corresponding to row *i* and column *j*. On the left side, “A higher than B” indicates the probability that method *i* performs better than classifier *j*. The center indicates the probability of practical equivalence between methods *i* and *j*. Similarly, on the right side, “B higher than A” indicates the probability that method *j* is better than classifier *i*. These experiments confirmed that ChatGPT performed better than all the other alternatives. The OPT models showed similar performance to methods other than ChatGPT and could be considered as a viable second option.

Table 3. Results on the National Institutes of Health (NIH) test data set.

Application	Precision	Recall	*F*-measure	Accuracy
LIWC2015[a]	0.2733	0.5226	0.3587	0.4540
SentiStrength	0.5732	0.6006	0.5814	0.6480
TextBlob	0.4505	0.4776	0.4053	0.4340
VADER[b]	0.6302	0.7036	0.6097	0.6580
Stanza	0.6178	0.5758	0.5886	0.6300
TweetEval	0.7818	*0.8318* [c]	0.7898	0.7840
Pysentimiento	0.7738	0.7780	0.7699	0.7760
NLPTown	0.4338	0.5173	0.4210	0.4520
OPT[d] 1.3B (few-shot)	0.8032	0.8000	0.7992	0.8000
OPT 2.7B (few-shot)	*0.8061*	0.8040	*0.8050*	*0.8040*
ChatGPT (zero-shot)	*0.8526*	*0.8926*	*0.8668*	*0.8600*
All negative	0.1487	0.3333	0.2056	0.4460
All neutral	0.1547	0.3333	0.2113	0.4640
All positive	0.0300	0.3333	0.0550	0.0900

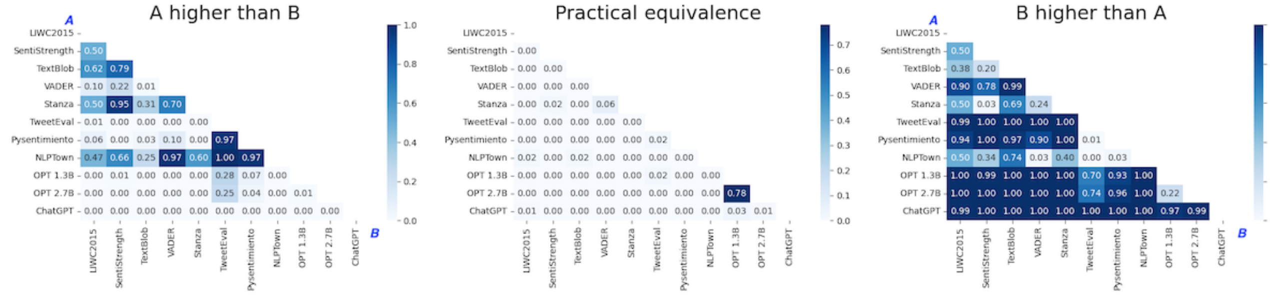
[a]LIWC2015: Linguistic Inquiry and Word Count 2015.
[b]VADER: Valence Aware Dictionary and Sentiment Reasoner.
[c]Italicization represents the top 2 performance results.
[d]OPT: Open Pre-Trained Transformers.

Table 4. Results on Stanford test data set.

Application	Precision	Recall	*F*-measure	Accuracy
LIWC2015[a]	0.3752	0.4391	0.3890	0.5400
SentiStrength	0.5738	0.5561	0.5335	0.5420
TextBlob	0.4757	0.4872	0.4527	0.4600
VADER[b]	0.5875	0.5919	0.5755	0.5840
Stanza	0.5975	0.4987	0.4859	0.5040
TweetEval	0.7366	0.7178	0.7090	0.7200
Pysentimiento	0.6731	0.6362	0.6267	0.6440
NLPTown	0.5163	0.5192	0.5056	0.5420
OPT[c] 1.3B (few-shot)	*0.8323* [d]	*0.8160*	*0.8211*	*0.8160*
OPT 2.7B (few-shot)	0.8288	0.8100	0.8147	0.8100
ChatGPT (zero-shot)	*0.8632*	*0.8779*	*0.8662*	*0.8740*
All negative	0.1560	0.3333	0.2125	0.4680
All neutral	0.0780	0.3333	0.1264	0.2340
All positive	0.0993	0.3333	0.1531	0.2980

[a]LIWC2015: Linguistic Inquiry and Word Count 2015.
[b]VADER: Valence Aware Dictionary and Sentiment Reasoner.
[c]OPT: Open Pre-Trained Transformers.
[d]Italicization represents the top 2 performance results.

Figure 5. Bayesian analysis conducted on accuracy performances of 11 sentiment analysis methods across 6 different subsets. LIWC2015: Linguistic Inquiry and Word Count 2015; OPT: Open Pre-Trained Transformers; VADER: Valence Aware Dictionary and Sentiment Reasoner.



Discussion

Principal Findings

Our primary objective was to assess various sentiment analysis tools for the purposed of predicting the sentiments of survey responses during the COVID-19 pandemic. Obtaining a thorough understanding of the tools' degree of agreement, as shown in Figure 4, was crucial for determining whether they could be used as surrogates for human labeling. The disagreement between tools led us to try ensemble methods to produce more reliable ratings. Fine-tuned BERT models such as TweetEval and Pysentimiento outperformed the other baseline methods. Fine-tuned methods have the ability to learn domain-specific patterns from text, resulting in better performance than lexicon- and rule-based methods. However, these techniques often require large training data sets to achieve optimal performance, such as the 40k tweet data set used to train TweetEval and Pysentimiento.

As part of the process of determining agreement between tools, we labeled a small data set (260 sentences), which is what prompted us to consider the possibility of using few-shot and zero-shot learning techniques. We then investigated the performance of OPT, which is unexplored in sentiment analysis, for few-shot learning using a small training data set (260 sentences). The OPT-1.3B and OPT-2.7B models surpassed all the baseline methods as well as the fine-tuned BERT models. This highlighted the potential of few-shot learning in dealing with scarce annotated data and the effectiveness of few-shot learning. Although better results could have been achieved with a larger training set, these experiments primarily aimed to investigate the potential of OPT using limited annotated data. The potential is to be able to produce models tailored to specific research applications, with only a small time investment by domain experts. We believe that these models can significantly contribute to the sentiment analysis of health- and clinical-related surveys and can be further fine-tuned with additional data and optimized hyperparameters.

Our investigation also encompassed zero-shot learning with ChatGPT, which exhibited remarkable performance compared with all other models, including few-shot learning with OPT, as presented in Tables 3 and 4. Note that GPT-3.5—the model behind ChatGPT—is trained on related tasks, such as language modeling or machine translation. This enabled it to understand the underlying structure of sentiment-related language and the context in which it is used. Moreover, the necessity for manual text annotations in sentiment analysis tasks makes ChatGPT and other LLMs particularly attractive. As demonstrated by Ziems et al [108], LLMs can alleviate the workload of human annotators in a zero-shot manner, thereby enhancing the efficiency of social-science analysis. In addition, a study [109] found that ChatGPT outperformed crowd workers in various text annotation tasks, including assessing relevance, stance, topics, and frame detection. These findings suggest that there may be potential in using ChatGPT and other recent LLMs for annotation in clinical NLP and reserving human input for quality control. Sentiment analysis tools based on LLMs, such as ChatGPT, automatically identify relevant features, reducing the need for manual engineering, which is a common requirement in tools such as LIWC 2015 and VADER. In addition, LLMs enable fine-tuning, allowing for potential adaptation to different sentiment analysis tasks (eg, in new domains) without the need for complete retraining. LLM-based tools can also capture longer-range context for more accurate sentiment assessment.

Limitations

There exist several limitations and risks of ChatGPT and other non–open-source LLMs regarding protected health information (PHI). Non–open-source LLMs require sending information to an external server and do not provide transparency into how they handle PHI, making it difficult to assess how the model is processing and protecting sensitive information. They may also have security vulnerabilities that can be exploited to gain unauthorized access to PHI. Note also that LLMs are not specifically designed for sentiment analysis, which may sometimes lead to errors, for instance, subtle sarcasm such as "Oh yes, great job!," context-dependent negation as in "The vaccine was not as bad as I thought," and idiomatic expressions such as "It's a piece of cake." They may encounter difficulties with nuanced health-related terminology and concepts. Therefore, specialized health terminology may require additional adaptation beyond general text fine-tuning, for instance, medical abbreviations and acronyms such as "The patient teared up because of a significant increase in their CD4 count" and "So, my mom's HbA1c levels have improved after insulin therapy." In addition, although several outputs may sound plausible, they may occasionally be incorrect. In our view, the output of LLMs should not be used without a plan for human quality control (eg, via sampling) or mitigation (eg, repeated validation). This is crucial for ensuring the accuracy and reliability of the generated content, as LLMs may produce results that require refinement or correction before dissemination. Moreover, there are constraints on the ability to access ChatGPT via its application programming interface, and this may make it too costly or time-consuming to do so. Therefore, researchers and health care practitioners might also opt to use an open-source language model for their NLP-related projects, such as OPT, which can be run on site and perform well on sentiment analysis.

Finally, our study focused on using surveys to understand people's feelings, specifically regarding COVID-19, which was a very important topic at the time. Thus, our conclusions apply specifically to discussions about COVID-19 and may not be true for other subjects. In addition, it is important to highlight that the Stanford data set has an implicit polarity bias: it specifically asks for positive effects ("Although this is a challenging time, can you tell us about any positive effects or 'silver linings' you have experienced during this crisis?") and difficulties ("Have you experienced any difficulties due to the coronavirus crisis?"). The NIH data set poses a single, less-biased question. Therefore, it is crucial to be careful when generalizing our findings beyond the scope of COVID-19 during the studied time frame.

Acknowledgments

The authors would like to thank Kelsey Yeh and Nathaniel Menon, who worked as research assistants in the Social Media Lab at Stanford University, for their assistance in annotating the Stanford test data set. The authors would also like to express gratitude to the many respondents who shared their experiences in both studies. The research reported in this publication was supported in part by the Intramural Research Programs of the National Institute of Mental Health (ZIAMH002922: Principal Investigator JC and ZIC-MH002968: Principal Investigator FP), National Center for Complementary and Integrative Health (ZIA-AT000035: Principal Investigator LA), and National Institute of Arthritis and Musculoskeletal and Skin Diseases (K24AR075060: Principal Investigator EL). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The test data sets generated and analyzed during this study are deidentified and freely available in the FigShare repository [89]. The source code for fine-tuning the OPT models and using ChatGPT in the experiments conducted in this study is publicly accessible on GitHub [110].

Authors' Contributions

JALV and FP contributed to conceiving the study idea and design. LA and JC led the collection of the National Institutes of Health (NIH) data set, whereas EL led the collection of the Stanford data set. The annotation of the training NIH and Stanford data sets was conducted by RW and AYL, respectively. The annotation of the NIH test data set was conducted by LA, RW, and EPG, whereas AYL and 2 research assistants annotated the Stanford test data set. JALV set up the applications and performed the evaluation. JALV and FP wrote the initial draft and revised the subsequent versions. All authors read, revised, and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Feldman R. Techniques and applications for sentiment analysis. Commun ACM 2013 Apr;56(4):82-89 [FREE Full text] [doi: 10.1145/2436256.2436274]
2. Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J 2014 Dec;5(4):1093-1113. [doi: 10.1016/j.asej.2014.04.011]
3. Birjali M, Kasri M, Beni-Hssane A. A comprehensive survey on sentiment analysis: approaches, challenges and trends. Knowl Based Syst 2021 Aug;226:107134. [doi: 10.1016/j.knosys.2021.107134]
4. Tsao SF, Chen H, Tisseverasinghe T, Yang Y, Li L, Butt ZA. What social media told us in the time of COVID-19: a scoping review. Lancet Digit Health 2021 Mar;3(3):e175-e194. [doi: 10.1016/s2589-7500(20)30315-0]
5. Alamoodi AH, Zaidan BB, Zaidan AA, Albahri OS, Mohammed KI, Malik RQ, et al. Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: a systematic review. Expert Syst Appl 2021 Apr 01;167:114155 [FREE Full text] [doi: 10.1016/j.eswa.2020.114155] [Medline: 33139966]
6. He L, He C, Reynolds TL, Bai Q, Huang Y, Li C, et al. Why do people oppose mask wearing? a comprehensive analysis of U.S. tweets during the COVID-19 pandemic. J Am Med Inform Assoc 2021 Jul 14;28(7):1564-1573 [FREE Full text] [doi: 10.1093/jamia/ocab047] [Medline: 33690794]
7. Wang X, Hegde S, Son C, Keller B, Smith A, Sasangohar F. Investigating mental health of US college students during the COVID-19 pandemic: cross-sectional survey study. J Med Internet Res 2020 Sep 17;22(9):e22817 [FREE Full text] [doi: 10.2196/22817] [Medline: 32897868]

8. Daly M, Sutin AR, Robinson E. Longitudinal changes in mental health and the COVID-19 pandemic: evidence from the UK Household Longitudinal Study. Psychol Med 2022 Oct;52(13):2549-2558 [FREE Full text] [doi: 10.1017/S0033291720004432] [Medline: 33183370]
9. Schuster C, Weitzman L, Sass Mikkelsen K, Meyer-Sahling J, Bersch K, Fukuyama F, et al. Responding to COVID-19 through surveys of public servants. Public Adm Rev 2020;80(5):792-796 [FREE Full text] [doi: 10.1111/puar.13246] [Medline: 32836447]
10. Lossio-Ventura JA, Lee AY, Hancock JT, Linos N, Linos E. Identifying silver linings during the pandemic through natural language processing. Front Psychol 2021 Sep 3;12:712111 [FREE Full text] [doi: 10.3389/fpsyg.2021.712111] [Medline: 34539512]
11. TextBlob: simplified text processing. TextBlob. URL: https://textblob.readthedocs.io/en/ [accessed 2022-03-15]
12. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proc Int AAAI Conf Web Soc Media 2014 May 16;8(1):216-225. [doi: 10.1609/icwsm.v8i1.14550]
13. Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD. Stanza: a python natural language processing toolkit for many human languages. 2020 Presented at: 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations; July 5-10, 2020; Online URL: https://aclanthology.org/2020.acl-demos.14 [doi: 10.18653/v1/2020.acl-demos.14]
14. He L, Zheng K. How do general-purpose sentiment analyzers perform when applied to health-related online social media data? Stud Health Technol Inform 2019 Aug 21;264:1208-1212 [FREE Full text] [doi: 10.3233/SHTI190418] [Medline: 31438117]
15. He L, Yin T, Hu Z, Chen Y, Hanauer DA, Zheng K. Developing a standardized protocol for computational sentiment analysis research using health-related social media data. J Am Med Inform Assoc 2021 Jun 12;28(6):1125-1134 [FREE Full text] [doi: 10.1093/jamia/ocaa298] [Medline: 33355353]
16. He L, Yin T, Zheng K. They may not work! an evaluation of eleven sentiment analysis tools on seven social media datasets. J Biomed Inform 2022 Aug;132:104142 [FREE Full text] [doi: 10.1016/j.jbi.2022.104142] [Medline: 35835437]
17. Liu S, Liu J. Public attitudes toward COVID-19 vaccines on English-language Twitter: a sentiment analysis. Vaccine 2021 Sep 15;39(39):5499-5505 [FREE Full text] [doi: 10.1016/j.vaccine.2021.08.058] [Medline: 34452774]
18. ChatGPT homepage. ChatGPT. URL: https://chat.openai.com/ [accessed 2023-04-17]
19. Zagibalov T, Carroll J. Automatic seed word selection for unsupervised sentiment classification of Chinese text. In: Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1. 2008 Presented at: COLING '08; August 18-22, 2008; Manchester, UK. [doi: 10.3115/1599081.1599216]
20. Ofoghi B, Mann M, Verspoor K. Towards early discovery of salient health threats: a social media emotion classification technique. 2016 Presented at: Biocomputing 2016: Proceedings of the Pacific Symposium; January 4-8, 2016; Kohala Coast, HI. [doi: 10.1142/9789814749411_0046]
21. Davis MA, Zheng K, Liu Y, Levy H. Public response to Obamacare on Twitter. J Med Internet Res 2017 May 26;19(5):e167 [FREE Full text] [doi: 10.2196/jmir.6946] [Medline: 28550002]
22. Pavlopoulos J, Androutsopoulos I. Multi-granular aspect aggregation in aspect-based sentiment analysis. 2014 Presented at: 14th Conference of the European Chapter of the Association for Computational Linguistics; April 26-30, 2014; Gothenburg, Sweden. [doi: 10.3115/v1/e14-1009]
23. Livas C, Delli K, Pandis N. "My invisalign experience": content, metrics and comment sentiment analysis of the most popular patient testimonials on YouTube. Prog Orthod 2018 Jan 22;19(1):3 [FREE Full text] [doi: 10.1186/s40510-017-0201-1] [Medline: 29354889]
24. Pennebaker JW, Francis ME, Booth RJ. Linguistic Inquiry and Word Count: LIWC. Mahwah, NJ: Lawrence Erlbaum Associates; 2001.
25. Bradley MM, Lang PJ. Affective norms for English words (ANEW): instruction manual and affective ratings. The Center for Research in Psychophysiology, University of Florida. 1999. URL: https://pdodds.w3.uvm.edu/teaching/courses/2009-08UVM-300/docs/others/everything/bradley1999a.pdf [accessed 2024-01-09]
26. Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). 2010 Presented at: LREC'10; May 17-23, 2010; Valletta, Malta URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/769
27. Lu Y, Castellanos M, Dayal U, Zhai C. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In: Proceedings of the 20th international conference on World Wide Web. 2011 Presented at: WWW '11; March 28-April 1, 2011; Hyderabad, India URL: https://doi.org/10.1145/1963405.1963456 [doi: 10.1145/1963405.1963456]
28. Huang S, Niu Z, Shi C. Automatic construction of domain-specific sentiment lexicon based on constrained label propagation. Knowl Based Syst 2014 Jan;56:191-200. [doi: 10.1016/j.knosys.2013.11.009]
29. Miller GA. WordNet: a lexical database for English. Commun ACM 1995 Nov;38(11):39-41. [doi: 10.1145/219717.219748]
30. Pedersen T. Rule-based and lightly supervised methods to predict emotions in suicide notes. Biomed Inform Insight 2012 Jan 30;5s1. [doi: 10.4137/bii.s8953]
31. Wang W, Chen L, Tan M, Wang S, Sheth AP. Discovering fine-grained sentiment in suicide notes. Biomed Inform Insight 2012 Jan 30;5s1. [doi: 10.4137/bii.s8963]

32. Desai T, Shariff A, Shariff A, Kats M, Fang X, Christiano C, et al. Tweeting the meeting: an in-depth analysis of Twitter activity at Kidney Week 2011. PLoS One 2012 Jul 5;7(7):e40253 [FREE Full text] [doi: 10.1371/journal.pone.0040253] [Medline: 22792254]
33. Cole-Lewis H, Varghese A, Sanders A, Schwarz M, Pugatch J, Augustson E. Assessing electronic cigarette-related tweets for sentiment and content using supervised machine learning. J Med Internet Res 2015 Aug 25;17(8):e208 [FREE Full text] [doi: 10.2196/jmir.4392] [Medline: 26307512]
34. Daniulaityte R, Chen L, Lamy FR, Carlson RG, Thirunarayan K, Sheth A. "When 'bad' is 'good'": identifying personal communication and sentiment in drug-related tweets. JMIR Public Health Surveill 2016 Oct 24;2(2):e162 [FREE Full text] [doi: 10.2196/publichealth.6327] [Medline: 27777215]
35. Appel O, Chiclana F, Carter J, Fujita H. A hybrid approach to the sentiment analysis problem at the sentence level. Knowl Based Syst 2016 Sep;108:110-124. [doi: 10.1016/j.knosys.2016.05.040]
36. Ahmad M, Aftab S, Ali I. Sentiment analysis of tweets using SVM. Int J Comput Appl 2017 Nov 15;177(5):25-29. [doi: 10.5120/ijca2017915758]
37. Gupta I, Joshi N. Enhanced Twitter sentiment analysis using hybrid approach and by accounting local contextual semantic. J Intell Syst 2019;29(1):1611-1625. [doi: 10.1515/jisys-2019-0106]
38. Zhang L, Wang S, Liu B. Deep learning for sentiment analysis: a survey. WIREs Data Min Knowl 2018;8(4):e1253. [doi: 10.1002/widm.1253]
39. Yadav A, Vishwakarma DK. Sentiment analysis using deep learning architectures: a review. Artif Intell Rev 2019 Dec 02;53(6):4335-4385. [doi: 10.1007/s10462-019-09794-5]
40. Zhou X, Wan X, Xiao J. Attention-based LSTM network for cross-lingual sentiment classification. 2016 Presented at: 2016 Conference on Empirical Methods in Natural Language Processing; November 1-5, 2016; Austin, Texas. [doi: 10.18653/v1/d16-1024]
41. Xu J, Chen D, Qiu X, Huang X. Cached long short-term memory neural networks for document-level sentiment classification. 2016 Presented at: 2016 Conference on Empirical Methods in Natural Language Processing; November 1-5, 2016; Austin, Texas URL: https://aclanthology.org/D16-1172 [doi: 10.18653/v1/d16-1172]
42. Ma Y, Peng H, Cambria E. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. 2018 Presented at: Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18); February 2-7, 2018; New Orleans, LA. [doi: 10.1609/aaai.v32i1.12048]
43. Rehman AU, Malik AK, Raza B, Ali W. A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis. Multimed Tools Appl 2019 Jun 11;78(18):26597-26613. [doi: 10.1007/s11042-019-07788-7]
44. Ouyang X, Zhou P, Li CH, Liu L. Sentiment analysis using convolutional neural network. 2015 Presented at: IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing; October 26-28, 2015; Liverpool, UK. [doi: 10.1109/cit/iucc/dasc/picom.2015.349]
45. Tang D, Qin B, Liu T. Document modeling with gated recurrent neural network for sentiment classification. 2015 Presented at: 2015 Conference on Empirical Methods in Natural Language Processing; September 17-21, 2015; Lisbon, Portugal URL: https://aclanthology.org/D15-1167 [doi: 10.18653/v1/d15-1167]
46. Chen T, Xu R, He Y, Wang X. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. Expert Syst Appl 2017 Apr;72:221-230. [doi: 10.1016/j.eswa.2016.10.065]
47. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AL, et al. Attention is all you need. arXiv Preprint posted online on June 12, 2017. [FREE Full text]
48. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); June 2-7, 2019; Minneapolis, MN URL: https://aclanthology.org/N19-1423 [doi: 10.18653/v1/n18-2]
49. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv Preprint posted online on July 26, 2019. [doi: 10.1090/mbk/121/79]
50. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. 2018 Presented at: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers); June 1-6, 2018; New Orleans, LA URL: https://aclanthology.org/N18-1202 [doi: 10.18653/v1/n18-1202]
51. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. OpenAI. URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf [accessed 2024-01-08]
52. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, et al. PaLM: scaling language modeling with pathways. arXiv Preprint posted online on April 5, 2022.
53. Sun C, Huang L, Qiu X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. arXiv Preprint posted online on March 22, 2019. [FREE Full text] [doi: 10.18653/v1/n18-2043]

54. Bataa E, Wu J. An investigation of transfer learning-based sentiment analysis in Japanese. 2019 Presented at: 57th Annual Meeting of the Association for Computational Linguistics; July 28-August 2, 2019; Florence, Italy URL: https://aclanthology.org/P19-1458 [doi: 10.18653/v1/p19-1458]
55. Yin D, Meng T, Chang KW. SentiBERT: a transferable transformer-based architecture for compositional sentiment semantics. 2020 Presented at: 58th Annual Meeting of the Association for Computational Linguistics; July 5-10, 2020; Online URL: https://aclanthology.org/2020.acl-main.341 [doi: 10.18653/v1/2020.acl-main.341]
56. Büyüköz B, Hürriyetoğlu A, Özgür A. Analyzing ELMo and DistilBERT on socio-political news classification. 2020 Presented at: Workshop on Automated Extraction of Socio-political Events from News 2020; May 12, 2020; Marseille, France URL: https://aclanthology.org/2020.aespen-1.4
57. Zhang L, Fan H, Peng C, Rao G, Cong Q. Sentiment analysis methods for HPV vaccines related tweets based on transfer learning. Healthcare (Basel) 2020 Aug 28;8(3):307 [FREE Full text] [doi: 10.3390/healthcare8030307] [Medline: 32872330]
58. Dai J, Yan H, Sun T, Liu P, Qiu X. Does syntax matter? A strong baseline for aspect-based sentiment analysis with RoBERTa. 2021 Presented at: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 6-11, 2021; Online URL: https://aclanthology.org/2021.naacl-main.146 [doi: 10.18653/v1/2021.naacl-main.146]
59. You L, Han F, Peng J, Jin H, Claramunt C. ASK-RoBERTa: a pretraining model for aspect-based sentiment classification via sentiment knowledge mining. Knowl Based Syst 2022 Oct 11;253:109511. [doi: 10.1016/j.knosys.2022.109511]
60. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. arXiv Preprint posted online on May 28, 2020. [FREE Full text]
61. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. arXiv Preprint posted online on May 24, 2022. [FREE Full text]
62. Rae JW, Borgeaud S, Cai T, Millican K, Hoffmann J, Song F, et al. Scaling language models: methods, analysis and insights from training gopher. arXiv Preprint posted online on December 8, 2021. [FREE Full text]
63. Rodríguez-Ibánez M, Casánez-Ventura A, Castejón-Mateos F, Cuenca-Jiménez PM. A review on sentiment analysis from social media platforms. Expert Syst Appl 2023 Aug;223:119862. [doi: 10.1016/j.eswa.2023.119862]
64. Zunic A, Corcoran P, Spasic I. Sentiment analysis in health and well-being: systematic review. JMIR Med Inform 2020 Jan 28;8(1):e16023 [FREE Full text] [doi: 10.2196/16023] [Medline: 32012057]
65. Bian J, Zhao Y, Salloum RG, Guo Y, Wang M, Prosperi M, et al. Using social media data to understand the impact of promotional information on laypeople's discussions: a case study of lynch syndrome. J Med Internet Res 2017 Dec 13;19(12):e414 [FREE Full text] [doi: 10.2196/jmir.9266] [Medline: 29237586]
66. Lim S, Tucker CS, Kumara S. An unsupervised machine learning model for discovering latent infectious diseases using social media data. J Biomed Inform 2017 Feb;66:82-94 [FREE Full text] [doi: 10.1016/j.jbi.2016.12.007] [Medline: 28034788]
67. Gohil S, Vuik S, Darzi A. Sentiment analysis of health care tweets: review of the methods used. JMIR Public Health Surveill 2018 Apr 23;4(2):e43 [FREE Full text] [doi: 10.2196/publichealth.5789] [Medline: 29685871]
68. Shofiya C, Abidi S. Sentiment analysis on COVID-19-related social distancing in Canada using Twitter data. Int J Environ Res Public Health 2021 Jun 03;18(11):5993 [FREE Full text] [doi: 10.3390/ijerph18115993] [Medline: 34204907]
69. Es-Sabery F, Es-Sabery K, Qadir J, Sainz-De-Abajo B, Hair A, Garcia-Zapirain B, et al. A MapReduce opinion mining for COVID-19-related tweets classification using enhanced ID3 decision tree classifier. IEEE Access 2021;9:58706-58739. [doi: 10.1109/access.2021.3073215]
70. Basiri ME, Nemati S, Abdar M, Asadi S, Acharrya UR. A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets. Knowl Based Syst 2021 Sep 27;228:107242 [FREE Full text] [doi: 10.1016/j.knosys.2021.107242] [Medline: 36570870]
71. Yan C, Law M, Nguyen S, Cheung J, Kong J. Comparing public sentiment toward COVID-19 vaccines across Canadian cities: analysis of comments on Reddit. J Med Internet Res 2021 Sep 24;23(9):e32685 [FREE Full text] [doi: 10.2196/32685] [Medline: 34519654]
72. Chandra R, Krishna A. COVID-19 sentiment analysis via deep learning during the rise of novel cases. PLoS One 2021 Aug 19;16(8):e0255615 [FREE Full text] [doi: 10.1371/journal.pone.0255615] [Medline: 34411112]
73. Zang S, Zhang X, Xing Y, Chen J, Lin L, Hou Z. Applications of social media and digital technologies in COVID-19 vaccination: scoping review. J Med Internet Res 2023 Feb 10;25:e40057 [FREE Full text] [doi: 10.2196/40057] [Medline: 36649235]
74. Loria S. textblob Documentation. Release 0.15, 2. TextBlob. 2018. URL: https://textblob.readthedocs.io/en/dev/changelog.html [accessed 2024-01-09]
75. Yu S, Eisenman D, Han Z. Temporal dynamics of public emotions during the COVID-19 pandemic at the epicenter of the outbreak: sentiment analysis of Weibo posts from Wuhan. J Med Internet Res 2021 Mar 18;23(3):e27078 [FREE Full text] [doi: 10.2196/27078] [Medline: 33661755]
76. Melton CA, Olusanya OA, Ammar N, Shaban-Nejad A. Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: a call to action for strengthening vaccine confidence. J Infect Public Health 2021 Oct;14(10):1505-1512 [FREE Full text] [doi: 10.1016/j.jiph.2021.08.010] [Medline: 34426095]

77. Huangfu L, Mo Y, Zhang P, Zeng DD, He S. COVID-19 vaccine tweets after vaccine rollout: sentiment-based topic modeling. J Med Internet Res 2022 Feb 08;24(2):e31726 [FREE Full text] [doi: 10.2196/31726] [Medline: 34783665]
78. Boukobza A, Burgun A, Roudier B, Tsopra R. Deep neural networks for simultaneously capturing public topics and sentiments during a pandemic: application on a COVID-19 tweet data set. JMIR Med Inform 2022 May 25;10(5):e34306 [FREE Full text] [doi: 10.2196/34306] [Medline: 35533390]
79. Ali SH, Foreman J, Capasso A, Jones AM, Tozan Y, DiClemente RJ. Social media as a recruitment platform for a nationwide online survey of COVID-19 knowledge, beliefs, and practices in the United States: methodology and feasibility analysis. BMC Med Res Methodol 2020 May 13;20(1):116 [FREE Full text] [doi: 10.1186/s12874-020-01011-0] [Medline: 32404050]
80. Geldsetzer P. Use of rapid online surveys to assess people's perceptions during infectious disease outbreaks: a cross-sectional survey on COVID-19. J Med Internet Res 2020 Apr 2;22(4):e18790 [FREE Full text] [doi: 10.2196/18790] [Medline: 32240094]
81. Hlatshwako TG, Shah SJ, Kosana P, Adebayo E, Hendriks J, Larsson EC, et al. Online health survey research during COVID-19. Lancet Digit Health 2021 Feb;3(2):e76-e77. [doi: 10.1016/s2589-7500(21)00002-9]
82. Shuja J, Alanazi E, Alasmary W, Alashaikh A. COVID-19 open source data sets: a comprehensive survey. Appl Intell (Dordr) 2021;51(3):1296-1325 [FREE Full text] [doi: 10.1007/s10489-020-01862-6] [Medline: 34764552]
83. Ben-Ezra M, Hamama-Raz Y, Goodwin R, Leshem E, Levin Y. Association between mental health trajectories and somatic symptoms following a second lockdown in Israel: a longitudinal study. BMJ Open 2021 Sep 02;11(9):e050480 [FREE Full text] [doi: 10.1136/bmjopen-2021-050480] [Medline: 34475179]
84. de Koning R, Egiz A, Kotecha J, Ciuculete AC, Ooi SZ, Bankole ND, et al. Survey fatigue during the COVID-19 pandemic: an analysis of neurosurgery survey response rates. Front Surg 2021 Aug 12;8:690680 [FREE Full text] [doi: 10.3389/fsurg.2021.690680] [Medline: 34458314]
85. Wang Z, Xie Q, Ding Z, Feng Y, Xia R. Is ChatGPT a good sentiment analyzer? a preliminary study. arXiv Preprint posted online on April 10, 2023. [FREE Full text]
86. Chung JY, Gibbons A, Atlas L, Ballard E, Ernst M, Japee S, et al. COVID-19 and mental health: predicted mental health status is associated with clinical symptoms and pandemic-related psychological and behavioral responses. medRxiv Preprint posted online on October 14, 2021 (forthcoming) [FREE Full text] [doi: 10.1101/2021.10.12.21264902] [Medline: 34671781]
87. Weger R, Lossio-Ventura JA, Rose-McCandlish M, Shaw JS, Sinclair S, Pereira F, et al. Trends in language use during the COVID-19 pandemic and relationship between language use and mental health: text analysis based on free responses from a longitudinal study. JMIR Ment Health 2023 Mar 01;10:e40899 [FREE Full text] [doi: 10.2196/40899] [Medline: 36525362]
88. Nelson LM, Simard JF, Oluyomi A, Nava V, Rosas LG, Bondy M, et al. US public concerns about the COVID-19 pandemic from results of a survey given via social media. JAMA Intern Med 2020 Jul 01;180(7):1020-1022 [FREE Full text] [doi: 10.1001/jamainternmed.2020.1369] [Medline: 32259192]
89. Lossio-Ventura JA, Weger R, Lee A, Guinee E, Chung J, Atlas L, et al. Sentiment analysis test dataset created from two COVID-19 surveys: National Institutes of Health (NIH) and Stanford University. FigShare. 2023 Nov 23. URL: http://tinyurl.com/2s39yutm [accessed 2024-01-09]
90. Nakov P, Ritter A, Rosenthal S, Sebastiani F, Stoyanov V. SemEval-2016 task 4: sentiment analysis in Twitter. 2016 Presented at: 10th International Workshop on Semantic Evaluation (SemEval-2016); June 16-17, 2016; San Diego, California URL: https://aclanthology.org/S16-1001/ [doi: 10.18653/v1/s16-1001]
91. Rosenthal S, Farra N, Nakov P. SemEval-2017 task 4: sentiment analysis in Twitter. 2017 Presented at: 11th International Workshop on Semantic Evaluation (SemEval-2017); August 3-4, 2017; Vancouver, Canada URL: https://aclanthology.org/S17-2088 [doi: 10.18653/v1/s17-2088]
92. Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015. University of Texas at Austin. 2015. URL: https://repositories.lib.utexas.edu/server/api/core/bitstreams/b0d26dcf-2391-4701-88d0-3cf50ebee697/content [accessed 2024-01-09]
93. Introducing LIWC-22. Linguistic Inquiry and Word Count. URL: https://www.liwc.app/ [accessed 2022-03-15]
94. Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A. Sentiment strength detection in short informal text. J Am Soc Inf Sci Technol 2010 Dec 15;61(12):2544-2558. [doi: 10.1002/asi.21416]
95. Thelwall M, Buckley K, Paltoglou G. Sentiment strength detection for the social web. J Am Soc Inf Sci Technol 2011 Oct 13;63(1):163-173. [doi: 10.1002/asi.21662]
96. SentiStrength- sentiment strength detection in short texts. SentiStrength. URL: http://sentistrength.wlv.ac.uk/ [accessed 2022-03-15]
97. VADER-sentiment-analysis. GitHub. URL: https://github.com/cjhutto/vaderSentiment [accessed 2022-03-15]
98. Stanza – a python NLP package for many human languages. Stanza. URL: https://stanfordnlp.github.io/stanza/ [accessed 2022-03-15]
99. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. arXiv Preprint posted online on October 16, 2013. [FREE Full text]

100. Barbieri F, Camacho-Collados J, Espinosa-Anke L, Neves L. TweetEval: unified benchmark and comparative evaluation for tweet classification. arXiv Preprint posted online on October 23, 2020. [FREE Full text] [doi: 10.18653/v1/2020.findings-emnlp.148]
101. Pérez JM, Rajngewerc M, Giudici JC, Furman DA, Luque F, Alemany LA, et al. pysentimiento: a python toolkit for opinion mining and social NLP tasks. arXiv Preprint posted online on June 17, 2021. [FREE Full text]
102. pysentimiento. GitHub. URL: https://github.com/pysentimiento/pysentimiento [accessed 2022-03-15]
103. Nguyen DQ, Vu T, Nguyen AT. BERTweet: a pre-trained language model for English tweets. arXiv Preprint posted online on May 20, 2020. [FREE Full text] [doi: 10.18653/v1/2020.emnlp-demos.2]
104. nlptown / bert-base-multilingual-uncased-sentiment. Hugging Face. URL: https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment [accessed 2022-03-15]
105. Zhang S, Roller S, Goyal N, Artetxe M, Chen M, Chen S, et al. OPT: open pre-trained transformer language models. arXiv Preprint posted online on May 2, 2022. [FREE Full text]
106. Benavoli A, Corani G, Dems?ar J, Zaffalon M. Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. arXiv Preprint posted online on June 14, 2016.
107. Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 2006;7:1-30 [FREE Full text]
108. Ziems C, Held W, Shaikh O, Chen J, Zhang Z, Yang D. Can large language models transform computational social science? arXiv Preprint posted online on April 12, 2023 [FREE Full text] [doi: 10.1162/coli_a_00502]
109. Gilardi F, Alizadeh M, Kubli M. ChatGPT outperforms crowd workers for text-annotation tasks. Proc Natl Acad Sci USA 2023 Jul 25;120(30):e2305016120 [FREE Full text] [doi: 10.1073/pnas.2305016120] [Medline: 37463210]
110. juanlossio / sentiment_analysis. GitHub. URL: https://github.com/juanlossio/sentiment_analysis [accessed 2024-01-12]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
DNN: deep neural network
ELMo: Embeddings from Language Models
GPT: Generative Pre-trained Transformers
LIWC: Linguistic Inquiry and Word Count
LIWC2015: Linguistic Inquiry and Word Count 2015
LLM: large language model
NIH: National Institutes of Health
NLP: natural language processing
OPT: Open Pre-Trained Transformers
PaLM: Pathways Language Model
PHI: protected health information
RoBERTa: Robustly optimized Bidirectional Encoder Representations from Transformers approach
VADER: Valence Aware Dictionary and Sentiment Reasoner

Edited by J Torous; submitted 21.06.23; peer-reviewed by L He, R Gore; comments to author 02.09.23; revised version received 16.11.23; accepted 17.11.23; published 25.01.24.

Please cite as:
Lossio-Ventura JA, Weger R, Lee AY, Guinee EP, Chung J, Atlas L, Linos E, Pereira F
A Comparison of ChatGPT and Fine-Tuned Open Pre-Trained Transformers (OPT) Against Widely Used Sentiment Analysis Tools: Sentiment Analysis of COVID-19 Survey Data
JMIR Ment Health 2024;11:e50150
URL: https://mental.jmir.org/2024/1/e50150
doi:10.2196/50150
PMID:38271138

Original Paper

Patient Satisfaction With a Coach-Guided, Technology-Based Mental Health Treatment: Qualitative Interview Study and Theme Analysis

Ashley Helm Smith[1,2], MA; Hilary Touchett[1,2,3], RN, PhD; Patricia Chen[1,3], PhD; Terri Fletcher[1,2,4], PhD; Jennifer Arney[5], PhD; Julianna Hogan[1,2,4], PhD; Miryam Wassef[1,2,4], LCSW; Marylene Cloitre[6], PhD; Jan A Lindsay[1,2,4,7], PhD

[1]Houston Veteran Affairs Health Services Research and Development Center for Innovations in Quality, Effectiveness and Safety, Michael E DeBakey Veteran Affairs Medical Center, Houston, TX, United States

[2]South Central Mental Illness Research, Education and Clinical Center, Houston, TX, United States

[3]Department of Medicine, Baylor College of Medicine, Houston, TX, United States

[4]Menninger Department of Psychiatry and Behavioral Sciences, Baylor College of Medicine, Houston, TX, United States

[5]Department of Sociology, College of Human Sciences and Humanities, University of Houston Clear Lake, Houston, TX, United States

[6]National Center for Post-Traumatic Stress Disorder Dissemination and Training Division, Veteran Affairs Palo Alto Health Care System, Palo Alto, CA, United States

[7]Rice University's Baker Institute for Public Policy, Houston, TX, United States

Corresponding Author:
Ashley Helm Smith, MA
Houston Veteran Affairs Health Services Research and Development Center for Innovations in Quality, Effectiveness and Safety
Michael E DeBakey Veteran Affairs Medical Center
2002 Holcombe Blvd. (152)
Houston, TX, 77030
United States
Phone: 1 713 794 8601
Email: ashley.smith30@va.gov

Abstract

Background: Technology-based mental health interventions address barriers rural veterans face in accessing care, including provider scarcity and distance from the hospital or clinic. webSTAIR is a 10-module, web-based treatment based on Skills Training in Affective and Interpersonal Regulation, designed to treat posttraumatic stress disorder and depression in individuals exposed to trauma. Previous work has demonstrated that webSTAIR is acceptable to participants and effective at reducing symptoms of posttraumatic stress disorder and depression when delivered synchronously or asynchronously (over 5 or 10 sessions).

Objective: This study explored factors that lead to greater patient satisfaction with webSTAIR, a web-based, coach-guided intervention.

Methods: We analyzed qualitative interview data to identify themes related to patient satisfaction with webSTAIR delivered with synchronous video-based coaching.

Results: Four themes emerged from the data: (1) coaching provides accountability and support, (2) self-pacing offers value that meets individual needs, (3) participants like the comfort and convenience of the web-based format, and (4) technical issues were common but not insurmountable.

Conclusions: We conclude that participants valued the accountability, flexibility, and convenience of tech-based interventions with video-delivered coaching.

(JMIR Ment Health 2024;11:e50977) doi:10.2196/50977

KEYWORDS

coaching; digital treatment; interview; mental health; patient satisfaction; PTSD; qualitative assessment; qualitative methods; sentiment analysis; technology-based; telehealth; trauma; veterans; video telehealth; web-based treatment

Introduction

More than 1.7 million US veterans receive mental health care through the Veterans Health Administration (VHA) annually [1]. However, US veterans underuse mental health care and often experience barriers that impact access, such as provider scarcity [2], distance to clinic [3], inability to take off from work or school, and dependent care responsibilities [4-6]. Veterans living in rural areas make up about 24% of all veterans in the United States [7], and they have similar mental health care needs as urban veterans; however, rural veterans experience access barriers at higher rates [8,9]. Technology-based mental health interventions are a potential solution to address some of the barriers that rural veterans in the United States face.

Technology-based mental health interventions can be accessed from a smart phone or tablet. They range from fully automated to therapist-guided in person or by video, telephone, or SMS text messages [10-13]. Technology-based mental health interventions have demonstrated effectiveness in clinical trials, with moderate-to-high effect sizes [14-17]. However, satisfaction with a treatment plays a significant role in treatment effectiveness. Patient satisfaction with treatment is linked to greater adherence [18,19] and greater perceived improvement in clinical status [20,21]. Even when adherence to treatment is controlled for, satisfied patients benefit more from care than less satisfied patients [22,23]. In mental health settings, a strong therapeutic alliance is linked to greater patient satisfaction [24]. Further understanding of factors that contribute to patient satisfaction is critical to designing future tech-based interventions that demonstrate better engagement and retention and greater clinical benefit [25,26].

Studying satisfaction with mental health treatments is complex. There are a number of variables that can influence patient satisfaction with mental health treatment, such as patient, disease, provider, therapy, and environmental factors (eg, technology-mediated treatments) [22,27,28]. Although many teams evaluate overall satisfaction by collecting survey data and compiling ratings to quantify patients' levels of satisfaction, qualitative data provides a nuanced and contextualized understanding of satisfaction [22]. In 2 recent studies, our team found that having therapist support and the convenience of being able to access treatment on the web contributed to patient satisfaction with technology-based mental health interventions [29,30]. This study uses qualitative interviews with patients who completed an internet-based mental health treatment to explore factors that lead to greater patient satisfaction in such a program. This study builds on previous findings by exploring the benefits of using a coach and the ways web-based formats provide comfort and convenience, as well as increased access in some cases.

Methods

Study Design

webSTAIR is an internet-based intervention adapted from Skills Training in Affective and Interpersonal Regulation (STAIR), a treatment for trauma-exposed individuals with symptoms of posttraumatic stress disorder (PTSD) and depression [31]. The program was developed to engage rural women veterans, who have been found to be underrepresented in mental health services [32]. However, male veterans were also enrolled if they were interested and met the study criteria. webSTAIR offers 10 web-based modules of self-directed skills training on emotion regulation (eg, emotional awareness, emotional management, and distress tolerance) and interpersonal skills (eg, assertiveness, flexibility, and compassion for oneself and others).

In addition to the asynchronously delivered web-based content, participants were offered 5 or 10 synchronous video coaching sessions, which took place using video-telehealth, after the completion of 1 or 2 modules. Coaches were licensed Veterans Affairs (VA) mental health providers who underwent training in the intervention and attended weekly supervision with a certified STAIR trainer. Sessions with the coach lasted approximately 45-50 minutes and involved reviewing module content, discussing the applicability of concepts in the veteran's life, and strategizing on how to integrate skills into daily practice. The webSTAIR program was delivered as routine care in mental health outpatient clinics at 9 sites serving rural patients within the VHA. The qualitative data presented were collected as part of a naturalistic evaluation of the program.

Recruitment and Sample

Participants in the webSTAIR program were US veterans recruited from 9 VHA facilities across the country that serve rural patients. Referrals were either clinician- or self-initiated. Participants completed the program between September 2018 and March 2020. Eligibility for the study was determined based on an initial telephone screening. Patients were considered eligible if they reported a history of trauma exposure and screened positive on the Primary Care PTSD Screen (PC-PTSD; positive on 3 of 5 items [33]) or the 2-item Patient Health Questionnaire (PHQ-2; positive on 1 of 2 items [34]), indicating clinically significant symptoms of PTSD or depression. The following resulted in ineligibility for enrollment: cognitive impairment or psychosis, mania, primary drug or alcohol abuse, current domestic violence, concurrent trauma-focused treatment, residential care for PTSD within 1 year, or an inability to attend telemental health appointments by video.

Measures and Data Collection

Demographic information for all patients in the program was collected at the initial intake phase of the study. The qualitative data for this study were collected during a 1-time interview conducted following the completion of the webSTAIR program. Patients who completed the program were sent a letter explaining the purpose of the interview and study details. This letter also included a contact number that patients could call to decline participation. All patients who did not actively call to decline were then contacted by study staff and asked if they would like to participate. Study staff made at least 3 attempts to contact participants. Outreach resulted in a response rate of 30%, with 74 participants agreeing to participate. The demographic information of interviewees was linked using a study ID number.

The interview guide was piloted and refined during the pilot intervention phase of the study [30,35]. It contained both categorical and open-ended questions. One member of the study staff conducted the interview, while 1-2 staff members took detailed notes during the interview. While the study team did not record or transcribe interviews, note-takers were often able to capture responses nearly verbatim. Quotes in this paper are drawn from these notes.

Participants were asked questions about their experience working through the program and about their satisfaction with the intervention. Interviews were conducted by telephone by study staff, including the first author (AHS) and second author (HT), as well as other members of the study team (MW). Interviews lasted approximately 60 minutes. At the conclusion of the interview, the interview notes were consolidated, and the finalized interview notes were entered into a database. Interview protocols were reviewed by the principal investigator's affiliated institutions' institutional review board, designated as "quality improvement."

Data Analysis

Interview data were analyzed using strategies from qualitative content analysis and thematic analysis [36-39]. The interview notes for each participant were entered into a matrix in Microsoft Excel (Microsoft Corp) and organized by interview question. A list of the interview questions used in the analysis for this study can be found in Multimedia Appendix 1. Data coding and analysis were conducted by a master's-level researcher with extensive qualitative experience who also functioned as a research assistant on the study (AHS) and a postdoctoral nurse trained in qualitative methods (HT). A codebook was developed based on interview guide questions and refined based on emergent findings. The second author (HT) reviewed the codes assigned by the first author, either agreeing or disagreeing. Coders met to resolve any disagreements until consensus was reached. The themes were analyzed and further defined to describe participants' satisfaction with the intervention. The data are presented here by satisfaction with the web-based, coach-guided format of webSTAIR.

Ethical Considerations

Study procedures were reviewed and found to be exempt by the institutional review board for Baylor College of Medicine and Affiliated Hospitals. Data were collected under a quality improvement designation. Participant confidentiality was upheld with the use of a randomized participant ID number.

Results

Sample

The exit interview was completed by 74 participants, the majority of whom were rural (46/74, 62%), White or Caucasian (48/74, 65%), female (42/74, 57%), aged between 35 and 44 years (25/74, 34%), and had some college or a 2-year college degree (32/74, 44%). A detailed list of participant characteristics can be found in Table 1.

Table 1. Participant demographics.

Demographics	Coach 5 (n=43), n (%)	Coach 10 (n=31), n (%)	Total (N=74), n (%)
Sex			
Male	14 (33)	17 (55)	31 (42)
Female	29 (67)	13 (42)	42 (57)
Transgender	0 (0)	1 (3)	1 (1)
Age (years)			
25-34	8 (19)	5 (16)	13 (18)
35-44	16 (37)	9 (29)	25 (34)
45-54	8 (19)	7 (23)	15 (20)
55-64	9 (21)	7 (23)	16 (22)
65-74	2 (4)	3 (9)	5 (6)
Rurality			
Urban	16 (37)	12 (39)	28 (38)
Rural	27 (63)	19 (61)	46 (62)
Race			
White or Caucasian	24 (56)	24 (78)	48 (65)
Black or African American	6 (14)	2 (6)	8 (11)
Hispanic or Latino	4 (9)	2 (6)	6 (8)
American Indian	1 (2)	1 (4)	2 (2)
Mixed race or ethnicity	8 (19)	2 (6)	10 (14)
Employment status			
Full-time	12 (28)	13 (42)	25 (34)
Part-time	6 (13)	4 (13)	10 (13)
Not currently working for pay	11 (26)	5 (16)	16 (22)
Retired	14 (33)	9 (29)	23 (31)
Educational level			
Some high school	1 (2)	0 (0)	1 (1)
Earned high school degree	4 (9)	4 (13)	8 (11)
Some college or 2-year degree	20 (47)	12 (39)	32 (44)
Earned 4-year degree	14 (33)	10 (32)	24 (32)
Postgraduate	3 (7)	5 (16)	8 (11)
Missing	1 (2)	0 (0)	1 (1)

Themes

Participants responded to categorical questions that they were generally satisfied with the webSTAIR program; that is, the majority of interview participants felt that webSTAIR met their needs and that they would use a similar program again in the future. Four themes emerged from the open-ended responses and are detailed below: (1) coaching provides accountability and support, (2) self-pacing offers value that meets individual needs, (3) participants like the comfort and convenience of the web-based format, and (4) technical issues were common but not insurmountable. Table 2 presents quotations exemplifying each theme.

Table 2. Domains and illustrative quotes.

Themes	Illustrative quotes
Theme 1: coaching provides accountability and support	• Because she was so good at bringing back the tools you get out of webSTAIR. Website tells you to apply it, but it's nice to hear it differently and she brings it back to my situation. That was really helpful for me (35-44–year-old White female). • Seemed like sessions were the practice/reiteration part. She really provided the application, tying it together and making it stick. Would stress this a lot. It really helped and challenged me. (35-44–year-old White male)
Theme 2: self-pacing offers value that meets individual needs	• It allowed me to go at my own pace. I had time to consider my answers and think about what I wanted to say. (55-64–year-old White male) • When you're meeting in person, you often are bobble heading even if you're not getting what they're telling you. Reading for yourself and being able to go over it and process… gives you time to process. (55-64–year-old mixed-race female) • Some lessons needed more time to practice and to mentally digest. I'm still going over them because I printed them out, but I feel like I needed more time... (45-54–year-old White female)
Theme 3: participants like the comfort and convenience of the web-based format	• I work, helped me not be stressed about getting appointments without taking off work. (25-34–year-old African American female) • With PTSD, I don't sleep well. To not have to get up and drive. Closest facility is 45 min-1 hour away…To not have to drive is great. I live in the country on purpose. I wouldn't go to therapy without webSTAIR. (35-44–year-old mixed-race female) • I liked that it was online and for whatever reason felt like a safer environment than sitting face to face with someone. (25-34–year-old White female) • It's easy to get complacent too, catch 22 at your home. Easy to get distracted, too comfortable at home so easily distracted. (25-34–year-old White male)
Theme 4: technical issues are common but not insurmountable	• If you're working on something and push the back button you delete everything you did. You have to start over. (35-44–year-old African American female) • I didn't like the fact the info I wrote down in the modules, therapist couldn't see, so when we were reviewing, I would have to go back in and start from the beginning. It would be helpful if she could see what I wrote. (35-44–year-old African American female) • Sometimes couldn't hear each other or it would freeze. Normal facetime issues. More issues with computer, iPad or iPhone crystal clear, perfect. (35-44–year-old American Indian or Alaskan Native female)

Theme 1: Coaching Provides Accountability and Support

The presence of regular check-ins with a coach kept participants accountable and motivated them to do the work and stay on track by providing guidance as to when they should complete the next module. Only 1 participant noted that the regular check-ins with their coach and expectations about content completion each week felt like too much pressure, but this was uncommonly reported. Most respondents indicated they appreciated check-ins.

> *I knew that we were going to have our sessions. So it helped with accountability – like actually doing the work because we're going to talk about it. And she keeps me on track because she tells me this is when you're going to do the next module.* [35-44-year-old Black or African-American female]

Coaches provided emotional support and helped participants understand and apply the webSTAIR content to their own lives. Participants often noted a good rapport with their coach, and many would have liked to talk to them more simply because they enjoyed talking with them. As a mixed-race male in his early 30s commented, "She was awesome. She always remembered what was going on with me. She helped with materials."

The majority of participants were satisfied with the number of sessions with their coach, regardless of the number of coaching sessions they received. Those that did express a preference for more sessions cited a variety of reasons, most commonly the desire for more emotional support and clarification about the content. Participants perceived coaching sessions as a time to process their feelings and better understand how content applied to their individual situation. The coaching sessions were considered an essential component of webSTAIR and were highly valued.

Theme 2: Self-Pacing Offers Value That Meets Individual Needs

Participants generally liked the format of webSTAIR, which allowed them to take their time and interact with and respond to the web-based content. Having time between reviewing the material and meeting with their coach helped participants feel less pressure and less rushed than standard face-to-face psychotherapy using evidence-based psychotherapy protocols, as in "No pressure to feel like you get an answer right or wrong or in a hurry to get it done" (35-44–year-old White male). They were able to reflect on and practice the skills webSTAIR aimed to teach them. In many cases, this led to more thoughtful reactions to the content, which informed meetings with their coach and made them more productive. A White male in his mid-forties commented:

> *It made you use your mind and think about stuff. You can go to counseling all day long and you don't get as much accomplished in an hour... [I] think on your own time, allows you to think outside the box.* [35-44–year-old White male]

In contrast to those who liked the self-paced format, a small number of participants reported they would have liked the pace to be more personalized or more self-paced. For example, a few participants would have liked to speed through some of the content that they felt they did not need or take more time (eg, more time in between video coaching sessions) to go deeper into the content they thought they needed most without feeling rushed. However, the desire for the pace to be more personalized did not impact their satisfaction with the program.

A minority of participants noted difficulty understanding or remembering the content between sessions.

> *By the time I got to the second module, I would forget about the first one. I would forget notes I had for my coach and was often unable to write down my questions. I would forget before next time.* [45-54-year-old White female]

Some participants successfully prepared for coaching sessions by taking detailed notes about the content and their questions, or by reviewing the webSTAIR content immediately before meeting with their coach. This strategy was acceptable for some webSTAIR participants, but others found it tedious or "annoying" to have to refresh between sessions to avoid forgetting the content, though this was uncommonly reported.

Theme 3: Comfort and Convenience of the Web-Based Format Made it Easier to Access Care

Participants liked the convenience of the coach-guided, web-based format, which allowed flexible scheduling and reduced travel time. The ability to meet on the web allowed veterans to engage in care they may not have otherwise engaged in due to distance or other logistical factors, such as difficulty scheduling around work or childcare obligations. A Hispanic male in his late 30s explained how and why the web-based format was more convenient:

> *Able to do it from home. I cannot take advantage of many mental health services at the VA; VA is 30-40 minutes from my home. I was glad I was able to take advantage of this.*

In addition to the logistical benefits, participants liked that they did not have to confront crowded waiting rooms at their facility. Instead, they could participate in the program from home, where they felt most comfortable and, in some cases, where they felt most safe. A minority of participants found it more difficult to find privacy in their own homes, away from the people they lived with, and others found themselves more prone to distractions at home. According to 1 participant, a White female in her mid-forties, "…the same thing that made it nice also made it not so nice." However, privacy issues or distractions at home were not commonly reported.

Theme 4: Technical Issues Are Common but not Insurmountable

The majority of participants experienced some technical challenges with the video coaching sessions, and almost half experienced difficulty with the website. Participants spontaneously reported that about a third of all the technical challenges they experienced occurred only 1 or 2 times throughout the entire program. For issues that did not resolve on their own, participants were able to troubleshoot, either independently or with their coach, to complete their sessions and web-based content. In cases where coaches were able to help solve problems or troubleshoot, their guidance seemed to ameliorate the negative impact on patient satisfaction.

> *[Had] difficulties quite a few times, but she always made it work.* [35–44-year-old Hispanic female]

Technical issues included being unable to hear the other person, issues with the link to the video session, and difficulty with the video freezing or closing unexpectedly. Sometimes the issue resolved quickly on its own and occurred infrequently. Other times, the issue could be attributed to a recent update or a setting that needed to be changed on their device. Strategies that were frequently used by coaches to help troubleshoot included pivoting to the phone for audio, sending different pieces of equipment to use, and offering emotional support that someone would help them resolve the issue. A White female in her mid-forties said:

> *One time we [experienced video issues]; [we] just turned off the audio, [and] video was still there and [we] used phone for audio. Work around; it worked just fine.*

Only 1 participant spontaneously reported having had to reschedule a session with their coach due to connection issues. More often, participants shared that they were still able to meet on the web with their coach despite issues with the video connection, in part because of their coach's persistence.

> *It didn't connect twice but she worked hard to get it connected, so we still had our sessions.* [35–44-year-old African American female]

Participants reported a wide range of issues with the website, including trouble accessing the tools or videos, glitches such as issues with the "back" or "next" button, difficulty with the website kicking them out or freezing up, and issues with the equipment that made accessing all the content from the modules difficult. Several participants spontaneously reported that they resolved the issue with the website through independent troubleshooting, usually by closing and reopening the browser: "Sometimes the role play [exercise] wasn't working right. Closed out the browser and then it would work fine" (35-44–year-old Hispanic female). In at least 1 case, an issue with the website was resolved when the coach sent the participant a VA iPad (Apple Inc). Issues that were most easily resolved included the website or web content freezing up and issues with the "back" button not working.

More persistent issues with the website included typed responses not saving, being unable to access program materials, such as worksheets or role plays, and difficulty printing. There was only

1 case in which a participant spontaneously shared that they had to reschedule a session due to technical issues with the website and an inability to access program materials. More often, participants found workarounds, like taking screenshots instead of printing and writing down their responses.

> *Typing info into forms/worksheets. Happened 2 times, then I quit. I started writing it down.* [45–54–year-old White female]

Discussion

Principal Results

This study contributes to the literature on satisfaction with web-based mental health treatments by qualitatively exploring mechanisms that lead to better or worse patient experiences with technology-based mental health tools, in addition to satisfaction with the patient-provider relationship. Results from this study may help clinicians and researchers tailor interventions in a way that increases satisfaction with web-based mental health treatments. Our chief findings were as follows: (1) the presence of a coach or therapist to help guide a web-based intervention was vital for accountability and support; (2) a flexible, self-paced format enabled participants to fit the treatment into their schedule more easily and give them time to reflect on the content in between sessions; (3) participants liked the comfort and convenience of the web-based format, as it decreased travel time and opened up access to mental health services that may otherwise have been too difficult to obtain; and, (4) despite most participants experiencing technical difficulties, a common issue with web-based treatments, participants in our sample were able to troubleshoot those issues and complete the program and it did not appear to impact their overall satisfaction. Although participants generally reported positive responses to the intervention, some reported a desire to have more control over the pacing of modules based on their individual mastery of the module's content.

Comparison With Previous Work

Consistent with the literature, the coach-guided support contributed to the high level of satisfaction observed in this study. Participants noted the importance of the coach for their support and accountability. Coaches helped keep participants on track, knowing that they are going to discuss the module in their next session. Coaching helped participants apply the content to their own lives and made it relatable, enhancing the program's value for patients. Consistent with previous work by this team, participants in this study often noted a high level of satisfaction with their coach, and many commented on the likeability or effectiveness of their coach [40]. Previous work suggests that engagement and efficacy in technology-based mental health interventions are often low without therapist support [41], which may be why the value of coaching was such a prevalent theme in our exploration of satisfaction in this study. Like participants in other coach-guided, internet-based studies, webSTAIR participants experienced excellent relationships with their coaches, and they were able to achieve this over video [30,42-44].

Previous studies of technology-based mental health interventions suggest that the most frequently reported reason for dissatisfaction is related to intervention pacing that moves too quickly [45]. Participants in this study often noted the benefit of the self-paced format of webSTAIR, as it allows more time and space to reflect on, practice, and absorb the content compared to standard, synchronous, evidence-based psychotherapy sessions in which participants may feel pressure to respond and process material in the moment. Additional time to review the material between professionally guided sessions contributed to the considerable satisfaction that we have observed in this study.

Participants in this study generally appreciated the comfort and convenience of the web-based format. Approximately 62% (46/74) of our participants were rural, so decreasing drive time to their provider, in some cases, may have enabled participants to receive care they would not have otherwise been able to access. About 57% (42/74) of our participants were women, and all had a history of trauma. This may have contributed to why participants felt more comfortable receiving their care in their own homes and why they appreciated avoiding crowded VHA waiting rooms. These findings echo previous studies that found similar benefits of telehealth for patients with obsessive-compulsive disorder and veterans receiving mental health care over video telehealth [29,46].

When using video telehealth, about a third of participants noted they only encountered 1-2 technical issues, which they were able to resolve with minimal interruption. Others with more significant interruptions were able to troubleshoot on their own or with their coach to eventually get connected or access the web-based material. Since this study describes the experiences of participants who completed webSTAIR, we can presume that they were able to navigate technical issues and complete the program. Given these findings, we can infer that if a technical issue is eventually resolved, the presence of the technical issue may not impact overall satisfaction. For this reason, resilience, supportive troubleshooting, and having a backup plan (eg, using the phone for sound when audio does not work) may contribute to greater satisfaction and the successful completion of internet-based interventions.

Limitations

The primary limitation of our study is selection bias. Our interview sample consisted of participants who completed the webSTAIR program and were prescreened to have adequate access to video telehealth. These factors may result in more positive experiences with the program, as reflected in the data. The 37% (74/202) response rate may have further exacerbated selection bias and resulted in more positive feedback. These patients were more likely to have the time and motivation to participate in an interview. The experiences of those who may not have felt the program's impact or did not feel inclined to participate in an interview were not captured. Additionally, the study was not randomized, and we do not have a control group to compare satisfaction among those who dropped out of the study. However, as a qualitative study, our aim was to provide a detailed description of participant experiences and explore how common themes across experiences may be relevant to clinicians and researchers interested in web-based mental health treatments.

Conclusions

Our findings suggest that a combination of self-paced, web-based content and a coach-guided intervention may integrate the best of both worlds, that is, the convenience and flexibility of a web-based modality and the structure, guidance, and accountability of a traditional psychotherapy approach. Our results indicate that providers and researchers should design tech-based interventions that are sensitive to individual differences and that integrate coaching support, as participants highly value coaching support for their digital interventions. Future research should focus on the impact of satisfaction on engagement, retention, and the therapeutic benefits of web-based mental health interventions.

Acknowledgments

The authors would like to acknowledge the South Central Mental Illness Research, Education, and Clinical Center (MIRECC) and the Center for Innovations in Quality, Effectiveness, and Safety (IQuESt, CIN13-413), which provided facility and resource support, as well as the Veterans Affairs (VA) Palo Alto National Center for PTSD, which assisted in the design, collection of data, and preparation and editing of the manuscript. The opinions expressed are those of the authors and do not necessarily reflect those of the Department of Veterans Affairs, the US government, or Baylor College of Medicine.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview questions.
[DOCX File , 14 KB - mental_v11i1e50977_app1.docx]

References

1. Office of research and development. U.S. Department of Veterans Affairs. URL: https://www.research.va.gov/topics/mental_health.cfm [accessed 2023-01-18]
2. Health Resources and Services Administration. Designated health professional shortage area statistics: first quarter of fiscal year 2023. Bureau of Health Workforce Health Resources and Services Administration (HRSA) US Department of Health and Human Services. URL: https://data.hrsa.gov/Default/GenerateHPSAQuarterlyReport [accessed 2023-12-23]
3. Elnitsky CA, Andresen EM, Clark ME, McGarity S, Hall CG, Kerns RD. Access to the US Department of Veterans Affairs Health System: self-reported barriers to care among returnees of operations enduring freedom and iraqi freedom. BMC Health Serv Res 2013;13:498 [FREE Full text] [doi: 10.1186/1472-6963-13-498] [Medline: 24289747]
4. Marshall V, Stryczek KC, Haverhals L, Young J, Au DH, Ho PM, et al. The focus they deserve: improving women veterans' health care access. Womens Health Issues 2021;31(4):399-407 [FREE Full text] [doi: 10.1016/j.whi.2020.12.011] [Medline: 33582001]
5. Godier-McBard LR, Wood A, Kohomange M, Cable G, Fossey M. Barriers and facilitators to mental healthcare for women veterans: a scoping review. J Ment Health 2023;32(5):951-961. [doi: 10.1080/09638237.2022.2118686] [Medline: 36062860]
6. Possemato K, Wray LO, Johnson E, Webster B, Beehler GP. Facilitators and barriers to seeking mental health care among primary care veterans with posttraumatic stress disorder. J Trauma Stress 2018;31(5):742-752. [doi: 10.1002/jts.22327] [Medline: 30338576]
7. Teich J, Ali MM, Lynch S, Mutter R. Utilization of mental health services by veterans living in rural areas. J Rural Health 2017;33(3):297-304. [doi: 10.1111/jrh.12221] [Medline: 27701791]
8. Douthit N, Kiv S, Dwolatzky T, Biswas S. Exposing some important barriers to health care access in the rural USA. Public Health 2015;129(6):611-620. [doi: 10.1016/j.puhe.2015.04.001] [Medline: 26025176]
9. Weeks WB, Kazis LE, Shen Y, Cong Z, Ren XS, Miller D, et al. Differences in health-related quality of life in rural and urban veterans. Am J Public Health 2004;94(10):1762-1767 [FREE Full text] [doi: 10.2105/ajph.94.10.1762] [Medline: 15451747]
10. Büscher R, Sander LB, Nuding M, Baumeister H, Teismann T. Blending video therapy and digital self-help for individuals with suicidal ideation: intervention design and a qualitative study within the development process. JMIR Form Res 2023;7:e49043 [FREE Full text] [doi: 10.2196/49043] [Medline: 37733414]
11. Wong HW, Lo B, Shi J, Hollenberg E, Abi-Jaoude A, Johnson A, et al. Postsecondary student engagement with a mental health app and online platform (thought spot): qualitative study of user experience. JMIR Ment Health 2021;8(4):e23447 [FREE Full text] [doi: 10.2196/23447] [Medline: 33797395]
12. Shalaby R, Vuong W, Eboreime E, Surood S, Greenshaw AJ, Agyapong VIO. Patients' expectations and experiences with a mental health-focused supportive text messaging program: mixed methods evaluation. JMIR Form Res 2022;6(1):e33438 [FREE Full text] [doi: 10.2196/33438] [Medline: 35014972]

13. Venville A, O'Connor S, Roeschlein H, Ennals P, McLoughlan G, Thomas N. Mental health service user and worker experiences of psychosocial support via telehealth through the COVID-19 pandemic: qualitative study. JMIR Ment Health 2021;8(8):e29671 [FREE Full text] [doi: 10.2196/29671] [Medline: 34182461]
14. Lange A, Rietdijk D, Hudcovicova M, van de Ven JP, Schrieken B, Emmelkamp PMG. Interapy: a controlled randomized trial of the standardized treatment of posttraumatic stress through the internet. J Consult Clin Psychol 2003;71(5):901-909. [doi: 10.1037/0022-006X.71.5.901] [Medline: 14516238]
15. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019;21(7):e12869 [FREE Full text] [doi: 10.2196/12869] [Medline: 31333198]
16. Griffiths KM, Christensen H. Review of randomised controlled trials of internet interventions for mental disorders and related conditions. Clinical Psychologist 2007;10(1):16-29. [doi: 10.1080/13284200500378696]
17. Andrews G, Basu A, Cuijpers P, Craske MG, McEvoy P, English CL, et al. Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: an updated meta-analysis. J Anxiety Disord 2018;55:70-78 [FREE Full text] [doi: 10.1016/j.janxdis.2018.01.001] [Medline: 29422409]
18. Chiolero A, Burnier M, Santschi V. Improving treatment satisfaction to increase adherence. J Hum Hypertens 2016;30(5):295-296. [doi: 10.1038/jhh.2015.89] [Medline: 26290276]
19. Wade SL, Cassedy AE, Sklut M, Taylor HG, McNally KA, Kirkwood MW, et al. The relationship of adolescent and parent preferences for treatment modality with satisfaction, attrition, adherence, and efficacy: the Coping with Head Injury Through Problem-Solving (CHIPS) Study. J Pediatr Psychol 2019;44(3):388-401 [FREE Full text] [doi: 10.1093/jpepsy/jsy087] [Medline: 30452665]
20. Ong JJ. Parental satisfaction and perception of progress in influencing the practice of complementary health approaches in children with autism: a cross sectional survey from Negeri Sembilan, Malaysia. BMC Complement Altern Med 2019;19(1):250 [FREE Full text] [doi: 10.1186/s12906-019-2672-8] [Medline: 31500611]
21. Peyrot M, Rubin RR. How does treatment satisfaction work?: modeling determinants of treatment satisfaction and preference. Diabetes Care 2009;32(8):1411-1417 [FREE Full text] [doi: 10.2337/dc08-2256] [Medline: 19470837]
22. Priebe S, Miglietta E. Assessment and determinants of patient satisfaction with mental health care. World Psychiatry 2019;18(1):30-31 [FREE Full text] [doi: 10.1002/wps.20586] [Medline: 30600613]
23. Swift JK, Callahan JL. The impact of client treatment preferences on outcome: a meta-analysis. J Clin Psychol 2009;65(4):368-381. [doi: 10.1002/jclp.20553] [Medline: 19226606]
24. Kim SC, Kim S, Boren D. The quality of therapeutic alliance between patient and provider predicts general satisfaction. Mil Med 2008;173(1):85-90 [FREE Full text] [doi: 10.7205/milmed.173.1.85] [Medline: 18251337]
25. Vuori H. Patient satisfaction—an attribute or indicator of the quality of care? QRB Qual Rev Bull 1987;13(3):106-108. [doi: 10.1016/s0097-5990(16)30116-6] [Medline: 3106873]
26. Johansson P, Oléni M, Fridlund B. Patient satisfaction with nursing care in the context of health care: a literature study. Scand J Caring Sci 2002;16(4):337-344. [doi: 10.1046/j.1471-6712.2002.00094.x] [Medline: 12445102]
27. Pauselli L, Galletti C, Verdolini N, Paolini E, Gallucci D, Balducci PM, et al. Predictors of client satisfaction with outpatient mental health clinic services in Italy and New York. Community Ment Health J 2018;54(5):562-570. [doi: 10.1007/s10597-017-0196-6] [Medline: 29147978]
28. Mazziotti R, Rutigliano G. Tele-mental health for reaching out to patients in a time of pandemic: provider survey and meta-analysis of patient satisfaction. JMIR Ment Health 2021;8(7):e26187 [FREE Full text] [doi: 10.2196/26187] [Medline: 34114956]
29. Fletcher TL, Boykin DM, Helm A, Dawson DB, Ecker AH, Freshour J, et al. A pilot open trial of video telehealth-delivered exposure and response prevention for obsessive-compulsive disorder in rural Veterans. Military Psychology 2021;34(1):83-90. [doi: 10.1080/08995605.2021.1970983]
30. Fletcher TL, Amspoker AB, Wassef M, Hogan JB, Helm A, Jackson C, et al. Increasing access to care for trauma-exposed rural veterans: a mixed methods outcome evaluation of a web-based skills training program with telehealth-delivered coaching. J Rural Health 2022;38(4):740-747. [doi: 10.1111/jrh.12628] [Medline: 34648188]
31. Cloitre M, Cohen LR, Ortigo KM, Jackson C, Koenen KC. Treating Survivors of Childhood Abuse and Interpersonal Trauma: STAIR Narrative Therapy. New York, NY, US: Guilford Publications; 2020.
32. Brooks E, Dailey N, Bair B, Shore J. Rural women veterans demographic report: defining VA users' health and health care access in rural areas. J Rural Health 2014;30(2):146-152. [doi: 10.1111/jrh.12037] [Medline: 24689540]
33. Prins A, Ouimette P, Kimerling R, Camerond RP, Hugelshofer DS, Shaw-Hegwer J, et al. The primary care PTSD screen (PC-PTSD): development and operating characteristics. Prim Care Psych 2004;9(1):9-14. [doi: 10.1185/135525703125002360]
34. Kroenke K, Spitzer RL, Williams JBW. The patient health questionnaire-2: validity of a two-item depression screener. Med Care 2003;41(11):1284-1292. [doi: 10.1097/01.MLR.0000093487.78664.3C] [Medline: 14583691]
35. Bauer A, Amspoker AB, Fletcher TL, Jackson C, Jacobs A, Hogan J, et al. A resource building virtual care programme: improving symptoms and social functioning among female and male rural veterans. Eur J Psychotraumatol 2021;12(1):1860357 [FREE Full text] [doi: 10.1080/20008198.2020.1860357] [Medline: 34025913]

36. Elo S, Kyngäs H. The qualitative content analysis process. J Adv Nurs 2008;62(1):107-115. [doi: 10.1111/j.1365-2648.2007.04569.x] [Medline: 18352969]
37. Lindgren BM, Lundman B, Graneheim UH. Abstraction and interpretation during the qualitative content analysis process. Int J Nurs Stud 2020;108:103632. [doi: 10.1016/j.ijnurstu.2020.103632] [Medline: 32505813]
38. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
39. Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: implications for conducting a qualitative descriptive study. Nurs Health Sci 2013;15(3):398-405 [FREE Full text] [doi: 10.1111/nhs.12048] [Medline: 23480423]
40. Roddy MK, Chen PV, Fletcher TL, Wassef M, Lindsay JA, Cloitre M, et al. Coaching with internet-based mental health program: patient and coach experiences. Telemed J E Health 2023;29(12):1853-1861. [doi: 10.1089/tmj.2023.0116] [Medline: 37219868]
41. Mohr DC, Cuijpers P, Lehman K. Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions. J Med Internet Res 2011;13(1):e30 [FREE Full text] [doi: 10.2196/jmir.1602] [Medline: 21393123]
42. Chen PV, Helm A, Caloudas SG, Ecker A, Day G, Hogan J, et al. Evidence of phone vs video-conferencing for mental health treatments: a review of the literature. Curr Psychiatry Rep 2022;24(10):529-539 [FREE Full text] [doi: 10.1007/s11920-022-01359-8] [Medline: 36053400]
43. Donaghy E, Atherton H, Hammersley V, McNeilly H, Bikker A, Robbins L, et al. Acceptability, benefits, and challenges of video consulting: a qualitative study in primary care. Br J Gen Pract 2019;69(686):e586-e594 [FREE Full text] [doi: 10.3399/bjgp19X704141] [Medline: 31160368]
44. Agha Z, Schapira RM, Laud PW, McNutt G, Roter DL. Patient satisfaction with physician-patient communication during telemedicine. Telemed J E Health 2009;15(9):830-839. [doi: 10.1089/tmj.2009.0030] [Medline: 19919189]
45. Olthuis JV, Watt MC, Bailey K, Hayden JA, Stewart SH. Therapist-supported internet cognitive behavioural therapy for anxiety disorders in adults. Cochrane Database Syst Rev 2015(3):CD011565 [FREE Full text] [doi: 10.1002/14651858.CD011565] [Medline: 25742186]
46. Lindsay JA, Day SC, Amspoker AB, Fletcher TL, Hogan J, Day G, et al. Personalized implementation of video telehealth. Psychiatr Clin North Am 2019;42(4):563-574. [doi: 10.1016/j.psc.2019.08.001] [Medline: 31672207]

Abbreviations

PC-PTSD: Primary Care PTSD Screen
PHQ-2: 2-item Patient Health Questionnaire
PTSD: posttraumatic stress disorder
STAIR: Skills Training in Affective and Interpersonal Regulation
VA: Veterans Affairs
VHA: Veterans Health Administration

Edited by J Torous; submitted 18.07.23; peer-reviewed by S Rennick-Egglestone, K Possemato; comments to author 06.09.23; revised version received 20.10.23; accepted 21.10.23; published 02.02.24.

Please cite as:
Smith AH, Touchett H, Chen P, Fletcher T, Arney J, Hogan J, Wassef M, Cloitre M, Lindsay JA
Patient Satisfaction With a Coach-Guided, Technology-Based Mental Health Treatment: Qualitative Interview Study and Theme Analysis
JMIR Ment Health 2024;11:e50977
URL: https://mental.jmir.org/2024/1/e50977
doi:10.2196/50977
PMID:38306167

Original Paper

Remote Short Sessions of Heart Rate Variability Biofeedback Monitored With Wearable Technology: Open-Label Prospective Feasibility Study

Robert P Hirten[1,2], MD; Matteo Danieletto[2,3], PhD; Kyle Landell[2,3], BA; Micol Zweig[2,3], MPH; Eddye Golden[3], MPH; Renata Pyzik[4], MS; Sparshdeep Kaur[2,3], BBA; Helena Chang[5], MS; Drew Helmus[1], MPH; Bruce E Sands[1], MD; Dennis Charney[6,7], MD; Girish Nadkarni[8], MD; Emilia Bagiella[5], PhD; Laurie Keefer[1], PhD; Zahi A Fayad[4,9], PhD

[1]The Dr Henry D Janowitz Division of Gastroenterology, Icahn School of Medicine at Mount Sinai, New York, NY, United States
[2]Windreich Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY, United States
[3]The Hasso Plattner Institute for Digital Health at the Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, NY, United States
[4]The BioMedical Engineering and Imaging Institute, Icahn School of Medicine at Mount Sinai, New York, NY, United States
[5]Center for Biostatistics, Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY, United States
[6]Office of the Dean, Icahn School of Medicine at Mount Sinai, New York, NY, United States
[7]Nash Family Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY, United States
[8]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, United States
[9]Department of Diagnostic, Molecular and Interventional Radiology, Icahn School of Medicine at Mount Sinai, New York, NY, United States

Corresponding Author:
Robert P Hirten, MD
The Dr Henry D Janowitz Division of Gastroenterology
Icahn School of Medicine at Mount Sinai
1 Gustave L Levy Place
New York, NY, 10029
United States
Phone: 1 212 241 6500
Email: robert.hirten@mountsinai.org

Abstract

Background: Heart rate variability (HRV) biofeedback is often performed with structured education, laboratory-based assessments, and practice sessions. It has been shown to improve psychological and physiological function across populations. However, a means to remotely use and monitor this approach would allow for wider use of this technique. Advancements in wearable and digital technology present an opportunity for the widespread application of this approach.

Objective: The primary aim of the study was to determine the feasibility of fully remote, self-administered short sessions of HRV-directed biofeedback in a diverse population of health care workers (HCWs). The secondary aim was to determine whether a fully remote, HRV-directed biofeedback intervention significantly alters longitudinal HRV over the intervention period, as monitored by wearable devices. The tertiary aim was to estimate the impact of this intervention on metrics of psychological well-being.

Methods: To determine whether remotely implemented short sessions of HRV biofeedback can improve autonomic metrics and psychological well-being, we enrolled HCWs across 7 hospitals in New York City in the United States. They downloaded our study app, watched brief educational videos about HRV biofeedback, and used a well-studied HRV biofeedback program remotely through their smartphone. HRV biofeedback sessions were used for 5 minutes per day for 5 weeks. HCWs were then followed for 12 weeks after the intervention period. Psychological measures were obtained over the study period, and they wore an Apple Watch for at least 7 weeks to monitor the circadian features of HRV.

Results: In total, 127 HCWs were enrolled in the study. Overall, only 21 (16.5%) were at least 50% compliant with the HRV biofeedback intervention, representing a small portion of the total sample. This demonstrates that this study design does not feasibly result in adequate rates of compliance with the intervention. Numerical improvement in psychological metrics was observed over the 17-week study period, although it did not reach statistical significance (all P>.05). Using a mixed effect cosinor model, the mean midline-estimating statistic of rhythm (MESOR) of the circadian pattern of the SD of the interbeat interval of normal sinus beats (SDNN), an HRV metric, was observed to increase over the first 4 weeks of the biofeedback intervention in HCWs who were at least 50% compliant.

Conclusions: In conclusion, we found that using brief remote HRV biofeedback sessions and monitoring its physiological effect using wearable devices, in the manner that the study was conducted, was not feasible. This is considering the low compliance rates with the study intervention. We found that remote short sessions of HRV biofeedback demonstrate potential promise in improving autonomic nervous function and warrant further study. Wearable devices can monitor the physiological effects of psychological interventions.

(JMIR Ment Health 2024;11:e55552) doi:10.2196/55552

KEYWORDS

biofeedback; digital health; digital technology; health care worker; HCW; heart rate variability; mHealth; mobile health; mobile phone; remote monitoring; smartphone; wearable devices

Introduction

Overview

Mental health conditions are common, with approximately 25% of the population in the United States experiencing a mental health disorder in a given year [1]. Since the COVID-19 pandemic, there have been increasing rates of anxiety, depression, and other psychological conditions [2]. This has disproportionately impacted health care workers (HCWs) who are at a higher risk of depression, anxiety, insomnia, and distress compared to the general population [3-6]. Over half of the physicians and approximately 40% of the nurses in the United States experience burnout, almost twice that of other professions [7]. Additionally, during the COVID-19 pandemic, approximately 1 in 5 HCWs were experiencing some degree of posttraumatic stress disorder [8]. Thus, HCWs represent a vulnerable population in which further study of mental health interventions is needed.

Unfortunately, access to mental health services can be limited [9]. Digital technologies, including smartphone apps and wearable devices, provide an opportunity to improve health care access and aid mental health professionals in the management of psychological conditions. Collectively, they can assess subjective and objective metrics of psychological and physiological well-being. Apps can remotely collect validated psychological assessments while wearable devices are able to monitor physiological metrics such as heart rate variability (HRV), a hypothesized indirect measure of the autonomic nervous system (ANS) [10-13]. HRV is a measure of the physiological variation in the time intervals between adjacent heartbeats [14]. It is hypothesized to be generated by heart-brain interactions and ANS processes, reflecting the activity of the sympathetic and parasympathetic nervous system tone on heart rate [10,15].

Higher HRV has been associated with reduced frustration, higher performance, and positive psychological adjustments [16]. Reduced HRV has been associated with reduced self-regulation, variable degrees of psychological tension, and anxiety [17,18]. Oscillations in heart rate occur due to the influence of respiration on the sinoatrial node of the heart and central nervous system respiratory pacemaker fluctuations. Interestingly, at 1 resting respiratory rate, the relationship between breathing and heart rate is asynchronous, with the heart rate increasing following inhalation [19,20]. This respiratory sinus arrhythmia is controlled by the vagus nerve, with increased vagal output producing greater heart rate variation, thereby reflecting the parasympathetic influence on the heart [21]. It has been shown that the amplitude of HRV is related to breathing frequency with maximum effect at a breathing rate of 0.1 Hz or 6 breaths per minute [19].

Mind-body interventions, such as deep breathing exercises, can improve resilience, psychological well-being, physiological functions, autonomic imbalance, mood, cardiopulmonary output, and immune function [22-25]. Adaptive changes in the central nervous system, characterized as reduced sympathetic tone, have been described with these exercises [26-28]. Achieving deep breathing rates of 4.5-6.5 breaths per minute results in higher HRV indices compared to baseline, with higher parasympathetic and baroreflex function [29]. This has been shown to positively impact physical function; athletic performance; quality of life; and psychological features such as anxiety, depression, and resilience [30-32]. The individual breath per minute rate producing the optimal HRV effect (resonance frequency) can be determined from measures of the heart and respiratory rate in real-time biofeedback sessions [33]. Changes in HRV secondary to respiratory rate modification can create a positive feedback loop further increasing HRV respiratory changes, elicited through biofeedback [19].

Biofeedback is a self-regulatory behavioral method that trains individuals to control physiological function through real-time information about these physical parameters [34]. HRV biofeedback involves the real-time visualization of HRV metrics and breathing's effect on this metric. It has been shown to increase HRV in adults [31,35-37]. There is significant empirical support for the use of office- or laboratory-based HRV biofeedback programs for the improvement of psychological conditions. In a recent meta-analysis of 14 studies, HRV biofeedback was shown to improve depressive symptoms in several psychophysiological conditions, as well as increase psychological well-being [38]. Large reductions in self-reported stress and anxiety have been demonstrated with HRV-directed biofeedback [31], as well as positive impacts on anger, athletic performance, sleep, and quality of life [39]. A systematic review of HRV biofeedback further demonstrated significantly improved symptoms of anxiety, depression, panic disorders, and posttraumatic stress disorder in 70% of the included studies [40].

However, despite the effectiveness of HRV-directed biofeedback, there are limitations to the implementation of such a technique. These interventions often rely on structured training and computer- or laboratory-based practice sessions that are often performed in the laboratory setting. This makes it challenging to broadly implement such techniques, limiting access to populations that may be most likely to benefit. This has prompted several studies using HRV biofeedback remotely and outside the laboratory setting with computer-based programs that demonstrated effect [36,41,42]. An additional significant obstacle to HRV biofeedback is the length of time required for each session, which can last up to 40 minutes [43]. Most also incorporate at least 1 laboratory session per week in addition to the daily home sessions [44]. These long and structured sessions, however, limit the ability of individuals to institute an HRV biofeedback program into their daily lives. Short sessions of HRV biofeedback might therefore provide a greater impact if they are able to elicit an autonomic response. Interestingly, short sessions of HRV biofeedback can successfully modify HRV and improve the regulation of emotional reactivity and therefore warrant further evaluation [45,46]. Gross et al [47] used 5 short 3- to 5-minute HRV biofeedback sessions. However, these were led by in-person practitioners. They demonstrated that HRV was successfully moderated and increased during these sessions; however, it was not changed overall from before training to after training. Deschodt-Arsac et al [45] furthered the evaluation of short-session HRV biofeedback by evaluating a twice-daily 5-minute biofeedback session in athletes, demonstrating an increase in autonomic function and a decrease in anxiety levels.

HRV measurements during and after biofeedback sessions evaluating physiological effects are often over brief periods and are in the clinic or laboratory setting. This limits the evaluation of its effectiveness on an individual's physiological status and further restricts biofeedback sessions to the office setting. Wearable devices provide a potential means to assess HRV remotely, passively, and outside the laboratory setting and, thus, a possible means to monitor HRV biofeedback in a real-world setting. Wearable-based HRV assessment can be performed through either electrocardiography (ECG) or photoplethysmography (PPG). ECG is the gold standard for HRV assessment as the graphical representation of cardiac activity enables the calculation of beat-to-beat intervals with reliability to the millisecond level [48]. Most commercially available wearables and all wrist- or hand-worn devices that measure HRV rely on PPG technology. PPG tracks heartbeats by measuring the alterations of light from an LED that reaches a photodiode created by pressure changes in veins with each heartbeat [49]. Several studies have used wearable devices to assess response to HRV-directed biofeedback sessions. However, these have primarily used wearables that both monitor and implement biofeedback at the same time. Chung et al [50] demonstrated in a small pilot study that the Lief Smart Patch can assess and deliver HRV-directed biofeedback to effectively modify HRV. However, HRV assessments, generated from an ECG tracing, were over very brief periods around the biofeedback sessions. Similarly, Lin et al [51] demonstrated that using an HRV biofeedback wearable device for a least 4 weeks was needed to demonstrate an effect on HRV. However, studies that have evaluated sensor-type preference in biofeedback have found that participants prefer wrist- or arm-worn sensors for monitoring [52]. Given the ubiquitous use of commercial smartwatches, many of which measure HRV, there is an opportunity to expand HRV-directed biofeedback monitoring with such devices. Commercial devices such as the Apple Watch [12,53,54], fPolar V800 [13,55], Empatica E4 wristband [56], and Fitbit Charge HR [57] have been shown to generate valid and reliable assessments of heart rate and HRV, with high agreement with ECGs. Furthermore, the use of HRV calculated through PPG signatures has been shown to be a reliable and valid method for the assessment of HRV in the setting of HRV-directed biofeedback [58].

Thus, the potential benefits of short sessions of HRV biofeedback coupled with the growth of digital technologies and wearable devices present an opportunity to expand the application and monitoring of HRV-directed biofeedback. To evaluate this approach, we launched a feasibility study to evaluate smartphone-based short sessions of HRV biofeedback in HCWs and monitored its impact using common commercially available wearable devices.

Objectives

The primary aim of the study was to determine the feasibility of fully remote, self-administered short sessions of HRV-directed biofeedback in a diverse population of HCWs. We hypothesized that fully remote HRV-directed biofeedback would have high compliance rates by HCWs. The secondary aim was to determine whether a fully remote, HRV-directed biofeedback intervention significantly alters longitudinal HRV over the intervention period. We hypothesized that HRV-directed biofeedback would significantly alter longitudinal HRV measurements. The tertiary aim was to estimate the impact of this intervention on metrics of psychological well-being. It was hypothesized that psychological well-being would improve with HRV-directed biofeedback. Study feasibility will be assessed by the percentage of HCWs who are at least 50% compliant with the intervention over the study period.

Methods

Ethical Considerations

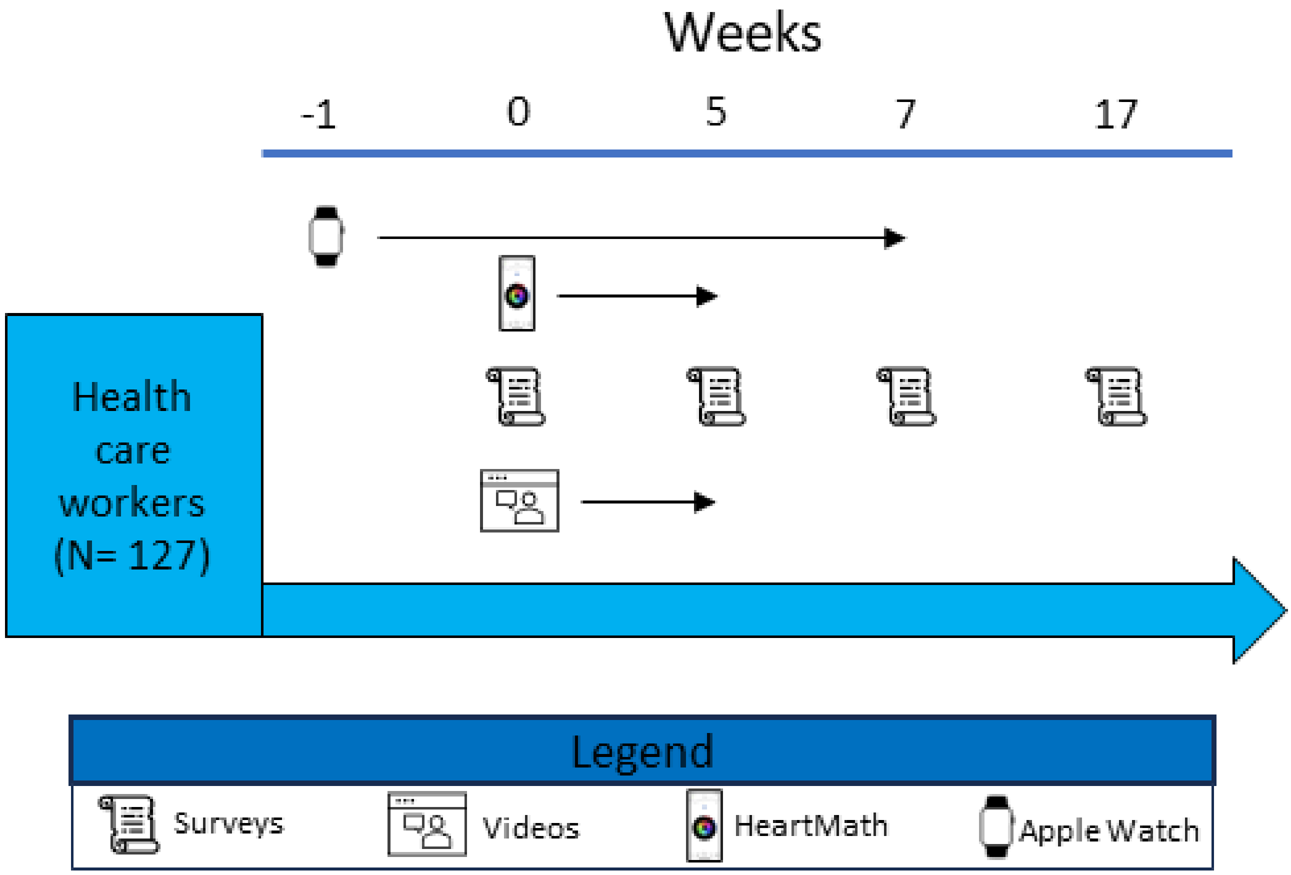
This study has been approved by the institutional review board at the Icahn School of Medicine at Mount Sinai (STUDY-21-00596). The study was retrospectively registered on ClinicalTrials.gov (NCT05958329). All participants signed informed consent. All study procedures were performed in accordance with the ethical standards outlined in the Helsinki Declaration of 1975, as revised in 2000. The study data was deidentified, with each participants data linked to a unique study identification number. Additionally, all data was stored on Mount Sinai's HIPAA compliant servers.

Study Design

The Warrior Shield study was an open-label prospective pilot clinical trial that enrolled HCWs across 7 hospitals in New York City (Figure 1). Participants were recruited from The Mount Sinai Hospital, Morningside Hospital, Mount Sinai West, Mount Sinai Beth Israel, Mount Sinai Queens, New York Eye and Ear

Infirmary, and Mount Sinai Brooklyn. Eligible participants were aged 18 years or older, employees at 1 of the participating sites, had an iPhone series 6 or higher, and had or were willing to wear an Apple Watch 5 or greater. Potential participants were excluded if they had an underlying chronic disease or used a medication that is known to impact ANS function.

Figure 1. Participants were prospectively enrolled and followed for 17 weeks. Participants wore their Apple Watch for at least 7 days before starting the intervention period (week 0) and used it through week 7 of the study. The HeartMath device was used throughout the 5-week intervention period. Participants answered surveys at baseline, week 5, week 7, and week 17. In total, 5 weekly educational videos describing HeartMath and the basis behind the intervention were available for viewing through the 5-week intervention period.



HCWs were recruited from the participating hospitals through emails sent to hospital employees and through study flyers (Figure S1 in Multimedia Appendix 1) placed in hospital common areas, including cafeterias and lobbies. Furthermore, participants who completed other digital studies run by our group were messaged with information about this study. Participants were provided with a US $50 gift card after completing 6 weeks of study activities. If a participant did not have an Apple Watch, he or she was able to borrow 1 for the duration of the study. This was returned to the research staff on completion of the study. Additionally, the participants had to return the HeartMath Inner Balance biofeedback device (HeartMath, LLC) at the end of the study.

Study Procedures

Overview

Participants downloaded our ehive study app to their smartphones and self-verified inclusion and exclusion criteria before signing the electronic consent. Participants then electronically requested their HeartMath Inner Balance biofeedback device and an Apple Watch if they did not have 1 of their own wearable devices. Participants were recommended to wear the Apple Watch for a minimum of 8 hours per day. On receiving the Inner Balance device and after at least 7 days of wearing the Apple Watch, participants started the HRV biofeedback intervention, as described below. This continued for 5 weeks. Validated surveys to assess psychological well-being were completed at baseline in the ehive app. Surveys were repeated at week 5, week 7, and week 17. Participants were asked to wear the Apple Watch for at least 7 weeks after starting the intervention to enable HRV monitoring. They were reminded to participate in the study through regularly scheduled push notifications to their smartphones and automated email reminders sent by the study team.

ehive App

The ehive app is the centralized digital research platform of The Hasso Plattner Institute for Digital Health at Mount Sinai Hospital, New York, New York. The patient-facing portion of the platform is a smartphone app that enables electronic consenting of participants. Customizable patient-reported outcomes measures and other tasks such as the study surveys and weekly study videos are embedded in the app and can be tracked for compliance. This app can track participant compliance and engage participants through light touch measures such as customized push notifications and customized emails to participants to maintain engagement. It has been downloaded by over 1484 participants and has been used to collect over 51 million wearable-based data points and over 132,241 surveys [11].

Survey Instruments

Several validated surveys were evaluated throughout the intervention period. The 10-item Connor-Davidson Resilience Scale (CD-RISC 10) is a 10-question survey that measures resilience. Higher scores reflect higher resilience, with each question graded on a 5-point Likert scale [59]. The following question is an example of what is included in the survey: “I am able to adapt when changes occur.” The emotional support 2-item Patient-Reported Outcomes Measurement Information System (PROMIS) questionnaire is graded from 2 to 10 points, with higher scores reflecting higher perceived emotional support. It measures whether an individual has someone who will listen to them and with whom they can discuss their feelings [60]. An example of a question included in this survey is “I have someone who will listen to me when I need to talk.” The Perceived Stress Scale 10 (PSS-10) is a validated survey assessing perceived stress. It is 10 questions scored from 0 to 40, with higher scores correlating with elevated perceived stress [61]. An example item in this survey includes “In the last month, how often have you been upset because of something that happened unexpectedly?” The 2-Item Global Health and Quality of Life Scale asks participants to grade how their quality of life and health are in general. Higher scores correlate with lower health and quality of life [60]. The following is an example question included in this survey: “In general, would you say your health is excellent, very good, good, fair, or poor?” The Patient Health Questionnaire-4 (PHQ-4) is a 4-question survey that screens for anxiety and depression and is graded from 0 to 12 points. Higher scores reflect more severe impairment [62]. An example question from this survey is “Over the last week, how often have you been bothered by the following problem? Feeling nervous, anxious, or on edge.” The National Institutes of Health (NIH) PROMIS positive affect and well-being scale is a 23-question survey graded on a 5-point Likert scale. Higher scores reflect higher degrees of positive affect and well-being [63]. The following question is an example from this survey: “Lately I had a sense of well-being never, rarely, sometimes, often, or always?”

HeartMath Intervention

The HeartMath biofeedback system is developed by the HeartMath Institute, which is a nonprofit research and educational organization that develops and provides easy-to-use self-regulation tools focused on HRV biofeedback [64]. Its tools and techniques have been tested in a range of settings with good efficacy and uptake in conditions ranging from blood pressure, heart failure, stress, and trauma syndromes [65-68]. It is used in a range of settings and has been widely implemented in the health care industry, being offered to HCWs and patients at institutions such as Kaiser Permanente and the Veteran Administration Hospitals and Clinics [69,70]. The Inner Balance app combines a smartphone app (Figure S2 in Multimedia Appendix 1) with an optical ear sensor enabling real-time HRV visualization, assessment, and optimization during biofeedback sessions. Participants downloaded the Inner Balance app to their smartphones and set up an account using the login information provided by the study team. HeartMath’s Inner Balance pulse sensor clips on the participant’s ear and links through Bluetooth directly to an individual’s smartphone. The sensor contains an optical photodetector that samples up to 125 Hz providing real-time HRV assessment. Clip-on ear sensors have been shown to provide an accurate assessment of HRV compared to ECG [71]. Through HRV calculations, it produces an index of coherence, as a percentage of time in high, medium, or low coherence, through breathing and self-generated positive emotions [72]. A flower-shaped central visual pacer is present in the app, which paces a participant’s breathing. Through integration with sensed HRV, the app is able to reinforce the correct technique for HRV optimization [65,73,74].

Participants used the Inner Balance app for one 5-minute session per day for 5 weeks. Compliance was tracked remotely through the HeartMath system. HRV biofeedback sessions are usually supplemented with in-person or structured education sessions. To enable learning remotely, weekly educational videos were provided to participants in the custom ehive app. Five weekly videos provided information on (1) how to use the technology; (2) an introduction to HRV, biofeedback, and coherence; (3) a description of what coherence is and how it works; (4) how to incorporate biofeedback techniques into everyday life; and (5) reinforcement of what is learned in prior videos. Each video was less than 20 minutes in length and could be watched over the week.

Wearable Device

HRV was measured by the Apple Watch Series 5 or 6 that was worn by participants throughout the intervention and postintervention period. The Apple Watch contains a PPG optical sensor with both a green light diode and a light-sensitive photodiode [75]. This creates time series peaks that are filtered for ectopic beats and used to generate interbeat intervals. HRV was automatically calculated by the Apple Watch using the SD of the interbeat interval of normal sinus beats (SDNN) [76]. SDNN is a time-domain HRV metric that reflects both sympathetic and parasympathetic nervous system activity [10]. The only HRV metric available from the Apple Watch is SDNN. Multiple HRV measurements were generated by the Apple Watch throughout each 24-hour period in which individuals were wearing the device. These data were retrieved through our ehive app. The Apple watch calculates each of these SDNN measurements over 60-second windows, with a bias toward nighttime measurements, to minimize artifacts in the readings. The algorithms used by the Apple Watch for artifact rejection and ectopic beat handling are proprietary and not publicly available. However, they likely use well-described algorithms in this space [64]. While this is a limitation, the PPG-based HRV calculations from Apple Watches have been validated against ECG [12,77].

Statistical Analysis

Data are presented descriptively as the mean and SD or frequency and percentages, as appropriate. Mean values for each psychological assessment were obtained at baseline; just before initiation of HRV biofeedback; and at week 5, week 7, and week 17. Changes over time in the psychological assessment were analyzed using mixed effects models with participants as random effects. Week-5, week-7, and week-17 survey results were each compared for statistical differences to the baseline values.

HRV is captured by the Apple Watch in a relatively sparse and nonuniform sampling and follows a circadian pattern [78,79]. To account for frequent daily measures of HRV that are collected from wearable devices over a several-week period, statistical methods that take into account these changes are needed. Daily circadian rhythms have been previously modeled by nonlinear cosinor methods [80]. This approach models the circadian HRV rhythm each day over 24 hours and enables the data to be described using circadian parameters (Figure S3 in Multimedia Appendix 1): (1) midline-estimating statistic of rhythm (MESOR): the midline of the rhythm, or a rhythm-adjusted mean, over the 24 hours; (2) acrophase: a measure of the time of the highest values that reoccur each day; and (3) amplitude: characterizes half the extent of the variation in every 24 hours. To fully use the cyclical nature of the physiological metrics, as well as the longitudinal measurements, mixed-effect cosinor models were used to model HRV over time based on the *cosinoRmixedeffects* R package (R Core Team) [81]. This expands the nonlinear cosinor methods to account for correlations of repeated measurements within a participant over time. As has been previously described, a cosinor model uses the nonlinear function $Y(t)=M + A\cos(2\pi t/\tau + \varphi) + e_i(t)$, where τ is the period (τ=24 h), M is the MESOR, A is the amplitude, and φ is the acrophase. This can be converted into the linear model with $x=\sin(2\pi t/\tau)$, $z=\sin(2\pi t/\tau)$. HRV can be written as $Y(t)=M + \beta x_t + \gamma z_t + e_i(t)$ [82]. The mixed-effect cosinor model extends the linear framework in a longitudinal setting through the inclusion of random effects θ_i that models the within-patient correlation and expressed as $Y_{it} = (M + \beta x_{it} + \gamma z_{it}) + W_{it}\theta_i + e_i(t)$, where $\theta_i \sim MVN(0, \Sigma)$ [81]. Bootstrapping procedures were used to calculate the confidence intervals of the model estimates. Age, sex, and BMI were included as covariates in the HRV analyses with participants as random effects.

HRV was evaluated using the above approach for each 7 days of the study. The baseline measurement reflects the 7 days preceding the initiation of the HRV-directed biofeedback. Each subsequent 7-day period, over the 7-week HRV observation period, was compared to this baseline value. All analyses were carried out at the 2-sided .05 significance level using SAS (version 9.4; SAS Institute Inc) and R (version 4.2.2; Foundation for Statistical Computing). Since this was a proof-of-concept study, there was no adjustment for the multiplicity of hypothesis testing.

Results

Quantitative Findings

In total, 127 participants consented to the study between July 2021 and April 2022. The mean age of these participants was 37.3 (SD 10.6) years, with 93 (73.8%) being female. In total, 72 (56.7%) of these participants started the intervention and used the Inner Balance device at least 1 time (>0% compliance), while 49 (38.6%) participants were at least 20% compliant, and 21 (16.5%) participants were at least 50% adherent over the 5-week intervention period (Table 1).

Table 1. Demographic information for participants signing consent, those who used the intervention at least 1 time, those with at least 20% compliance, and those with at least 50% compliance.

Characteristics	Signed consent (N=127)	Compliance >0% (n=72)	Compliance ≥20% (n=49)	Compliance ≥50% (n=21)
Age (years), mean (SD)	37.3 (10.6)	38.4 (11.0)	38.0 (11.0)	37.7 (12.1)
Male, n (%)	33 (26.2)	22 (31)	13 (26.5)	5 (23.8)
BMI (kg/m^2), mean (SD)	25.3 (5.5)	25.3 (5.5)	25.3 (5.6)	27.3 (6.5)
Race, n (%)				
Asian	32 (25.6)	15 (21.4)	11 (22.9)	2 (10)
Black	15 (12)	9 (12.9)	6 (12.5)	3 (15)
White	69 (55.2)	43 (61.4)	29 (60.4)	14 (70)
Native Hawaiian or Pacific Islander	2 (1.6)	1(1.4)	1 (2.1)	1 (5)
Unknown	4 (3.2)	2 (2.9)	1 (2.1)	1 (5)
Hispanic or Latino, n (%)	29 (23.8)	16 (23.5)	11 (23.4)	6 (30)
Smoking: never or rarely, n (%)	103 (81.1)	57 (79.2)	40 (81.6)	17 (81)
Anxiety, n (%)	29 (24.4)	16 (23.5)	10 (20.4)	6 (28.6)
Depression, n (%)	25 (19.8)	15 (20.8)	10 (20.4)	6 (28.6)

The percentage of participants who watched the entire weekly video decreased throughout the study. A video introducing the study at enrollment was watched in its entirety by 100% of participants. The video in week 2 was watched in its entirety by 54% (68/127) of participants, the week 3 video was watched by 47% (60/127) of participants, the week 4 video was watched by 42% (53/127) of participants, and the week 5 video was watched by 39% (49/127) of participants. There was a technology tutorial video that provided information about the Inner Balance system. This was watched in its entirety by 65% (83/127) of participants.

Overall, the acceptability of the study was good. Participants were asked how satisfied they were with the HeartMath Intervention on a scale of 1 (not satisfied) to 7 (very satisfied). In total, 79 participants answered the question with a median score of 5. Out of the 81 participants who answered the question as to whether they pursued additional learning about HeartMath outside of the study, 17.3% (n=14) reported in the affirmative. Participants who pursued additional learning about HeartMath were more satisfied with the HeartMath intervention, scoring their degree of satisfaction with a mean of 6.07 (SD 0.86) compared to those who did not pursue outside learning (mean 4.45, SD 1.52).

Psychological Assessment

In participants who were at least 50% compliant (n=21) with the Inner Balance device, resilience scores were noted to numerically increase between the baseline assessment, week 5, week 7, and week 17. However, none of these values differed significantly from the baseline assessment. Social support scores (emotional support 2-item PROMIS questionnaire) similarly demonstrated a numerical increase from baseline (mean 8.13, SD 1.46) to week 5 (mean 8.60, SD 0.89), week 7 (mean 9.80, SD 0.45), and week 17 (mean 8.67, SD 2.31). None of these increases were statistically significant compared to the baseline assessment. Stress scores (PSS-10) numerically decreased in the ≥50% compliant cohort, dropping from mean 20.63 (SD 5.95) at baseline to mean 10.67 (SD 7.77) at week 17. The change in stress scores at week 5 (P=.24), week 7 (P=.45), and week 17 (P=.26) were not significantly different compared to the baseline assessment. PHQ-4 scores, which increase when there is greater psychological impairment, decreased from baseline through week 17. In the ≥50% compliant cohort there was not a statistically significant change in these values, compared to the baseline assessment, at week 5 (P=.83), week 7 (P=.55), or week 17 (P=.38). NIH PROMIS positive affect and well-being scores rose as well over the 17 weeks in the ≥50% compliant cohort, reflecting increasing positive affect and well-being. Due to the small number of individuals in this cohort, we were not able to calculate P values for this comparison. The 2-Item Global Health and Quality of Life Scale increased over the observation period, demonstrating higher quality of life. This change, compared to baseline, did not reach the level of statistical significance at week 5 (P=.50), week 7 (P=.62), or week 17 (P=.36). Overall, psychological assessments additionally demonstrated numerical improvement over the 5-week intervention period and through the 17-week follow-up period in those who used the Inner Balance device at least once and in those who were ≥20% compliant, though they did not differ significantly from the baseline assessments (Table 2).

Table 2. The mean psychological assessments are presented at baseline, week 5, week 7, and week 17 in each compliance group. The mean scores for each survey at week 5, week 7, and week 17 are compared against the baseline scores. *P* values reflect the significance of this comparison. Compliance groups are defined as those performing the intervention at least 1 time, those with at least 20% compliance, and those with at least 50% compliance.

Measures	Compliance >0%		Compliance ≥20%		Compliance ≥50%	
	Mean (SD)	*P* value	Mean (SD)	*P* value	Mean (SD)	*P* value
CD-RISC 10[a]						
Baseline	27.05 (7.20)	—[b]	28.04 (7.80)	—	27.38 (5.66)	—
Week 5	27.67 (6.80)	.15	27.31 (5.92)	.79	23.80 (7.66)	.68
Week 7	27.94 (9.16)	.05	28.22 (9.00)	.12	30.40 (6.95)	.44
Week 17	32.50 (3.42)	.07	31.33 (3.06)	.37	29.00 (4.58)	.95
Emotional support 2-item PROMIS[c] questionnaire						
Baseline	8.45 (1.75)	—	8.48 (1.86)	—	8.13 (1.46)	—
Week 5	8.72 (1.13)	.41	8.77 (1.09)	.95	8.60 (0.89)	.65
Week 7	9.00 (2.0)	.12	9.00 (2.00)	.27	9.80 (0.45)	.27
Week 17	9.5 (1.0)	.70	9.33 (1.15)	.43	8.67 (2.31)	.66
PSS-10[d]						
Baseline	16.53 (6.22)	—	17.26 (6.31)	—	20.63 (5.95)	—
Week 5	18.61 (6.41)	.22	16.31 (5.34)	.64	17.00 (4.64)	.24
Week 7	18.82 (6.59)	.20	18.22 (6.24)	.53	16.80 (7.85)	.45
Week 17	12.63 (7.91)	.85	10.17 (7.59)	.23	10.67 (7.77)	.26
PHQ-4[e]						
Baseline	3.13 (2.53)	—	3.61 (2.59)	—	4.25 (2.92)	—
Week 5	2.61 (2.38)	.27	2.69 (2.53)	.19	3.60 (3.78)	.83
Week 7	2.71 (2.39)	.97	2.44 (1.01)	.40	2.60 (1.14)	.55
Week 17	0.88 (0.63)	.12	0.83 (0.76)	.08	1.33 (1.53)	.38
NIH[f] PROMIS positive affect and well-being scale						
Baseline	82.65 (18.08)	—	84.91 (18.00)	—	81.50 (10.88)	—
Week 5	86.16 (13.14)	.23	85.21 (11.89)	.99	83.80 (15.59)	N/A[g]
Week 7	88.99 (16.55)	.19	90.33 (20.37)	.32	93.60 (19.48)	N/A
Week 17	97.50 (10.66)	.10	97.33 (13.05)	.18	90.00 (4.24)	N/A
2-Item Global Health and Quality of Life Scale						
Baseline	7.58 (1.54)	—	7.43 (1.44)	—	7.63 (1.41)	—
Week 5	7.61 (1.72)	.78	7.31 (1.84)	.60	8.60 (1.34)	.50
Week 7	7.41 (0.80)	.70	7.33 (1.00)	.84	7.20 (0.84)	.62
Week 17	7.88 (1.44)	.63	8.50 (0.87)	.19	8.67 (0.58)	.36

[a]CD-RISC 10: Connor-Davidson Resilience Scale.

[b]Not available.

[c]PROMIS: Patient-Reported Outcomes Measurement Information System.

[d]PSS-10: Perceived Stress Scale 10.

[e]PHQ-4: Patient Health Questionnaire-4.

[f]NIH: National Institutes of Health.

[g]N/A: not applicable.

Physiological Metrics

There was an average of 4.7 (SD 3.5) HRV measurements obtained per participant per day. The average length of time of each sample was 59 seconds. The median SDNN value obtained in the full cohort was 38 milliseconds with a minimum and maximum value of 10 milliseconds and 200 milliseconds, respectively. We fit a cosinor model evaluating differences in HRV (SDNN) each week over the 5-week intervention period and over the 2 weeks following the intervention period. There were no significant changes from baseline in the amplitude or acrophase of the circadian pattern of SDNN in all 3 compliance groups (Table 3). Significant changes were noted in the MESOR of the circadian pattern of SDNN in participants who are ≥50% compliant with the intervention. In this group, the mean MESOR was 50.20 (95% CI 41.16-58.78) during the baseline 7-day period. A numerical but not significant rise (*P*=.12) in the MESOR was observed during week 1 of the intervention (mean 52.59; 95% CI 43.65-61.08). There were significant changes in the mean MESOR of the circadian pattern of SDNN found during week 2 (mean 55.00; 95% CI 46.11-63.37; difference 4.80; 95% CI 1.63-7.91; *P*<.001), week 3 (mean 54.25; 95% CI 45.27-62.76; difference 4.04; 95% CI 0.64-7.00; *P*=.01), and week 4 (mean 55.70; 95% CI 46.77-63.94; difference 5.50; 95% CI 2.31-8.60; *P*<.001) compared to baseline (Figure 2). The MESOR during week 5 of the intervention and the 2 weeks after the end of the intervention did not demonstrate significant changes compared to baseline.

In the participants who used the Inner Balance device at least once and in those who were ≥20% compliant with the intervention, there was only 1 significant change in the MESOR observed over the 7-week follow-up period. There was a significant change in the MESOR of the circadian HRV pattern in participants with >0% compliance with the intervention during week 1 (mean 45.46; 95% CI 39.30-51.58; difference 1.48; 95% CI 0.10-2.88; *P*=.04), compared to the baseline 7-day period.

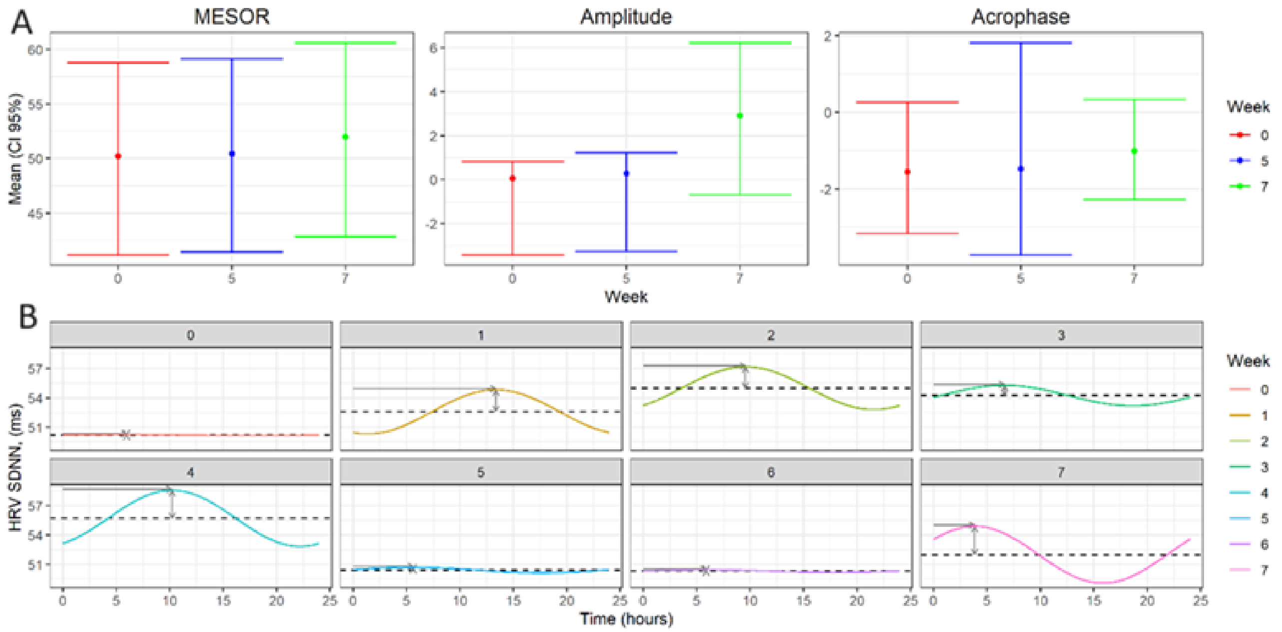
Table 3. The mean midline-estimating statistic of rhythm (MESOR), amplitude, and acrophase are presented for each week of the observation period. heart rate variability (HRV) circadian parameters were calculated for each 7 days of the study, with the baseline readings representing the 7-day preintervention period. Comparisons between each HRV metric 7-day period and the baseline 7-day period were performed. P-values reflect the significance of each comparison. Compliance groups are defined as those performing the intervention at least 1 time, those with at least 20% compliance, and those with at least 50% compliance.

Parameters	Compliance >0%			Compliance ≥20%			Compliance ≥50%		
	Mean (95% CI)	Difference (95% CI)	*P* values	Mean (95% CI)	Difference (95% CI)	*P* values	Mean (95% CI)	Difference (95% CI)	*P* values
MESOR									
Baseline	43.98 (37.84 to 49.93)	—[a]	—	45.71 (38.60 to 52.93)	—	—	50.20 (41.16 to 58.78)	—	—
Week 1	45.46 (39.30 to 51.58)	1.48 (0.10 to 2.88)	.04	46.83 (39.65 to 54.11)	1.12 (–0.53 to 2.97)	.19	52.59 (43.65 to 61.08)	2.38 (–0.84 to 5.14)	.12
Week 2	45.30 (39.46 to 51.70)	1.62 (0.13 to 3.31)	.05	46.93 (39.75 to 54.20)	1.22 (–0.68 to 2.88)	.18	55.00 (46.11 to 63.37)	4.80 (1.63 to 7.91)	<.001
Week 3	44.83 (38.80 to 50.85)	0.85 (–0.60 to 2.50)	.26	47.19 (40.19 to 54.36)	1.48 (–0.35 to 3.30)	.10	54.25 (45.27 to 62.76)	4.04 (0.64 to 7.00)	.01
Week 4	45.28 (39.08 to 51.42)	1.30 (–0.17 to 2.74)	.10	47.24 (40.03 to 54.51)	1.53 (–0.33 to 3.17)	.10	55.70 (46.77 to 63.94)	5.50 (2.31 to 8.60)	<.001
Week 5	45.31 (39.22 to 51.43)	1.33 (–0.14 to 2.95)	.09	47.25 (40.08 to 54.52)	1.54 (–0.17 to 3.46)	.09	50.43 (41.43 to 59.14)	0.22 (–3.47 to 3.69)	.89
Week 6	45.36 (39.23 to 51.42)	1.38 (–0.13 to 2.88)	.08	47.20 (40.07 to 54.36)	1.49 (–0.22 to 3.26)	.11	50.33 (41.41 to 58.84)	0.12 (–2.82 to 2.95)	.93
Week 7	45.19 (39.17 to 51.22)	1.22 (–0.44 to 2.75)	.14	46.66 (39.53 to 53.82)	0.95 (–0.87 to 2.66)	.31	52.00 (42.84 to 60.60)	1.79 (–1.46 to 4.90)	.27
Amplitude									
Baseline	4.30 (2.61 to 5.94)	—	—	3.55 (1.57 to 5.61)	—	—	0.06 (–3.43 to 0.82)	—	—
Week 1	4.37 (2.63 to 6.15)	0.07 (–2.08 to 2.10)	.94	4.45 (2.49 to 6.33)	0.90 (–1.23 to 3.12)	.43	2.25 (–1.17 to 5.22)	2.20 (–2.52 to 4.98)	.25
Week 2	4.83 (3.12 to 6.58)	0.53 (–1.76 to 2.58)	.63	4.75 (2.88 to 6.62)	1.20 (–1.06 to 3.62)	.28	2.18 (– 0.99 to 5.03)	2.13 (–2.63 to 5.20)	.27
Week 3	4.45 (2.59 to 6.18)	0.15 (–2.02 to 2.34)	.88	3.75 (1.80 to 5.71)	0.20 (–2.35 to 2.58)	.88	1.05 (– 2.20 to 2.87)	0.99 (–2.87 to 3.70)	.55
Week 4	4.49 (2.69 to 6.23)	0.19 (–2.18 to 2.01)	.87	4.73 (2.63 to 6.79)	1.18 (–1.31 to 3.99)	.34	2.87 (– 0.62 to 6.16)	2.81 (–2.23 to 5.65)	.17
Week 5	3.91 (2.03 to 5.67)	–0.39 (– 2.56 to 1.75)	.71	3.67 (1.55 to 5.68)	0.12 (–2.01 to 2.66)	.92	0.29 (– 3.27 to 1.23)	0.23 (–2.98 to 3.32)	.89
Week 6	4.28 (2.56 to 5.96)	–0.02 (– 2.26 to 1.95)	.98	4.08 (2.16 to 5.95)	0.53 (–2.00 to 3.12)	.64	0.12 (– 3.54 to 1.06)	0.06 (–3.04 to 3.45)	.96
Week 7	2.84 (0.99 to 4.60)	–1.46 (– 3.50 to 0.69)	.16	2.10 (0.08 to 4.16)	–1.45 (– 3.73 to 0.91)	.21	2.91 (– 0.69 to 6.23)	2.85 (–2.21 to 5.61)	.15
Acrophase									
Baseline	–2.93 (–3.35 to –2.53)	—	—	–3.03 (–3.65 to –2.33)	—	—	–1.56 (–3.16 to 0.27)	—	—
Week 1	–2.79 (–3.19 to –2.40)	0.14 (–0.44 to 0.61)	.59	–3.25 (–3.71 to –2.81)	–0.22 (–0.93 to 0.49)	.49	–3.50 (–5.35 to –1.70)	–1.94 (–4.31 to 1.18)	.11
Week 2	–2.69 (–3.05 to –2.37)	0.24 (–0.24 to 0.75)	.29	–3.23 (–3.69 to 2.79)	–0.19 (–0.82 to 0.42)	.53	–2.50 (–4.17 to –0.50)	–0.95 (–3.32 to 1.24)	.26
Week 3	–2.56 (–2.88 to –2.24)	0.37 (–0.07 to 0.93)	.15	–2.47 (–2.89 to –2.05)	0.56 (–0.04 to 1.30)	.10	–1.74 (–2.33 to –1.08)	–0.18 (–1.35 to 2.03)	.70
Week 4	–3.32 (–3.72 to –2.92)	–0.39 (–0.96 to 0.14)	.16	–3.37 (–3.86 to –2.91)	–0.34 (–1.07 to 0.39)	.32	–2.67 (–4.10 to –1.01)	–1.12 (–3.35 to 1.33)	.20

Parameters	Compliance >0%			Compliance ≥20%			Compliance ≥50%		
	Mean (95% CI)	Difference (95% CI)	*P* values	Mean (95% CI)	Difference (95% CI)	*P* values	Mean (95% CI)	Difference (95% CI)	*P* values
Week 5	–3.16 (–3.61 to –2.72)	–0.23 (–0.84 to 0.33)	.41	–3.16 (–3.81 to –2.54)	–0.13 (–0.91 to 0.66)	.70	–1.47 (–3.72 to 1.81)	0.09 (–4.60 to 2.00)	.90
Week 6	–2.78 (–3.19 to –2.37)	0.15 (–0.27 to 0.69)	.54	–3.03 (–3.56 to –2.50)	0.0004 (–0.79 to 0.69)	>.99	–1.54 (–3.26 to 0.56)	0.02 (–3.26 to 1.55)	.98
Week 7	–2.89 (–3.50 to –2.20)	0.04 (–0.67 to 0.68)	.91	–2.87 (–4.04 to –1.67)	0.16 (–0.93 to 1.33)	.73	–1.01 (–2.27 to 0.34)	0.55 (–1.39 to 2.77)	.32

[a]Not available.

Figure 2. (A) Plots show the mean (95% CIs) heart rate variability (HRV) midline-estimating statistic of rhythm, amplitude, and acrophase for participants at baseline, week 5, and week 7. (B) Plots show the average weekly circadian HRV rhythm for participants at baseline and over the first 7 weeks of the study period for participants with at least 50% compliance (n=21). midline-estimating statistic of rhythm; SDNN: SD of the interbeat interval of normal sinus beats.



Discussion

Overview

Overall, we found that using brief remote HRV biofeedback sessions and monitoring its physiological effect using wearable devices, in the manner that the study was conducted, was not feasible. This is considering the low compliance rates with the study intervention. However, there was a numerical improvement in all psychological metrics over the intervention period, and compliant participants had a measurable physiological change in wearable assessed HRV. In addition, participants were in general satisfied with the HeartMath system that was used. This supports the potential for at-home, HRV-directed biofeedback and wearable-based monitoring to be effective, but only when participants are engaged. The findings highlight the challenges with maintaining engagement in large remote intervention studies.

This study built on the existing literature supporting the use of short sessions of HRV biofeedback by using a short 5-minute HRV biofeedback session that could be performed on an individual's smartphone. Furthermore, it took the structured education that often accompanies biofeedback and divided it into easily digestible short videos, which individuals could absorb at their own pace. This framework pilots an approach that enables the intervention to be used by individuals who might not have the time to engage in a more structured program. Furthermore, while the physiological effects of biofeedback are often evaluated through brief HRV assessments, we used a commonly used commercial wearable device to monitor its impact. While HRV data were available from the HeartMath device during the short biofeedback sessions, this represented only a very brief assessment of the physiological effect in a relatively small number of compliant participants (n=21). These measurements do not assess the intervention's sustained effect on an individual's physiological parameters, which is of primary interest in this study. Therefore, our focus was on analyzing and leveraging the longitudinal HRV data provided by the Apple Watch. The benefit of this approach is 2-fold in that it can unobtrusively monitor the intervention's effect and evaluate the intervention's impact over longer periods through its assessment of circadian features of autonomic function. Importantly, we demonstrated that the MESOR of the circadian pattern, which reflects the mean HRV reading over the observation period, increased in participants compliant with the intervention, reflecting increased parasympathetic tone. Previous studies have demonstrated that commercially available wearable devices may be able to monitor and identify psychological states through HRV monitoring [82,83]. The results of this study extend these observations by demonstrating that commonly used wearable devices can potentially be used to monitor the physiological effects of psychological interventions and warrant further evaluation.

We used the HeartMath system, which uses a well-studied HRV biofeedback tool, as described above. We found that psychological metrics were numerically improved with the intervention. However, these changes did not meet statistical significance. A primary driver of this observation is likely the low rate of adherence, as the number of people who were at least 50% compliant with the intervention was only 21 individuals. While the trend in improvement was evident in all 3 adherence groups, statistical significance may have been met if the number of participants ≥50% compliant was larger or if rates of adherence were well over 50%. Given the limited number of participants, we were not able to perform a sensitivity analysis to determine the minimum adherence or engagement rate needed to elicit an effect. However, the trends we observed in psychological metrics warrant further study of this approach. Another potential hypothesis as to why we did not see statistical improvement in the psychological metrics may be that the cohort is relatively healthy compared to other groups undergoing psychological interventions. However, when we look at psychological metrics such as resilience, we see that the mean CD-RISC 10 score for the entire cohort was 27.05, compared to the general population's mean of 31.8 (SD 5.4) [84]. Therefore, the cohort is less resilient at baseline and presumably would benefit from such an intervention. Interestingly, we did demonstrate that short sessions of HRV biofeedback are able to significantly modify HRV. The performance of just half or more of the 5-minute biofeedback sessions in 1 week significantly impacted the circadian features of HRV and increased parasympathetic tone. While we did not observe this significant difference during the fifth week of the intervention in this cohort, the sample size was small, and engagement varied week by week, likely explaining the drop in effectiveness in the final week. During the 2 weeks after the intervention period, HRV was not significantly different from baseline. This observation demonstrates that sustained employment of short sessions of biofeedback is required for ongoing physiological effects. This is an important finding, as there are scarce studies evaluating the long-term impact of HRV biofeedback on HRV metrics, with few studies demonstrating sustained short-term effects [85,86]. Further studies evaluating the duration of physiological effects are needed.

While 127 participants initially joined the study, only 21 (16.5%) participants used the intervention at least half of the time. While low rates of persistent engagement can be seen across remote digital psychological intervention studies, future work using this remote biofeedback intervention should focus on direct means to maintain engagement. This could include coaching models or community-based engagement such as "leaderboards." [87] Additionally, dedicated study coordinators checking in with each participant could potentially improve adherence and participant engagement. Adherence may be increased by focusing recruitment efforts on individuals most interested in biofeedback programs. Our recruitment methods opened the study up to any HCW across multiple hospitals. However, focusing recruitment efforts on individuals interacting with hospital psychological support systems would engage individuals more likely to be interested in performing psychological interventions. Furthermore, we could hypothesize that the most engaged participants may have some degree of knowledge or interest in digital technologies, given the employment of apps and wearable devices. Therefore, such programs may be most effectively deployed in tech-savvy populations. While 79 participants rated the acceptability of the Heart Math system and were overall satisfied with it, we, unfortunately, did not have qualitative data regarding its acceptability or feasibility.

There are several additional limitations to this study. One important limitation is the limited external evaluation of Apple Watch–generated HRV measurements. There have been several studies evaluating and validating the Apple Watch's accuracy in measuring HRV. These studies have compared calculations derived from metrics collected from the Apple Watch with ECG measures. Turki et al [54] demonstrated that in 6 healthy participants, HRV acquired from R-R interval estimates derived from Apple Watch measures of heart rate are reasonable estimates of HRV derived from an ECG. Similarly, Hernando et al [12] validated the R-R intervals derived from the Apple Watch and the HRV metrics calculated from these series against readings derived from a single lead ECG acquired from the Polar band in 20 participants. Khushhal et al [53] performed a similar study on 21 individuals during exercise, demonstrating agreement in HRV metrics calculated from Apple Watch outputs compared to the Polar HR monitor. These studies demonstrate that Apple Watch metrics, used in the calculation of HRV by the device, are valid. However, the algorithms describing how the Apple Watch cleans the PPG data for ectopic beats and artifacts are not publicly available, and therefore limited data demonstrating how Apple calculated HRV metrics compared to ECG-derived measures. While this is a limitation of the study, we still incorporated the Apple Watch as it also serves the important purpose of demonstrating the potential for commonly used commercial devices to monitor the effect of HRV-directed biofeedback. HRV was only assessed in 1-time domain metric (SDNN) in this study, limiting the evaluation of other HRV metrics with the study outcomes. However, SDNN is 1 of the most common HRV features evaluated when studying resilience or the impact of HRV on psychological or physiological features [19]. Further study evaluating other HRV metrics is needed in the future to determine how other HRV parameters are impacted by short sessions of remote HRV-biofeedback. Another limitation is that we did not have exit surveys to understand why certain individuals were not compliant with study components, such as watching weekly videos or using the HeartMath device consistently. A final limitation is that HRV is not specific and can be impacted by many environmental factors beyond the covariates we controlled in the analysis. This is an important limitation to recognize as there is the potential for unmeasured covariates to impact the results, including such things as ongoing tobacco use and menstrual cycles.

Conclusion

We demonstrated that fully remote, short HRV biofeedback sessions, using light touch engagement measures, have low compliance rates. However, we did find numerical improvement in psychological assessments over the intervention and follow-up period and alterations of wearable assessed HRV measures in compliant individuals. This supports the need for further evaluation of remotely used short sessions of HRV biofeedback and of the use of wearable devices to monitor response if higher rates of engagement can be achieved.

Acknowledgments

RPH is the guarantor of this study. Support for this study was provided by the Ehrenkranz Lab for Human Resilience, the BioMedical Engineering and Imaging Institute, The Hasso Plattner Institute for Digital Health at Mount Sinai, The Mount Sinai Clinical Intelligence Center, The Dr. Henry D. Janowitz Division of Gastroenterology and K23DK129835 (RPH).

Data Availability

The data sets generated or analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

RPH is an advisor to Bristol Meyers Squibb. LK is a consultant/advisor to Pfizer, AbbVie, Ardelyx, Coprata Health, Trellus Health. LK owns equity in Trellus Health and receives royalties from MetaMe Health.

BES is a consultant or received speaker's fees from AbbVie, Abivax, Adiso Therapeutics, Alimentiv, Amgen, Arena Pharmaceuticals, Artizan Biosciences, Artugen Therapeutics, AstraZeneca, Bacainn Therapeutics, Biora Therapeutics, Boehringer Ingelheim, Boston Pharmaceuticals, Bristol Myers Squibb, Calibr, Celltrion, ClostraBio, Connect Biopharm, Cytoki Pharma, Eli Lilly and Company, Enthera, Evommune, Ferring, Fresenius Kabi, Galapagos, Gilead Sciences, Genentech, Glaxo SmithKline, Gossamer Bio, HMP Acquisition, Imhotex, Immunic, InDex Pharmaceuticals, Innovation Pharmaceuticals, Inotrem, Ironwood Pharmaceuticals, Janssen, Johnson & Johnson, Kaleido, Kalyope, Merck, MiroBio, Morphic Therapeutic, MRM Health, OSE Immunotherapeutics, Pfizer, Progenity, Prometheus Biosciences, Prometheus Laboratories, Protagonist Therapeutics, Q32 Bio, RedHill Biopharma, Sun Pharma Global, Surrozen, Synlogic Operating Company, Takeda, Target RWE, Theravance Biopharma R&D, TLL Pharmaceutical, USWM Enterprises, Ventyx Biosciences, Viela Bio, and stock options from Ventyx Biosciences.

DC is a coinventor on patents filed by the Icahn School of Medicine at Mount Sinai (ISMMS) relating to the treatment for treatment-resistant depression, suicidal ideation, and other disorders. ISMMS has entered into a licensing agreement with Janssen Pharmaceuticals, Inc, and it has received and will receive payments from Janssen under the license agreement related to these patents for the treatment of treatment-resistant depression and suicidal ideation. Consistent with the ISMMS Faculty Handbook (the medical school policy), DC is entitled to a portion of the payments received by the ISMMS. Because SPRAVATO has received regulatory approval for treatment-resistant depression, through the ISMMS, DC will be entitled to additional payments beyond those already received under the license agreement. DC is a named coinventor on several patents filed by ISMMS for a cognitive training intervention to treat depression and related psychiatric disorders. The ISMMS has entered into a licensing agreement with Click Therapeutics, Inc and has received and will receive payments related to the use of this cognitive training intervention for the treatment of psychiatric disorders. In accordance with the ISMMS Faculty Handbook, DC has received a portion of these payments and is entitled to a portion of any additional payments that the medical school may receive from this license with Click Therapeutics. DC is a named coinventor on a patent application filed by the ISMMS for the use of intranasally administered Neuropeptide Y for the treatment of mood and anxiety disorders. This intellectual property has not been licensed. DC is a named coinventor on a patent application in the United States and several issued patents outside the United States filed by the ISMMS related to the use of ketamine for the treatment of posttraumatic stress disorder. This intellectual property has not been licensed.

GN reports employment with, consultancy agreements with, and ownership interest in Pensieve Health and Renalytix AI; receiving consulting fees from AstraZeneca, BioVie, GLG Consulting, and Reata; and serving as a scientific advisor or member of Pensieve Health and Renalytix AI. ZAF discloses consulting fees from Rockley Photonics and Matter Neuroscience. ZAF receives financial compensation as a board member and advisor to Trained Therapeutix Discovery and owns equity in Trained Therapeutix Discovery as a cofounder. All other authors declare no relevant conflicts of interest.

Multimedia Appendix 1

Study recruitment flyer, screenshot of HeartMath's Inner Balance smartphone app, and daily circadian pattern of hear rate variability measures.

[DOCX File , 390 KB - mental_v11i1e55552_app1.docx]

References

1. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the national comorbidity survey replication. Arch Gen Psychiatry 2005;62(6):593-602 [FREE Full text] [doi: 10.1001/archpsyc.62.6.593] [Medline: 15939837]
2. Policy brief: COVID-19 and the need for action on mental health. United Nations Sustainable Development Group. 2020. URL: https://unsdg.un.org/resources/policy-brief-covid-19-and-need-action-mental-health [accessed 2024-03-28]
3. Lai J, Ma S, Wang Y, Cai Z, Hu J, Wei N, et al. Factors associated with mental health outcomes among health care workers exposed to coronavirus disease 2019. JAMA Netw Open 2020;3(3):e203976 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.3976] [Medline: 32202646]
4. Liang Y, Wu K, Zhou Y, Huang X, Zhou Y, Liu Z. Mental health in frontline medical workers during the 2019 novel coronavirus disease epidemic in China: a comparison with the general population. Int J Environ Res Public Health 2020;17(18):6550 [FREE Full text] [doi: 10.3390/ijerph17186550] [Medline: 32916836]
5. Rossi R, Socci V, Pacitti F, Di Lorenzo G, Di Marco A, Siracusano A, et al. Mental health outcomes among frontline and second-line health care workers during the coronavirus disease 2019 (COVID-19) pandemic in Italy. JAMA Netw Open 2020;3(5):e2010185 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.10185] [Medline: 32463467]
6. Saragih ID, Tonapa SI, Saragih IS, Advani S, Batubara SO, Suarilah I, et al. Global prevalence of mental health problems among healthcare workers during the Covid-19 pandemic: a systematic review and meta-analysis. Int J Nurs Stud 2021;121:104002 [FREE Full text] [doi: 10.1016/j.ijnurstu.2021.104002] [Medline: 34271460]
7. Reith TP. Burnout in United States healthcare professionals: a narrative review. Cureus 2018;10(12):e3681 [FREE Full text] [doi: 10.7759/cureus.3681] [Medline: 30761233]
8. Li Y, Scherer N, Felix L, Kuper H. Prevalence of depression, anxiety and post-traumatic stress disorder in health care workers during the COVID-19 pandemic: a systematic review and meta-analysis. PLoS One 2021;16(3):e0246454 [FREE Full text] [doi: 10.1371/journal.pone.0246454] [Medline: 33690641]
9. Gunnell D, Appleby L, Arensman E, Hawton K, John A, Kapur N, et al. Suicide risk and prevention during the COVID-19 pandemic. Lancet Psychiatry 2020;7(6):468-471 [FREE Full text] [doi: 10.1016/S2215-0366(20)30171-1] [Medline: 32330430]
10. Shaffer F, Ginsberg JP. An overview of heart rate variability metrics and norms. Front Public Health 2017;5:258 [FREE Full text] [doi: 10.3389/fpubh.2017.00258] [Medline: 29034226]
11. Hirten RP, Danieletto M, Landell K, Zweig M, Golden E, Orlov G, et al. Development of the ehive digital health app: protocol for a centralized research platform. JMIR Res Protoc 2023;12:e49204 [FREE Full text] [doi: 10.2196/49204] [Medline: 37971801]
12. Hernando D, Roca S, Sancho J, Alesanco Á, Bailón R. Validation of the Apple watch for heart rate variability measurements during relax and mental stress in healthy subjects. Sensors (Basel) 2018;18(8):2619 [FREE Full text] [doi: 10.3390/s18082619] [Medline: 30103376]
13. Caminal P, Sola F, Gomis P, Guasch E, Perera A, Soriano N, et al. Validity of the Polar V800 monitor for measuring heart rate variability in mountain running route conditions. Eur J Appl Physiol 2018;118(3):669-677 [FREE Full text] [doi: 10.1007/s00421-018-3808-0] [Medline: 29356949]
14. McCraty R, Shaffer F. Heart rate variability: new perspectives on physiological mechanisms, assessment of self-regulatory capacity, and health risk. Glob Adv Health Med 2015;4(1):46-61 [FREE Full text] [doi: 10.7453/gahmj.2014.073] [Medline: 25694852]
15. Singh N, Moneghetti KJ, Christle JW, Hadley D, Plews D, Froelicher V. Heart rate variability: an old metric with new meaning in the era of using mHealth technologies for health and exercise training guidance. Part one: physiology and methods. Arrhythm Electrophysiol Rev 2018;7(3):193-198 [FREE Full text] [doi: 10.15420/aer.2018.27.2] [Medline: 30416733]
16. Aritzeta A, Aranberri-Ruiz A, Soroa G, Mindeguia R, Olarza A. Emotional self-regulation in primary education: a heart rate-variability biofeedback intervention programme. Int J Environ Res Public Health 2022;19(9):5475 [FREE Full text] [doi: 10.3390/ijerph19095475] [Medline: 35564869]
17. Chalmers JA, Quintana DS, Abbott MJA, Kemp AH. Anxiety disorders are associated with reduced heart rate variability: a meta-analysis. Front Psychiatry 2014;5:80 [FREE Full text] [doi: 10.3389/fpsyt.2014.00080] [Medline: 25071612]
18. Hildebrandt LK, McCall C, Engen HG, Singer T. Cognitive flexibility, heart rate variability, and resilience predict fine-grained regulation of arousal during prolonged threat. Psychophysiology 2016;53(6):880-890 [FREE Full text] [doi: 10.1111/psyp.12632] [Medline: 26899260]
19. Lehrer PM, Gevirtz R. Heart rate variability biofeedback: how and why does it work? Front Psychol 2014;5:756 [FREE Full text] [doi: 10.3389/fpsyg.2014.00756] [Medline: 25101026]
20. Vaschillo E, Lehrer P, Rishe N, Konstantinov M. Heart rate variability biofeedback as a method for assessing baroreflex function: a preliminary study of resonance in the cardiovascular system. Appl Psychophysiol Biofeedback 2002;27(1):1-27. [doi: 10.1023/a:1014587304314] [Medline: 12001882]
21. Berntson GG, Bigger JT, Eckberg DL, Grossman P, Kaufmann PG, Malik M, et al. Heart rate variability: origins, methods, and interpretive caveats. Psychophysiology 1997;34(6):623-648. [doi: 10.1111/j.1469-8986.1997.tb02140.x] [Medline: 9401419]

22. Brown RP, Gerbarg PL. Sudarshan Kriya yogic breathing in the treatment of stress, anxiety, and depression: part I-neurophysiologic model. J Altern Complement Med 2005;11(1):189-201. [doi: 10.1089/acm.2005.11.189] [Medline: 15750381]
23. Jerath R, Edry JW, Barnes VA, Jerath V. Physiology of long pranayamic breathing: neural respiratory elements may provide a mechanism that explains how slow deep breathing shifts the autonomic nervous system. Med Hypotheses 2006;67(3):566-571 [FREE Full text] [doi: 10.1016/j.mehy.2006.02.042] [Medline: 16624497]
24. Kaushik RM, Kaushik R, Mahajan SK, Rajesh V. Effects of mental relaxation and slow breathing in essential hypertension. Complement Ther Med 2006;14(2):120-126. [doi: 10.1016/j.ctim.2005.11.007] [Medline: 16765850]
25. Berrill JW, Sadlier M, Hood K, Green JT. Mindfulness-based therapy for inflammatory bowel disease patients with functional abdominal symptoms or high perceived stress levels. J Crohns Colitis 2014;8(9):945-955 [FREE Full text] [doi: 10.1016/j.crohns.2014.01.018] [Medline: 24529603]
26. Bernardi L, Spadacini G, Bellwon J, Hajric R, Roskamm H, Frey AW. Effect of breathing rate on oxygen saturation and exercise performance in chronic heart failure. Lancet 1998;351(9112):1308-1311. [doi: 10.1016/S0140-6736(97)10341-5] [Medline: 9643792]
27. Harinath K, Malhotra AS, Pal K, Prasad R, Kumar R, Kain TC, et al. Effects of Hatha yoga and Omkar meditation on cardiorespiratory performance, psychologic profile, and melatonin secretion. J Altern Complement Med 2004;10(2):261-268. [doi: 10.1089/107555304323062257] [Medline: 15165407]
28. Han JN, Stegen K, De Valck C, Clément J, Van de Woestijne KP. Influence of breathing therapy on complaints, anxiety and breathing pattern in patients with hyperventilation syndrome and anxiety disorders. J Psychosom Res 1996;41(5):481-493. [doi: 10.1016/s0022-3999(96)00220-6] [Medline: 9032712]
29. Lin IM, Tai LY, Fan SY. Breathing at a rate of 5.5 breaths per minute with equal inhalation-to-exhalation ratio increases heart rate variability. Int J Psychophysiol 2014;91(3):206-211. [doi: 10.1016/j.ijpsycho.2013.12.006] [Medline: 24380741]
30. Paul M, Garg K. The effect of heart rate variability biofeedback on performance psychology of basketball players. Appl Psychophysiol Biofeedback 2012;37(2):131-144. [doi: 10.1007/s10484-012-9185-2] [Medline: 22402913]
31. Goessl VC, Curtiss JE, Hofmann SG. The effect of heart rate variability biofeedback training on stress and anxiety: a meta-analysis. Psychol Med 2017;47(15):2578-2586. [doi: 10.1017/S0033291717001003] [Medline: 28478782]
32. Giardino ND, Chan L, Borson S. Combined heart rate variability and pulse oximetry biofeedback for chronic obstructive pulmonary disease: preliminary findings. Appl Psychophysiol Biofeedback 2004;29(2):121-133. [doi: 10.1023/b:apbi.0000026638.64386.89] [Medline: 15208975]
33. Lehrer P, Vaschillo B, Zucker T, Graves J, Katsamanis M, Aviles M, et al. Protocol for heart rate variability biofeedback training. Biofeedback 2013;41(3):98-109 [FREE Full text] [doi: 10.5298/1081-5937-41.3.08]
34. Schwartz NM, Schwartz MS. Definitions of biofeedback and applied psychophysiology. In: Schwartz MS, Andrasik F, editors. Biofeedback: A Practitioner's Guide, Third Edition. New York: Guilford Press; 2003:27-39.
35. Aritzeta A, Soroa G, Balluerka N, Muela A, Gorostiaga A, Aliri J. Reducing anxiety and improving academic performance through a biofeedback relaxation training program. Appl Psychophysiol Biofeedback 2017;42(3):193-202. [doi: 10.1007/s10484-017-9367-z] [Medline: 28623467]
36. Henriques G, Keffer S, Abrahamson C, Horst SJ. Exploring the effectiveness of a computer-based heart rate variability biofeedback program in reducing anxiety in college students. Appl Psychophysiol Biofeedback 2011;36(2):101-112. [doi: 10.1007/s10484-011-9151-4] [Medline: 21533678]
37. Prato CA, Yucha CB. Biofeedback-assisted relaxation training to decrease test anxiety in nursing students. Nurs Educ Perspect 2013;34(2):76-81. [doi: 10.5480/1536-5026-34.2.76] [Medline: 23763019]
38. Pizzoli SFM, Marzorati C, Gatti D, Monzani D, Mazzocco K, Pravettoni G. A meta-analysis on heart rate variability biofeedback and depressive symptoms. Sci Rep 2021;11(1):6650 [FREE Full text] [doi: 10.1038/s41598-021-86149-7] [Medline: 33758260]
39. Lehrer P, Kaur K, Sharma A, Shah K, Huseby R, Bhavsar J, et al. Heart rate variability biofeedback improves emotional and physical health and performance: a systematic review and meta analysis. Appl Psychophysiol Biofeedback 2020;45(3):109-129. [doi: 10.1007/s10484-020-09466-z] [Medline: 32385728]
40. Schoenberg PLA, David AS. Biofeedback for psychiatric disorders: a systematic review. Appl Psychophysiol Biofeedback 2014;39(2):109-135. [doi: 10.1007/s10484-014-9246-9] [Medline: 24806535]
41. Jazaieri H, Goldin PR, Werner K, Ziv M, Gross JJ. A randomized trial of MBSR versus aerobic exercise for social anxiety disorder. J Clin Psychol 2012;68(7):715-731 [FREE Full text] [doi: 10.1002/jclp.21863] [Medline: 22623316]
42. Cavanagh K, Strauss C, Cicconi F, Griffiths N, Wyper A, Jones F. A randomised controlled trial of a brief online mindfulness-based intervention. Behav Res Ther 2013;51(9):573-578. [doi: 10.1016/j.brat.2013.06.003] [Medline: 23872699]
43. Lehrer PM, Vaschillo E, Vaschillo B. Resonant frequency biofeedback training to increase cardiac variability: rationale and manual for training. Appl Psychophysiol Biofeedback 2000;25(3):177-191. [doi: 10.1023/a:1009554825745] [Medline: 10999236]
44. Lalanza JF, Lorente S, Bullich R, García C, Losilla JM, Capdevila L. Methods for heart rate variability biofeedback (HRVB): a systematic review and guidelines. Appl Psychophysiol Biofeedback 2023;48(3):275-297 [FREE Full text] [doi: 10.1007/s10484-023-09582-6] [Medline: 36917418]

45. Deschodt-Arsac V, Lalanne R, Spiluttini B, Bertin C, Arsac LM. Effects of heart rate variability biofeedback training in athletes exposed to stress of university examinations. PLoS One 2018;13(7):e0201388 [FREE Full text] [doi: 10.1371/journal.pone.0201388] [Medline: 30048519]
46. Peira N, Pourtois G, Fredrikson M. Learned cardiac control with heart rate biofeedback transfers to emotional reactions. PLoS One 2013;8(7):e70004 [FREE Full text] [doi: 10.1371/journal.pone.0070004] [Medline: 23894574]
47. Gross MJ, Shearer DA, Bringer JD, Hall R, Cook CJ, Kilduff LP. Abbreviated resonant frequency training to augment heart rate variability and enhance on-demand emotional regulation in elite sport support staff. Appl Psychophysiol Biofeedback 2016;41(3):263-274. [doi: 10.1007/s10484-015-9330-9] [Medline: 26782084]
48. Schäfer A, Vagedes J. How accurate is pulse rate variability as an estimate of heart rate variability? A review on studies comparing photoplethysmographic technology with an electrocardiogram. Int J Cardiol 2013;166(1):15-29. [doi: 10.1016/j.ijcard.2012.03.119] [Medline: 22809539]
49. Allen J. Photoplethysmography and its application in clinical physiological measurement. Physiol Meas 2007;28(3):R1-R39. [doi: 10.1088/0967-3334/28/3/R01] [Medline: 17322588]
50. Chung AH, Gevirtz RN, Gharbo RS, Thiam MA, Ginsberg JPJ. Pilot study on reducing symptoms of anxiety with a heart rate variability biofeedback wearable and remote stress management coach. Appl Psychophysiol Biofeedback 2021;46(4):347-358 [FREE Full text] [doi: 10.1007/s10484-021-09519-x] [Medline: 34308526]
51. Lin IM, Chen TC, Tsai HY, Fan SY. Four sessions of combining wearable devices and heart rate variability (HRV) biofeedback are needed to increase HRV indices and decrease breathing rates. Appl Psychophysiol Biofeedback 2023;48(1):83-95. [doi: 10.1007/s10484-022-09567-x] [Medline: 36350478]
52. Umair M, Chalabianloo N, Sas C, Ersoy C. HRV and stress: a mixed-methods approach for comparison of wearable heart rate sensors for biofeedback. IEEE Access 2021;9:14005-14024. [doi: 10.1109/access.2021.3052131]
53. Khushhal A, Nichols S, Evans W, Gleadall-Siddall DO, Page R, O'Doherty AF, et al. Validity and reliability of the Apple watch for measuring heart rate during exercise. Sports Med Int Open 2017;1(6):E206-E211 [FREE Full text] [doi: 10.1055/s-0043-120195] [Medline: 30539109]
54. Turki A, Behbehani K, Ding K, Zhang R, Li M, Bell K. Estimation of heart rate variability measures using Apple watch and evaluating their accuracy: estimation of heart rate variability measures using Apple watch. 2021 Presented at: PETRA '21: Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference; June 29-July 2, 2021; Corfu, Greece p. 565-574. [doi: 10.1145/3453892.3462647]
55. Giles D, Draper N, Neil W. Validity of the Polar V800 heart rate monitor to measure RR intervals at rest. Eur J Appl Physiol 2016;116(3):563-571 [FREE Full text] [doi: 10.1007/s00421-015-3303-9] [Medline: 26708360]
56. Schuurmans AAT, de Looff P, Nijhof KS, Rosada C, Scholte RHJ, Popma A, et al. Validity of the Empatica E4 wristband to measure heart rate variability (HRV) parameters: a comparison to electrocardiography (ECG). J Med Syst 2020;44(11):190 [FREE Full text] [doi: 10.1007/s10916-020-01648-w] [Medline: 32965570]
57. de Zambotti M, Baker FC, Willoughby AR, Godino JG, Wing D, Patrick K, et al. Measures of sleep and cardiac functioning during sleep using a multi-sensory commercially-available wristband in adolescents. Physiol Behav 2016;158:143-149 [FREE Full text] [doi: 10.1016/j.physbeh.2016.03.006] [Medline: 26969518]
58. van Dijk W, Huizink AC, Oosterman M, Lemmers-Jansen ILJ, de Vente W. Validation of photoplethysmography using a mobile phone application for the assessment of heart rate variability in the context of heart rate variability-biofeedback. Psychosom Med 2023;85(7):568-576. [doi: 10.1097/PSY.0000000000001236] [Medline: 37678565]
59. Connor KM, Davidson JRT. Development of a new resilience scale: the Connor-Davidson Resilience Scale (CD-RISC). Depress Anxiety 2003;18(2):76-82. [doi: 10.1002/da.10113] [Medline: 12964174]
60. Hays RD, Schalet BD, Spritzer KL, Cella D. Two-item PROMIS® global physical and mental health scales. J Patient Rep Outcomes 2017;1(1):2 [FREE Full text] [doi: 10.1186/s41687-017-0003-8] [Medline: 29757325]
61. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983;24(4):385-396. [doi: 10.2307/2136404] [Medline: 6668417]
62. Kroenke K, Spitzer RL, Williams JBW, Löwe B. An ultra-brief screening scale for anxiety and depression: the PHQ-4. Psychosomatics 2009;50(6):613-621. [doi: 10.1176/appi.psy.50.6.613] [Medline: 19996233]
63. Broderick JE, DeWitt EM, Rothrock N, Crane PK, Forrest CB. Advances in patient-reported outcomes: the NIH PROMIS(®) measures. EGEMS (Wash DC) 2013;1(1):1015 [FREE Full text] [doi: 10.13063/2327-9214.1015] [Medline: 25848562]
64. McCraty R. Science of the Heart: Exploring the Role of the Heart in Human Performance, Volume 2. Boulder Creek, CA: HeartMath Institute; 2016.
65. Lemaire JB, Wallace JE, Lewin AM, de Grood J, Schaefer JP. The effect of a biofeedback-based stress management tool on physician stress: a randomized controlled clinical trial. Open Med 2011;5(4):e154-e163 [FREE Full text] [Medline: 22567069]
66. Ginsberg JP, Berry ME, Powell DA. Cardiac coherence and posttraumatic stress disorder in combat veterans. Altern Ther Health Med 2010;16(4):52-60. [Medline: 20653296]
67. Luskin F, Reitz M, Newell K, Quinn TG, Haskell W. A controlled pilot study of stress management training of elderly patients with congestive heart failure. Prev Cardiol 2002;5(4):168-172 [FREE Full text] [doi: 10.1111/j.1520.037x.2002.01029.x] [Medline: 12417824]

68. McCraty R, Atkinson M, Tomasino D. Impact of a workplace stress reduction program on blood pressure and emotional health in hypertensive employees. J Altern Complement Med 2003;9(3):355-369 [FREE Full text] [doi: 10.1089/107555303765551589] [Medline: 12816624]
69. Mundell L. The gift of HeartMath. Look insideKP Northern California. 2023. URL: https://lookinside.kaiserpermanente.org/the-gift-of-heartmath/ [accessed 2023-11-09]
70. Mobile apps and online tools. U.S. Department of Veterans Affairs. URL: https://www.va.gov/WHOLEHEALTH/veteran-resources/MobileApps-OnlineTools.asp [accessed 2023-11-09]
71. Lu G, Yang F, Taylor JA, Stein JF. A comparison of photoplethysmography and ECG recording to analyse heart rate variability in healthy subjects. J Med Eng Technol 2009;33(8):634-641. [doi: 10.3109/03091900903150998] [Medline: 19848857]
72. Pruneti C, Ferrari S, Guidotti S. A narrative review of heart rate variability as a good index of psychophysical health in athletes and in biofeedback training. J Clin Sport Psychol 2023;2023:1-28. [doi: 10.1123/jcsp.2022-0016]
73. The science of HeartMath. HeartMath. URL: https://www.heartmath.com/science/ [accessed 2023-11-09]
74. HeartMath® Inner Balance™ trainer coherence scoring system. HeartMath. URL: https://cdn.heartmath.com/manuals/IB_Coherence_Achevment_rev1a.pdf [accessed 2023-11-09]
75. Monitor your heart rate with Apple watch. Apple Inc. URL: https://support.apple.com/en-us/HT204666 [accessed 2020-10-22]
76. Hirten RP, Danieletto M, Tomalin L, Choi KH, Zweig M, Golden E, et al. Use of physiological data from a wearable device to identify SARS-CoV-2 infection and symptoms and predict COVID-19 diagnosis: observational study. J Med Internet Res 2021;23(2):e26107 [FREE Full text] [doi: 10.2196/26107] [Medline: 33529156]
77. Zhang Y, Song S, Vullings R, Biswas D, Simões-Capela N, van Helleputte N, et al. Motion artifact reduction for wrist-worn photoplethysmograph sensors based on different wavelengths. Sensors (Basel) 2019;19(3):673 [FREE Full text] [doi: 10.3390/s19030673] [Medline: 30736395]
78. Malpas SC, Purdie GL. Circadian variation of heart rate variability. Cardiovasc Res 1990;24(3):210-213. [doi: 10.1093/cvr/24.3.210] [Medline: 2346954]
79. Lombardi F, Sandrone G, Mortara A, La Rovere MT, Colombo E, Guzzetti S, et al. Circadian variation of spectral indices of heart rate variability after myocardial infarction. Am Heart J 1992;123(6):1521-1529. [doi: 10.1016/0002-8703(92)90804-5] [Medline: 1595531]
80. Cornelissen G. Cosinor-based rhythmometry. Theor Biol Med Model 2014;11:16 [FREE Full text] [doi: 10.1186/1742-4682-11-16] [Medline: 24725531]
81. Hou R, Tomalin LE, Suárez-Fariñas M. cosinoRmixedeffects: an R package for mixed-effects cosinor models. BMC Bioinformatics 2021;22(1):553 [FREE Full text] [doi: 10.1186/s12859-021-04463-3] [Medline: 34773978]
82. Hirten RP, Danieletto M, Tomalin L, Choi KH, Zweig M, Golden E, et al. Factors associated with longitudinal psychological and physiological stress in health care workers during the COVID-19 pandemic: observational study using Apple watch data. J Med Internet Res 2021;23(9):e31295 [FREE Full text] [doi: 10.2196/31295] [Medline: 34379602]
83. Hirten R, Suprun M, Danieletto M, Zweig M, Golden E, Pyzik R, et al. A machine learning approach to determine resilience utilizing wearable device data: analysis of an observational cohort. JAMIA Open 2023;6(2):ooad029 [FREE Full text] [doi: 10.1093/jamiaopen/ooad029] [Medline: 37143859]
84. Campbell-Sills L, Forde DR, Stein MB. Demographic and childhood environmental predictors of resilience in a community sample. J Psychiatr Res 2009;43(12):1007-1012. [doi: 10.1016/j.jpsychires.2009.01.013] [Medline: 19264325]
85. Burch JB, Ginsberg JP, McLain AC, Franco R, Stokes S, Susko K, et al. Symptom management among cancer survivors: randomized pilot intervention trial of heart rate variability biofeedback. Appl Psychophysiol Biofeedback 2020;45(2):99-108. [doi: 10.1007/s10484-020-09462-3] [Medline: 32358782]
86. de Souza PM, de Cássia Souza M, Diniz LA, Araújo CRV, Lopez M, Volchan E, et al. Long-term benefits of heart rate variability biofeedback training in older adults with different levels of social interaction: a pilot study. Sci Rep 2022;12(1):18795 [FREE Full text] [doi: 10.1038/s41598-022-22303-z] [Medline: 36335140]
87. Christensen H, Griffiths KM, Farrer L. Adherence in internet interventions for anxiety and depression. J Med Internet Res 2009;11(2):e13 [FREE Full text] [doi: 10.2196/jmir.1194] [Medline: 19403466]

Abbreviations

ANS: autonomic nervous system
CD-RISC 10: 10-item Connor-Davidson Resilience Scale
ECG: electrocardiography
HCW: health care worker
HRV: heart rate variability
MESOR: midline-estimating statistic of rhythm
NIH: National Institutes of Health
PHQ-4: Patient Health Questionnaire-4
PPG: photoplethysmography

PROMIS: Patient-Reported Outcomes Measurement Information System
PSS-10: Perceived Stress Scale 10
SDNN: SD of the interbeat interval of normal sinus beats

Edited by J Torous; submitted 15.12.23; peer-reviewed by R Gevirtz, L Guadagnoli; comments to author 19.01.24; revised version received 06.02.24; accepted 20.02.24; published 25.04.24.

Please cite as:
Hirten RP, Danieletto M, Landell K, Zweig M, Golden E, Pyzik R, Kaur S, Chang H, Helmus D, Sands BE, Charney D, Nadkarni G, Bagiella E, Keefer L, Fayad ZA
Remote Short Sessions of Heart Rate Variability Biofeedback Monitored With Wearable Technology: Open-Label Prospective Feasibility Study
JMIR Ment Health 2024;11:e55552
URL: https://mental.jmir.org/2024/1/e55552
doi:10.2196/55552
PMID:38663011

Original Paper

Psychometric Assessment of an Item Bank for Adaptive Testing on Patient-Reported Experience of Care Environment for Severe Mental Illness: Validation Study

Sara Fernandes[1], PhD; Yann Brousse[1], MSc; Xavier Zendjidjian[1], MD, PhD; Delphine Cano[1], MSc; Jérémie Riedberger[1], MSc; Pierre-Michel Llorca[2], MD, PhD; Ludovic Samalin[2], MD, PhD; Daniel Dassa[1], MD, PhD; Christian Trichard[3], MD, PhD; Vincent Laprevote[4,5], MD, PhD; Anne Sauvaget[6], MD, PhD; Mocrane Abbar[7], MD, PhD; David Misdrahi[8], MD, PhD; Fabrice Berna[5], MD, PhD; Christophe Lancon[1], MD, PhD; Nathalie Coulon[9], MD, PhD; Wissam El-Hage[10], MD, PhD; Pierre-Emmanuel Rozier[11], MD; Michel Benoit[12], MD, PhD; Bruno Giordana[12], MD, PhD; Alejandra Caqueo-Urízar[13], PhD; Dong Keon Yon[14], MD, PhD; Bach Tran[1,15], PhD; Pascal Auquier[1], MD, PhD; Guillaume Fond[1], MD, PhD; Laurent Boyer[1], MD, PhD

[1]Assistance Publique-Hopitaux de Marseille, Aix-Marseille University, UR3279: Health Service Research and Quality of Life Center - CEReSS, Marseille, France

[2]Department of Psychiatry, Centre Hospitalier Universitaire de Clermont-Ferrand, University of Clermont Auvergne, Centre national de la recherche scientifique, Institut national polytechnique de Clermont Auvergne, Institut Pascal UMR 6602, Clermont-Ferrand, France

[3]Etablissement public de santé Barthélemy Durand, Etampes, France

[4]Department of Addictology and Psychiatry, Centre Psychothérapique de Nancy, Laxou, France

[5]Département de Psychiatrie, Centre Hospitalier Régional Universitaire de Strasbourg, Université de Strasbourg, Institut national de la santé et de la recherche médicale U1114, Fédération de Médecine Translationnelle de Strasbourg, Strasbourg, France

[6]Nantes Université, Centre Hospitalier Régional Universitaire de Nantes, Movement - Interactions - Performance - MIP UR 4334, Nantes, France

[7]Department of Psychiatry, Centre Hospitalier Universitaire de Nîmes, University of Montpellier, Nîmes, France

[8]National Centre for Scientific Research UMR 5287 - Institut de Neurosciences Cognitives et Intégratives d'Aquitaine, University of Bordeaux, Centre Hospitalier Charles Perrens, Bordeaux, France

[9]Centre Expert Schizophrénie, Centre Expert TSA-SDI et Centre Référent de Réhabilitation Psychosociale et de Remédiation Cognitive - C3R, Centre Hospitalier Alpes Isère, Grenoble, France

[10]Centre Hospitalier Régional Universitaire de Tours, Clinique Psychiatrique Universitaire, Tours, France

[11]Department of Psychiatry, Centre Hospitalier des Pyrénées, Pau, France

[12]Department of Psychiatry, Hopital Pasteur, University Hospital of Nice, Nice, France

[13]Instituto de Alta Investigación, Universidad de Tarapacá, Arica, Chile

[14]Center for Digital Health, Medical Science Research Institute, Kyung Hee University College of Medicine, Department of Pediatrics, Kyung Hee University Medical Center, Seoul, Republic of Korea

[15]Institute of Preventive Medicine and Public Health, Hanoi Medical University, Hanoi, Vietnam

Corresponding Author:
Sara Fernandes, PhD
Assistance Publique-Hopitaux de Marseille, Aix-Marseille University, UR3279: Health Service Research and Quality of Life Center - CEReSS
27, Boulevard Jean-Moulin
Marseille, 13385
France
Phone: 33 660185077
Email: sarah.fernandes@ap-hm.fr

Abstract

Background: The care environment significantly influences the experiences of patients with severe mental illness and the quality of their care. While a welcoming and stimulating environment enhances patient satisfaction and health outcomes, psychiatric facilities often prioritize staff workflow over patient needs. Addressing these challenges is crucial to improving patient experiences and outcomes in mental health care.

Objective: This study is part of the Patient-Reported Experience Measure for Improving Quality of Care in Mental Health (PREMIUM) project and aims to establish an item bank (PREMIUM-CE) and to develop computerized adaptive tests (CATs) to measure the experience of the care environment of adult patients with schizophrenia, bipolar disorder, or major depressive disorder.

Methods: We performed psychometric analyses including assessments of item response theory (IRT) model assumptions, IRT model fit, differential item functioning (DIF), item bank validity, and CAT simulations.

Results: In this multicenter cross-sectional study, 498 patients were recruited from outpatient and inpatient settings. The final PREMIUM-CE 13-item bank was sufficiently unidimensional (root mean square error of approximation=0.082, 95% CI 0.067-0.097; comparative fit index=0.974; Tucker-Lewis index=0.968) and showed an adequate fit to the IRT model (infit mean square statistic ranging between 0.7 and 1.0). DIF analysis revealed no item biases according to gender, health care settings, diagnosis, or mode of study participation. PREMIUM-CE scores correlated strongly with satisfaction measures (r=0.69-0.78; P<.001) and weakly with quality-of-life measures (r=0.11-0.21; P<.001). CAT simulations showed a strong correlation (r=0.98) between CAT scores and those of the full item bank, and around 79.5% (396/498) of the participants obtained a reliable score with the administration of an average of 7 items.

Conclusions: The PREMIUM-CE item bank and its CAT version have shown excellent psychometric properties, making them reliable measures for evaluating the patient experience of the care environment among adults with severe mental illness in both outpatient and inpatient settings. These measures are a valuable addition to the existing landscape of patient experience assessment, capturing what truly matters to patients and enhancing the understanding of their care experiences.

Trial Registration: ClinicalTrials.gov NCT02491866; https://clinicaltrials.gov/study/NCT02491866

(JMIR Ment Health 2024;11:e49916) doi:10.2196/49916

KEYWORDS

psychiatry; public mental health; schizophrenia; major depressive disorders; bipolar disorders; patient-reported experience measures; quality of care; health services research; computerized adaptive testing; real-world data

Introduction

The health care environment, which encompasses design features (ie, cleanliness, food, privacy, waiting time, basic amenities) and the overall atmosphere (or climate) [1], has been recognized as a significant factor influencing the experiences of patients with severe mental illness (SMI) [2-5]. It is an important factor in the quality of patient care [2,6-8], contributing to improved patient satisfaction [9] and improved health outcomes [10,11]. In a recent study, patients identified a welcoming environment as one of the most important aspects of their care [12]. Indeed, a calm and welcoming environment helps to improve patients' sense of control and empowerment and, consequently, reinforces their willingness to follow recommended treatments. In addition, the care environment is the patients' first impression and can lead to a positive image of the therapeutic process [13]. A supportive environment promotes communication between patients and staff, can help reduce stressful stimuli, and thus prevents relapses and risky behavior. The priority for psychiatric facilities is therefore to provide patients with a warm and safe atmosphere that allows for positive social interactions, with opportunities for stimulating activities, enabling patients to facilitate their recovery and transition to the community. Different theoretical models can shed light on the additional nonpharmacological and biopsychosocial effects of a patient's care experience, including the placebo response effects and the set and setting theory [14,15].

Recommended features to promote patient recovery [16-19] include smaller, home-like units with well-decorated common spaces, open designs, access to nature and daylight, and an environment that is clean, well laid out, and ensures privacy and security for personal effects. However, psychiatric facilities are often criticized for prioritizing staff workflow over patient needs [2], leading in some cases to a perceived "prison-like atmosphere" [16,20,21] characterized by conflicting routines and rules and a lack of stimulation [22,23]. Some patients have reported feelings of boredom, loneliness, and stigmatization in these environments [21-26]. The lack of stimulating activities and positive social interactions is a barrier to patients' successful recovery [24-30]. These negative experiences can contribute to decreased patient satisfaction, increased levels of anxiety and stress among patients, ineffective care, and signs of burnout among staff [27,31,32]. Emphasis should be placed on the design of psychiatric facilities, as a difficult environment is a barrier to care, and patients often perceive such an environment as a lack of attention from staff [30]. In psychiatry, patients cope with an unfamiliar and potentially stressful environment [33], and a better understanding of their experiences is essential to identify and improve current barriers.

Given this growing interest, it is necessary to provide a valid and reliable instrument for measuring patients' experience of the care environment, applicable to both inpatient and outpatient settings, as care pathways for patients with SMI often combine several care modalities. Previous research has demonstrated that patients with SMI can provide reliable and valid responses to self-administered questionnaires; the impact of psychiatric symptoms and cognitive deficits seems to be negligible [34,35]. The French group PREMIUM (Patient-Reported Experience Measure for Improving Quality of Care in Mental Health) is developing item banks and computerized adaptive tests (CATs) to improve the systematic use of patient-reported experience measures in mental health care [36]. The use of CATs significantly reduces measurement burden by administering a limited number of items targeted to the respondent's experience level, aiming to improve measurement accuracy.

The objective of this study was to calibrate an item bank and develop a CAT to assess the care environment experienced by adult patients with SMI. These measures will contribute to the current landscape of patient experience measures by providing a valuable complement to PREMIUM measures and capturing what really matters to patients.

Methods

Study Population and Procedure

This is a national, multicenter, cross-sectional study conducted between January 2016 and December 2021. Patients were recruited through in- and outpatient psychiatric settings of a French teaching hospital (Assistance Publique-Hôpitaux de Marseille), the FondaMental Foundation's expert centers [37], and through an online survey. In mental health settings, stable patients who met the inclusion criteria were identified and approached by a member of their usual care team to invite them to participate in the study. The link to the web survey was distributed through patient associations.

Inclusion criteria were as follows: age older than 18 years and younger than 65 years with a diagnosis of schizophrenia, bipolar disorder, or major depressive disorder (MDD), receiving inpatient or outpatient psychiatric care, and speaking or reading French. Vulnerable persons (ie, pregnant or nursing women, persons under legal protection) or those unable to complete a self-administered questionnaire were not included in the study.

Current recommendations suggest a sample size of 300-500 observations for multiparameter item response theory (IRT) models [38-40]. Consequently, we estimated that a sample of around 500 patients would be sufficient to obtain reasonably stable estimates.

Data Collection

Data were collected through paper questionnaires in health care settings and online through a web survey. Patients reported the following sociodemographic and clinical characteristics: gender, age, educational level, marital status, occupational status, main diagnosis (schizophrenia, bipolar disorders, or MDD), duration of illness, and quality of life (QoL) as measured using the medical outcome study 12-item Short Form (SF-12) [41], which describes 8 QoL dimensions: physical functioning, social functioning, role physical, role emotional, mental health, vitality, bodily pain, general health, and 2 composite scores for physical and mental QoL (ranging from 0 to 100, with higher scores indicating better QoL). Adequate psychometric properties of the SF-12 have been demonstrated among individuals with SMI [42], and the SF-12 has proven to be a good alternative to the SF-36 for minimizing response burden.

The PREMIUM for Care Environment (PREMIUM-CE) item bank consists of 16 items designed for patients with SMI and measures their experience regarding the care environment over the past 4 weeks. Participants respond to the items on a 5-point Likert scale ranging from "strongly disagree" to "strongly agree" with a "not applicable" response option. Additionally, an overall satisfaction item ("Overall, are you satisfied with the health care facilities in which you receive care?") and a visual analog scale (VAS; minimum 0 to maximum 10) were collected. PREMIUM-CE items were identified through face-to-face interviews with patients with SMI and a systematic review of existing patient-reported experience measure; then the item pool was refined based on an expert review and cognitive interviews with patients with SMI [4,5,36].

Statistical Analysis

Basic Descriptive Analysis

Descriptive statistics were calculated to describe participants' characteristics, including frequencies and percentages for categorical variables and means and SDs for continuous variables. Response rates, means and SDs, and ceiling and floor effects were also calculated for each item.

IRT Assumptions

Unidimensionality, local independence, and monotonicity are the 3 fundamental assumptions of IRT [43]. Data were randomly divided into 2 data sets (n=249 each), one for exploratory factor analysis and one for confirmatory factor analysis (CFA) with the weighted least squares mean and variance estimator to ensure that the PREMIUM-CE was sufficiently unidimensional [44]. Local independence was examined using residual correlations from the final CFA model. Monotonicity was examined using visual inspection of characteristic item curves.

Calibration and Fitting an IRT Model

Item parameters were estimated using the generalized partial credit model (GPCM) [45] and compared to the partial credit model [46]. IRT handles missing values by using full information maximum likelihood estimation, which uses all available information, and GPCM is recommended when the amount of missing data is high (20% or more) [38]. Item fit was assessed by examining the mean square infit statistics, which reflect the information-weighted mean squared residuals between the observed and expected response patterns. PREMIUM-CE scores (θ) were estimated by the Bayesian Expected a Posteriori estimation method [47], and a linear transformation was performed to obtain PREMIUM-CE scores ranging from 0 to 100 (higher scores indicate better experience with the care environment). The information curve of the final item bank was calculated, and high measurement precision was defined as an information score >10, corresponding to a reliability of >0.90 [39].

Differential Item Functioning Analysis

DIF was examined using ordinal logistic regression models [48,49] by gender (man vs woman), age (median distribution), care setting (outpatient vs inpatient), psychiatric diagnosis (schizophrenia vs bipolar disorder vs MDD), and mode of study participation (online survey vs health care settings).

External Validity

Construct validity was examined through convergent and discriminative validity assessments. For convergent validity, Spearman's rank correlations were computed between

PREMIUM-CE scores and both satisfaction (global satisfaction item and VAS) and QoL (SF-12 subscales and composite scores) scores. Our hypothesis was that PREMIUM-CE scores would have strong correlations with satisfaction scores (r>0.60), which are 2 related measures, and weak correlations with QoL scores (r<0.30). For discriminant validity, relationships between PREMIUM-CE scores and sociodemographic and clinical characteristics of the respondents were examined by using 2-tailed t tests, ANOVA, and Pearson correlations. The Q-Q plot was used to determine that the data are approximately normally distributed. Based on previous studies of the determinants of patient satisfaction with psychiatric services [3,50,51], our hypotheses were that higher levels of patient experience of the care environment were associated with older age, being female, being nonsingle, or being in an outpatient setting.

CAT Simulations

These simulations using participants' actual responses were run from the calibrated item bank and compared to identify the best performing CAT version. The stopping rules were based on standard error of measurement (SEM) values of 0.33, 0.44, and 0.55 (corresponding to a reliability between 0.90 and 0.70 [52]). The item administered at baseline was the one that offered the most information to the population mean (θ=0), and then items were administered according to the maximum Fisher information criterion [53].

The indicators used at each stage of the psychometric analyses are presented in Table S1 in Multimedia Appendix 1 [44,54-70]. All of the statistical analyses were performed using the following software: SPSS (version 20.0; IBM Corp), MPlus (version 7.0; Muthen & Muthen), and R (version 4.2.0; R Core Team), using packages *mirt* [71], *lordif* [72], *BifactorIndicesCalculator* [73], and *mirtCAT* [74]. A 2-tailed P<.05 was considered statistically significant.

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki and approved by the relevant ethics committee (2014-A01152-45). The study was registered in ClinicalTrials.gov (NCT02491866). All participants provided nonopposition, as required by French law. Additionally, all data were anonymized.

Results

Sample Characteristics

Of the 498 participants, 50.2% (250/498) were men, 72.3% (345/477) were unemployed, 73.5% (350/476) were single, and 70.5% (337/478) had an education level of a bachelor's degree or higher. The average age was 40.9 (SD 11.9) years, and the mean duration of illness was 12.9 (SD 9.3) years. In total, 51.8% (253/488) of the participants had a diagnosis of schizophrenia, 24.4% (119/488) had bipolar disorder, or 23.8% (116/488) had MDD, and 77.7% (387/497) of them were outpatients. The characteristics of the sample are presented in Table 1.

Table 1. Sample characteristics.

Characteristics	Values
Study participation, n/N (%)	
Health care setting	271/498 (54.4)
Online survey	227/498 (45.6)
Sociodemographic data	
Gender (man), n/N (%)	250/498 (50.2)
Age (years), mean (SD; n=496)	40.9 (11.9)
Marital status (single), n/N (%)	350/476 (73.5)
Educational level (<bachelor's degree), n/N (%)	141/478 (29.5)
Employment status (unemployed), n/N (%)	345/477 (72.3)
Clinical data	
Care setting, n/N(%)	
Outpatient	387/498 (77.7)
Inpatient	111/498 (22.3)
Inpatient with involuntary commitment	40/111 (36.1)
Main diagnosis (n=488), n (%)	
Schizophrenia	253 (51.8)
Bipolar disorder	119 (24.4)
Major depressive disorder	116 (23.8)
Duration of illness (years; n=469)	
Value, mean (SD)	12.9 (9.3)
<5 years, n (%)	105 (22.4)
≥5 years, n (%)	364 (77.6)
Quality of life (SF-12[a] scores), mean (SD)	
Physical functioning (n=490)	46.5 (11.4)
Social functioning (n=491)	34.3 (11.8)
Role physical (n=491)	40.5 (11.1)
Role emotional (n=491)	33.3 (12.4)
Mental health (n=493)	45.0 (11.1)
Vitality (n=491)	51.2 (10.3)
Bodily pain (n=493)	44.1 (12.8)
General health (n=492)	34.8 (10.5)
Physical composite (n=484)	43.8 (10.3)
Mental composite (n=484)	39.3 (11.5)

[a]SF-12: 12-item Short Form.

Basic Descriptive Statistics

The mean item scores ranged from 2.07 (SD 1.32) to 3.24 (SD 0.89), and most items had a missing data rate <10% (except items CE10, CE12, and CE15). The floor and ceiling effects ranged from 1.8% to 10.6% and from 10% to 45.2%, respectively. The interitem correlation values ranged from 0.01 to 0.79, and 3 pairs of items showed too high interitem correlations (>0.70): items CE3-CE4 (r=0.73), items CE3-CE5 (r=0.78), and items CE4-CE5 (r=0.79). Items CE3 and CE5 were excluded because their content was considered less relevant than the remaining items. The lowest scores were for item CE15 ("food was of good quality"), item CE12 ("you had access to media (telephone, computer, internet or Wi-Fi connection, etc)," and item CE10 ("the health care facilities were well equipped"). Table 2 summarizes the distribution of responses for each item.

Table 2. Descriptive statistics of PREMIUM-CE item bank.

Item number	Content item	Score, mean (SD)	Floor effect (%)	Ceiling effect (%)	Missing values (%)	Skewness coefficient	Interitem correlations (Range)
CE1	The health care facilities were easily accessible (distance from home, parking, etc)	3.13 (1.09)	4.6	45.2	2	–1.42	0.01-0.66
CE2	The health care facilities were easy to find (eg, signage present and adapted)	3.13 (1.04)	3.4	43.4	1.8	–1.34	0.10-0.51
CE3	The health care facilities were welcoming	2.86 (1.17)	6.2	35.1	0.2	–0.96	0.27-0.78
CE4	The health care facilities were well-laid-out	2.91 (1.09)	4.8	33.5	0.4	–1.04	0.26-0.79
CE5	The health care facilities were pleasant	2.68 (1.19)	6.8	28.7	0	–0.72	0.27-0.79
CE6	The health care facilities were quiet enough	2.83 (1.15)	7.0	30.9	0.4	–1.06	0.21-0.50
CE7	The health care facilities were comfortable (chairs, armchairs, beds, etc)	2.89 (1.07)	4.8	30.1	0.4	–1.09	0.31-0.64
CE8	The health care facilities were clean	3.24 (0.89)	1.8	44.8	0.2	–1.46	0.21-0.63
CE9	The health care facilities were adapted to your needs	3.01 (1.03)	3.4	36.3	1.2	–1.13	0.23-0.68
CE10	The health care facilities were well equipped (materials for activities, group rooms, etc)	2.63 (1.22)	5.8	21.3	20.5	–0.68	0.25-0.66
CE11	The waiting time was acceptable	2.76 (1.21)	7.2	30.7	1.8	–0.90	0.30-0.49
CE12	You had access to media (telephone, computer, internet or Wi-Fi connection, etc)	2.24 (1.39)	10.0	17.5	26.7	–0.21	0.18-0.52
CE13	The sanitary facilities (toilets, bathroom, etc) were clean	3.08 (1.05)	3.2	39.2	7.4	–1.24	0.19-0.63
CE14	The health care facilities guarantee the respect for your privacy	3.12 (1.06)	4.4	41.8	4.6	–1.41	0.25-0.59
CE15	The food was of good quality, if you had to eat	2.07 (1.32)	10.6	10.0	39.4	–0.15	0.01-0.31
CE16	The smoking ban was respected	3.16 (1.02)	3.2	42.6	6.8	–1.37	0.21-0.50

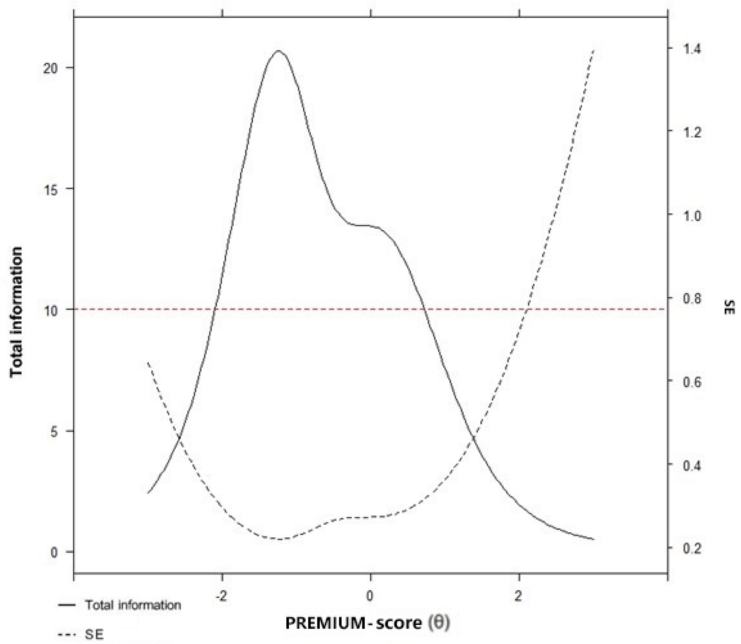
IRT Assumptions

In EFA, 2 factors had eigenvalue greater than 1, and the scree plot and parallel analysis indicated 2 factors. The eigenvalue of the first factor was 6.46 and explained 46.11% of the total variance; the second factor was 1.33, and the ratio was 4.86. Evaluations indicated that the 2 spatial accessibility items (CE1 and CE2) may form a separate factor, and after a content review, only item CE1 was kept as it was deemed the most relevant. The 1-factor CFA model provided evidence to support the unidimensionality of the remaining 13 items (root mean square error of approximation [RMSEA]=0.082; 95% CI 0.067-0.097; comparative fit index=0.974; Tucker-Lewis index=0.968) and no items showed local dependence (all residual correlations were above |0.20|). Of the 13 items in the bank, 10 were recoded to meet the monotonicity assumption (Table S2 in Multimedia Appendix 1), which improved the model fit (Akaike information criterion=–3343.78 and Bayes information criterion=–3428). Cronbach α was .91.

Calibration and Fitting an IRT Model

The GPCM was used to calibrate the item bank and showed superior fit compared to the partial credit model (10,192.60 and 10,367.43 for Akaike information criterion and 10,382.07 and 10,506.38 for Bayes information criterion; and χ^2=198.84; *P*<.001); item fit was good (infit values ranging between 0.74 and 1.00). IRT parameter estimates for the 13 items showed slopes ranging from 0.55 to 2.85 and thresholds ranging from –2.07 to 2.29. Item parameters and item fit are provided in Table S2 in Multimedia Appendix 1. As shown in Figure 1, PREMIUM-CE provided the most information in the scale range between –2.6 and 1.4 and had a high measurement accuracy (reliability >0.90) in a shorter range between –2.1 and 0.7 (which corresponds to 88.6% of total information). Item CE7 was the most informative of the bank—“the health care facilities were comfortable,” whereas item CE15 was the least informative—“the food was of good quality.”

Figure 1. The test information for the Patient-Reported Experience Measure for Improving Quality of Care in Mental Health for Care Environment (PREMIUM-CE) item bank.



Differential Item Functioning Analysis

Responses to items CE6 (quiet) and CE13 (sanitary) were flagged for overall DIF but with negligible magnitude according to health care settings. Likewise, the DIF magnitude was negligible for item CE16 (smoking ban) according to mode of study participation and for item CE15 (food) according to gender, mode of study participation, and diagnostic after pooling bipolar disorder and MDD (mood disorders vs schizophrenia; P=.02; ΔR^2=.013). None of the items showed significant DIF for age. DIF results are provided in Table S3 in Multimedia Appendix 1.

External Validity

As expected, there were strong correlations between the PREMIUM-CE item bank and overall satisfaction and VAS, supporting convergent validity. Similarly, all SF-12 dimensions were weakly correlated with the PREMIUM-CE item bank, except for bodily pain and vitality. Associations were found between better experience of the care environment (ie, higher PREMIUM-CE scores) and older age, being a woman, being voluntarily admitted to a hospital, and being recruited through health care facilities. There was no significant effect of educational level, marital status, employment status, diagnosis, or duration of illness. These results are presented in Table 3.

Table 3. Comparison of PREMIUM-CE scores with sociodemographic and clinical data and proxy measures of quality of care.

Characteristics	Correlation coefficient (*r*)	Mean (SD)	*P* value
Study participation			<.001
Health care setting	N/A[a]	63.12 (17.58)	
Online survey	N/A	52.53 (20.34)	
Sociodemographic data			
Age	0.19	N/A	<.001
Gender			.04
Man	N/A	56.52 (18.84)	
Woman	N/A	60.18 (20.19)	
Marital status			.51
Single	N/A	58.67 (19.12)	
Nonsingle	N/A	57.33 (21.37)	
Educational level			.07
<Bachelor's degree	N/A	60.88 (18.86)	
≥Bachelor's degree	N/A	57.34 (20.04)	
Employment status			.60
Employed	N/A	57.58 (19.36)	
Unemployed	N/A	58.64 (19.95)	
Clinical data			
Care setting			.01
Outpatient	N/A	57.34 (20.11)	
Inpatient voluntarily admitted	N/A	65.54 (17.61)	
Inpatient involuntarily admitted	N/A	54.33 (14.52)	
Main diagnosis			.58
Schizophrenia	N/A	57.46 (18.48)	
Bipolar disorder	N/A	59.73 (20.72)	
Major depressive disorder	N/A	58.22 (20.97)	
Duration of illness			.14
<5 years	N/A	60.86 (18.79)	
≥5 years	N/A	57.64 (20.08)	
Proxy measures			
Item of overall satisfaction	0.78	N/A	<.001
Visual analog scale	0.69	N/A	<.001
Quality of life (SF-12[b] scores)			
Physical functioning	0.14	N/A	.003
Social functioning	0.19	N/A	<.001
Role physical	0.21	N/A	<.001
Role emotional	0.19	N/A	<.001
Mental health	0.12	N/A	.01
Vitality	–0.04	N/A	.41
Bodily pain	0.09	N/A	.05
General health	0.13	N/A	.004

Characteristics	Correlation coefficient (*r*)	Mean (SD)	*P* value
Physical composite	0.14	N/A	.001
Mental composite	0.11	N/A	.01

[a]N/A: not applicable.

[b]SF-12: 12- items short form.

CAT Simulations

As reported in Table 4, the results of the CAT simulations based on SEM <.33 and <.44 were both acceptable in terms of accuracy and precision, but the scenario based on SEM <.33 (corresponding to a reliability of 0.90) was the most efficient. Of the 498 participants included in the simulation, 79.5% (396) achieved a reliable score with an average of 7 items administered.

Table 4. Mean scores and precision indicators for each computerized adaptive test simulation.

Precision level and indicators	Values
SEM[a]<0.33	
Mean (SD)	58.30 (19.39)
Correlation coefficient (*r*)	0.98
RMSE[b]	0.17
Mean number of items	6.95
SEM<0.44	
Mean (SD)	58.35 (18.75)
Correlation coefficient (*r*)	0.95
RMSE	0.29
Mean number of items	4.46
SEM<0.55	
Mean (SD)	50.57 (21.27)
Correlation coefficient (*r*)	0.92
RMSE	0.37
Mean number of items	3.10

[a]SEM: standard error of measurement.

[b]RMSE: root mean square error.

Discussion

Principal Findings

In this study, we report the calibration and initial evaluation of a new PREMIUM-CE item bank measuring patients' experience of the care environment that can be used for CATs. The PREMIUM-CE questionnaire is the first available questionnaire thus far to assess the quality of the care environment, applicable in outpatient and inpatient settings, for adults with SMI. This new measure covers different facets of the care environment, including ease of access in time and space, facility layout and basic amenities, food quality, comfort and cleanliness, respect for privacy, and smoking ban. PREMIUM-CE items address both concerns common to all patients (eg, cleanliness or food) and those more specific to psychiatric patients (eg, therapeutic workshops). Existing instruments measure more objective aspects (eg, checklists fulfilled by direct observation), and patients with SMI were not involved in the development and validation process [75].

PREMIUM-CE has undergone rigorous psychometric evaluation, consistent with previous studies conducted as part of the French PREMIUM initiative [36]. Although the RMSEA was slightly above the criterion of <.08, our results provide evidence of sufficient unidimensionality, and the item pool meets the assumptions for IRT modeling. Research has shown that the RMSEA statistic is problematic for assessing the unidimensionality of item banks measuring health concepts [76], as RMSEA is sensitive to model complexity (number of estimated parameters) and skewed data distributions [77]. These results are comparable to other calibration studies of item banks of patient-reported measures [78-83]. Overall, our results demonstrate that PREMIUM-CE has strong psychometric properties for patients with SMI, with negligible measurement bias by gender, health care settings, and mode of study participation. Items CE10, CE12, and CE15 had a higher rate of missing data than the other items, but this rate was below 40%, which remains acceptable by psychometrics standards [84]. In addition, these items had lower scores compared to others, meaning that efforts should be targeted on these aspects to improve the experience of patients with SMI. Future studies should examine whether changes to these items are required. The absence of a large DIF magnitude according to health care settings will make it possible to study changes in the experience of psychiatric patients over time, for whom care pathways often combine inpatient and outpatient care modalities. The 13 items in the final version of the PREMIUM-CE are listed in Multimedia Appendix 1, Table S4. In addition, the CAT version showed comparable measurement accuracy to the full item bank with high correlations between scores with an average of only 7 items administered.

External validity, explored using validated questionnaires and sociodemographic and clinical data, generally supported our initial hypotheses. Previous research has demonstrated that some factors, such as age, gender, marital status, and physical and mental health status, can influence individuals' experiences within a specific environment [2,3]. It is important to note, however, that the literature has not consistently established clear associations for age, gender, and marital status [3]. According to our results, older age, being female, being voluntarily admitted, and reporting a good physical and mental quality of life are associated with higher levels of patient experience of the care environment. As previously described [85], women reported higher levels of experience than men. Likewise, older people tend to be more accommodating, perhaps because they have fewer expectations than younger people [86]. Also, contrary to what might be expected, voluntarily admitted patients reported higher levels of experience than outpatients, although some patients reported a preference for community mental health treatment, which they considered less stigmatizing [87] and compatible with professional and social functioning. Furthermore, the literature has shown that hospitalization, particularly in the context of involuntary admission, can have a negative impact on patient experience [3], because it can be experienced as traumatic or particularly stressful for patients [88].

However, our results suggest that patients voluntarily admitted to the hospital may have a more holistic and structured experience compared to outpatients, conducive to positive therapeutic relationships with staff, whereas constraint has a negative effect on therapeutic relationships in the case of involuntarily admitted patients [88-90]. Finally, a positive but weak association was found between higher levels of patient experience and better QoL, as previously reported in other studies [51]. A calm and welcoming care environment contributes to patients feeling more comfortable and safer [50], which can reduce stress and anxiety and enhance relationships with staff, thereby promoting patients' recovery. Participants completing the online survey reported a poorer experience of the care environment than participants in health care settings because the latter may be more favorable due to fear of a negative effect on their relationships with staff, or this difference may be due to a possible recall bias.

The most poorly rated items by patients were related to accessing equipment (CE10), media (CE12), and food (CE15). Difficulties with access to equipment (eg, for art therapy) and media (eg, televisions or computers) are related to boredom, isolation, frustration, and higher levels of distress in patients [25]. A variety of individual or group activities could be offered to patients, such as therapeutic workshops in self-expression (ie, writing), art (ie, photography or painting), psychosocial rehabilitation (ie, cooking, which may also improve diet habits), or body awareness (ie, sophrology), to help patients develop social skills and promote social reintegration, improve confidence and self-esteem, build emotional resilience, and enjoy themselves. Facilities should have basic amenities such as affordable Wi-Fi and a working television in a common room accessible to all patients. Likewise, rooms should be equipped with a minimum package of free channels; however, not all facilities are equally equipped, and the cost of access to Wi-Fi and pay television channels can vary by as much as 2-fold. The content of what is broadcast on television should also be a therapeutic consideration. For example, it seems logical to avoid broadcasting distressing news or uninspiring programs and to favor the broadcasting of cultural works that could be the object of an exchange after viewing, such as a film club. The use of cell phones in health care settings presents challenges in terms of the potential risk of theft or breakage, as well as concerns about maintaining confidentiality. Additionally, it can be a source of tension with staff (eg, if the telephone credit is exceeded). There is no law that prohibiting the use of cell phones because communicating is a fundamental individual freedom, but the internal rules of the facilities can regulate their use by specifying the times and places of use and prohibit taking pictures of patients and staff. Furthermore, psychiatrists may occasionally prohibit a patient from keeping a cell phone, computer, or tablet as part of a medical decision, particularly in the case of placement in a seclusion room or for medical conditions. Previous studies have shown that a healthy diet is essential for good mental health and can prevent the worsening of symptoms [91,92], and that patients' satisfaction with hospital food services strongly influences their overall satisfaction with hospital care [93]. Diets such as the Mediterranean diet have been shown to improve patient outcomes [91]. Providing a menu tailored to patient preferences while focusing on food quality (taste, presentation, flavor, preparation, and variety), as well as the hospital environment, will help improve inpatient appetite and satisfaction [93]. In summary, the current challenges of hospital food service are to transition to a diet that is lower in meat, closer to the Mediterranean diet, without plastic packaging, and low in processed products while increasing the attractiveness of local and seasonal products, all while maintaining costs [91]. By contrast, the most highly rated items by patients were related to spatial accessibility (CE1), cleanliness (CE8), and smoking ban (CE16). Although health care facilities are under a total smoking ban throughout their whole facilities (including in specifically dedicated "smoking areas" or outside), the reality is often more flexible to accommodate patients who cannot leave the health care facilities, even temporarily (eg, patients under constraint). Proposals for smoking cessation assistance should be systematically offered to patients.

Limitations

Some limitations of this study are worth noting. Our sample size, while relatively modest, was sufficient to obtain accurate estimates. Current recommendations suggest that at least 300 observations are sufficient when using multiparameter models like the GPCM [38-40]. However, our results showed that the assumptions required for IRT calibration were met and that the model fit was adequate. In addition, some DIF analyses comparing subgroups with sample sizes smaller than those recommended for DIF analyses (at least 200 observations per group [94]) may have lacked the statistical power to detect a statistically significant DIF. These DIF findings should be regarded as preliminary, and future work with a larger sample will allow us to confirm these results. Although participants from the online survey and those from health care settings may have reported different levels of experience, this mixed survey design was chosen to ensure inclusivity across various subgroups, as supported by previous research on the equivalence of administration methods [95]. DIF analysis revealed that none of the items was flagged with a large DIF magnitude according to the patient's mode of study participation, suggesting that the data can be pooled without substantial bias. It was not possible to calculate a participation rate or to compare the characteristics of respondents and nonrespondents. This study was widely disseminated nationally, and our sample included inpatients and outpatients with diverse characteristics from different geographic regions of the country. Patients self-reported their diagnosis, and some data (488/498, 2.5%) were missing. However, the risk of misdiagnosis is considered minimal because all participants were fully informed about the study scope and diagnostic criteria. Additionally, this approach closely mirrors the real-world conditions of PREMIUM use. The title of the study mentioned general experience of care to limit the self-selection bias of patients with extreme care environment experiences. Future work will confirm the generalizability of our results. PREMIUM-CE has greater measurement accuracy for patients with scores between –2.1 and +0.7 (ie, reporting low to moderate levels of experience), and thus more items are needed to estimate scores for patients at both ends of the latent continuum. Future work should also reevaluate the precision and accuracy of the CAT in an independent sample and under real-world conditions. Finally, criterion validity could not be assessed because, to our knowledge, no gold standard was available and evidence for construct validity was limited. Future validation studies should examine the relationship between this new measure and objective assessments of the care environment (eg, evaluation by architects or other professionals).

Conclusion

The PREMIUM-CE item bank and its CAT version have demonstrated strong psychometric properties, making them robust measures for assessing patient experience of the care environment, applicable in both outpatient and inpatient settings, for adults with SMI. These measures contribute to the current landscape of patient experience measures by providing a valuable complement to PREMIUM measures of what really matters to patients.

Acknowledgments

The authors wish to thank Baumstarck Karine, Boucekine Mohamed, Loundou Anderson, and Michel Pierre for their statistical and methodological support. This research received funding through institutional grants from the French Programme de recherche sur la performance du système des soins and the Agence Technique de l'Information sur l'Hospitalisation. The sponsors had no role in study design, collection, analysis, and interpretation of data, report writing, or the decision to submit the study for publication.

Authors' Contributions

LB was responsible for the conceptualization, supervision, project administration, and funding acquisition. SF was responsible for the methodology, formal analysis, and writing the original draft. LB, SF, and GF are responsible for review and editing. All the authors read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Indicators of psychometric performance, parameter estimates (discrimination and thresholds) and fit statistics, differential item functioning results, and list of the 13-item of the Patient-Reported Experience Measure for Improving Quality of Care in Mental Health for Care Environment (PREMIUM-CE) item bank (English and French versions).

[DOCX File , 27 KB - mental_v11i1e49916_app1.docx]

References

1. Nicholls D, Kidd K, Threader J, Hungerford C. The value of purpose built mental health facilities: use of the Ward Atmosphere Scale to gauge the link between milieu and physical environment. Int J Ment Health Nurs 2015;24(4):286-294. [doi: 10.1111/inm.12138] [Medline: 25975388]
2. Daigle K, Frankel L. Collaboration to improve experience in hospital environments. In: Di Bucchianico G, Shin CS, Shim S, Fukuda S, Montagna G, Carvalho C, editors. Advances in Industrial Design. Cham, Switzerland: Springer; 2020:403-409.
3. Woodward S, Berry K, Bucci S. A systematic review of factors associated with service user satisfaction with psychiatric inpatient services. J Psychiatr Res 2017;92:81-93. [doi: 10.1016/j.jpsychires.2017.03.020] [Medline: 28412601]

4. Fernandes S, Fond G, Zendjidjian XY, Baumstarck K, Lançon C, Berna F, et al. Measuring the patient experience of mental health care: a systematic and critical review of patient-reported experience measures. Patient Prefer Adherence 2020;14:2147-2161 [FREE Full text] [doi: 10.2147/PPA.S255264] [Medline: 33192054]
5. Fernandes S, Fond G, Zendjidjian X, Michel P, Lançon C, Berna F, et al. A conceptual framework to develop a patient-reported experience measure of the quality of mental health care: a qualitative study of the PREMIUM project in France. J Mark Access Health Policy 2021;9(1):1885789 [FREE Full text] [doi: 10.1080/20016689.2021.1885789] [Medline: 33680364]
6. Henriksen K, Isaacson S, Sadler BL, Zimring CM. The role of the physical environment in crossing the quality chasm. Jt Comm J Qual Patient Saf 2007;33(Suppl 11):68-80. [doi: 10.1016/s1553-7250(07)33114-0] [Medline: 18173167]
7. Zhao Y, Mourshed M. Patients' perspectives on the design of hospital outpatient areas. Buildings 2017;7(4):117 [FREE Full text] [doi: 10.3390/buildings7040117]
8. Connellan K, Gaardboe M, Riggs D, Due C, Reinschmidt A, Mustillo L. Stressed spaces: mental health and architecture. HERD 2013;6(4):127-168. [doi: 10.1177/193758671300600408] [Medline: 24089185]
9. Jovanović N, Miglietta E, Podlesek A, Malekzadeh A, Lasalvia A, Campbell J, et al. Impact of the hospital built environment on treatment satisfaction of psychiatric in-patients. Psychol Med 2022;52(10):1969-1980. [doi: 10.1017/S0033291720003815] [Medline: 33087185]
10. Ulrich RS, Zimring C, Zhu X, DuBose J, Seo HB, Choi YS, et al. A review of the research literature on evidence-based healthcare design. HERD 2008;1(3):61-125. [doi: 10.1177/193758670800100306] [Medline: 21161908]
11. Jamshidi S, Parker JS, Hashemi S. The effects of environmental factors on the patient outcomes in hospital environments: a review of literature. Front Archit Res 2020;9(2):249-263 [FREE Full text] [doi: 10.1016/j.foar.2019.10.001]
12. Kelly EL, Davis L, Mendon S, Kiger H, Murch L, Pancake L, et al. Provider and consumer perspectives of community mental health services: implications for consumer-driven care. Psychol Serv 2019;16(4):572-584. [doi: 10.1037/ser0000244] [Medline: 29722997]
13. Ghaffari F, Shabak M, Norouzi N, Fallah SN. Hospital salutogenic public spaces: a conceptual framework of effective perceptional environment quality components on patients' satisfaction. Int J Build Pathol Adapt 2021;41(5):965-987. [doi: 10.1108/ijbpa-05-2021-0071]
14. Colloca L, Barsky AJ. Placebo and nocebo effects. N Engl J Med 2020;382(6):554-561. [doi: 10.1056/NEJMra1907805] [Medline: 32023375]
15. Hartogsohn I. Set and setting, psychedelics and the placebo response: an extra-pharmacological perspective on psychopharmacology. J Psychopharmacol 2016;30(12):1259-1267. [doi: 10.1177/0269881116677852] [Medline: 27852960]
16. Jovanović N, Campbell J, Priebe S. How to design psychiatric facilities to foster positive social interaction—a systematic review. Eur Psychiatry 2019;60:49-62 [FREE Full text] [doi: 10.1016/j.eurpsy.2019.04.005] [Medline: 31112827]
17. Ulrich RS, Bogren L, Gardiner SK, Lundin S. Psychiatric ward design can reduce aggressive behavior. J Environ Psychol 2018;57:53-66. [doi: 10.1016/j.jenvp.2018.05.002]
18. Shepley MM, Watson A, Pitts F, Garrity A, Spelman E, Kelkar J, et al. Mental and behavioral health environments: critical considerations for facility design. Gen Hosp Psychiatry 2016;42:15-21. [doi: 10.1016/j.genhosppsych.2016.06.003] [Medline: 27638966]
19. Borge L, Fagermoen MS. Patients' core experiences of hospital treatment: wholeness and self-worth in time and space. J Ment Health 2009;17(2):193-205. [doi: 10.1080/09638230701505996]
20. Staniszewska S, Mockford C, Chadburn G, Fenton SJ, Bhui K, Larkin M, et al. Experiences of in-patient mental health services: systematic review. Br J Psychiatry 2019;214(6):329-338 [FREE Full text] [doi: 10.1192/bjp.2019.22] [Medline: 30894243]
21. Shattell MM, Andes M, Thomas SP. How patients and nurses experience the acute care psychiatric environment. Nurs Inq 2008;15(3):242-250. [doi: 10.1111/j.1440-1800.2008.00397.x] [Medline: 18786217]
22. Molin J, Graneheim UH, Lindgren BM. Quality of interactions influences everyday life in psychiatric inpatient care--patients' perspectives. Int J Qual Stud Health Well-being 2016;11:29897 [FREE Full text] [doi: 10.3402/qhw.v11.29897] [Medline: 26806313]
23. Lindgren BM, Aminoff C, Graneheim UH. Features of everyday life in psychiatric inpatient care for self-harming: an observational study of six women. Issues Ment Health Nurs 2015;36(2):82-88. [doi: 10.3109/01612840.2014.941077] [Medline: 25625707]
24. Lilja L, Hellzén O. Former patients' experience of psychiatric care: a qualitative investigation. Int J Ment Health Nurs 2008;17(4):279-286. [doi: 10.1111/j.1447-0349.2008.00544.x] [Medline: 18666911]
25. Folke F, Hursti T, Kanter JW, Arinell H, Tungström S, Söderberg P, et al. Exploring the relationship between activities and emotional experience using a diary in a mental health inpatient setting. Int J Ment Health Nurs 2018;27(1):276-286. [doi: 10.1111/inm.12318] [Medline: 28220616]
26. Lindgren BM, Ringnér A, Molin J, Graneheim UH. Patients' experiences of isolation in psychiatric inpatient care: insights from a meta-ethnographic study. Int J Ment Health Nurs 2019;28(1):7-21. [doi: 10.1111/inm.12519] [Medline: 29975446]

27. Molin J, Graneheim UH, Ringnér A, Lindgren BM. From ideals to resignation—interprofessional teams perspectives on everyday life processes in psychiatric inpatient care. J Psychiatr Ment Health Nurs 2016;23(9-10):595-604. [doi: 10.1111/jpm.12349] [Medline: 27785861]
28. Donald F, Duff C, Lee S, Kroschel J, Kulkarni J. Consumer perspectives on the therapeutic value of a psychiatric environment. J Ment Health 2015;24(2):63-67. [doi: 10.3109/09638237.2014.954692] [Medline: 25915815]
29. Stewart D, Burrow H, Duckworth A, Dhillon J, Fife S, Kelly S, et al. Thematic analysis of psychiatric patients' perceptions of nursing staff. Int J Ment Health Nurs 2015;24(1):82-90. [doi: 10.1111/inm.12107] [Medline: 25382159]
30. Molin J, Strömbäck M, Lundström M, Lindgren BM. It's not just in the walls: patient and staff experiences of a new spatial design for psychiatric inpatient care. Issues Ment Health Nurs 2021;42(12):1114-1122 [FREE Full text] [doi: 10.1080/01612840.2021.1931585] [Medline: 34142934]
31. Lindgren BM, Molin J, Lundström M, Strömbäck M, Renberg ES, Ringnér A. Does a new spatial design in psychiatric inpatient care influence patients' and staff's perception of their care/working environment? A study protocol of a pilot study using a single-system experimental design. Pilot Feasibility Stud 2018;4:191 [FREE Full text] [doi: 10.1186/s40814-018-0383-4] [Medline: 30607254]
32. Laursen J, Danielsen A, Rosenberg J. Effects of environmental design on patient outcome: a systematic review. HERD 2014;7(4):108-119. [doi: 10.1177/193758671400700410] [Medline: 25303431]
33. Johansson IM, Skärsäter I, Danielson E. The meaning of care on a locked acute psychiatric ward: patients' experiences. Nord J Psychiatry 2009;63(6):501-507. [doi: 10.3109/08039480903118208] [Medline: 19688635]
34. Baumstarck K, Boyer L, Boucekine M, Aghababian V, Parola N, Lançon C, et al. Self-reported quality of life measure is reliable and valid in adult patients suffering from schizophrenia with executive impairment. Schizophr Res 2013;147(1):58-67. [doi: 10.1016/j.schres.2013.03.008] [Medline: 23566495]
35. Reininghaus U, Priebe S. Measuring patient-reported outcomes in psychosis: conceptual and methodological review. Br J Psychiatry 2012;201(4):262-267 [FREE Full text] [doi: 10.1192/bjp.bp.111.107615] [Medline: 23028084]
36. Fernandes S, Fond G, Zendjidjian X, Michel P, Baumstarck K, Lancon C, et al. The Patient-Reported Experience Measure for Improving qUality of care in Mental health (PREMIUM) project in France: study protocol for the development and implementation strategy. Patient Prefer Adherence 2019;13:165-177 [FREE Full text] [doi: 10.2147/PPA.S172100] [Medline: 30718945]
37. Schürhoff F, Fond G, Berna F, Bulzacka E, Vilain J, Capdevielle D, et al. A national network of schizophrenia expert centres: an innovative tool to bridge the research-practice gap. Eur Psychiatry 2015;30(6):728-735. [doi: 10.1016/j.eurpsy.2015.05.004] [Medline: 26072427]
38. Dai S, Vo TT, Kehinde OJ, He H, Xue Y, Demir C, et al. Performance of polytomous IRT models with rating scale data: an investigation over sample size, instrument length, and missing data. Front Educ 2021;6:721963 [FREE Full text] [doi: 10.3389/feduc.2021.721963]
39. Cappelleri JC, Lundy JJ, Hays RD. Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. Clin Ther 2014;36(5):648-662 [FREE Full text] [doi: 10.1016/j.clinthera.2014.04.006] [Medline: 24811753]
40. Nguyen TH, Han HR, Kim MT, Chan KS. An introduction to item response theory for patient-reported outcome measurement. Patient 2014;7(1):23-35 [FREE Full text] [doi: 10.1007/s40271-013-0041-0] [Medline: 24403095]
41. Ware J, Kosinski M, Keller S. How to score the SF-12 physical and mental health summary scales, 2nd edition. The Health Institute, New England Medical Center. Boston, MA; 1995. URL: https://www.researchgate.net/profile/John-Ware-6/publication/291994160_How_to_score_SF-12_items/links/58dfc42f92851c369548e04e/How-to-score-SF-12-items.pdf [accessed 2023-02-07]
42. Salyers MP, Bosworth HB, Swanson JW, Lamb-Pagone J, Osher FC. Reliability and validity of the SF-12 health survey among people with severe mental illness. Med Care 2000;38(11):1141-1150. [doi: 10.1097/00005650-200011000-00008] [Medline: 11078054]
43. Embretson SE, Reise SP. Item Response Theory for Psychologists. Mahwah, NJ: Lawrence Erlbaum Associates; 2000.
44. Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Med Care 2007;45(5 Suppl 1):S22-S31 [FREE Full text] [doi: 10.1097/01.mlr.0000250483.85507.04] [Medline: 17443115]
45. Muraki E. A generalized partial credit model: application of an EM algorithm. Appl Psychol Meas 1992;16(2):159-176. [doi: 10.1177/014662169201600206]
46. Masters GN. A rasch model for partial credit scoring. Psychometrika 1982;47(2):149-174. [doi: 10.1007/BF02296272]
47. Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika 1981;46(4):443-459. [doi: 10.1007/bf02293801]
48. Zieky M. Practical questions in the use of DIF statistics in test development. In: Wainer H, Holland PW, editors. Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum Associates; 1993:337-347.
49. Rogers HJ. Differential item functioning. In: Everitt BS, Howell DC, editors. Encyclopedia of Statistics in Behavioral Science. Chichester, UK: John Wiley & Sons; 2005:485-490.

50. Chen H, Li M, Wang J, Xue C, Ding T, Nong X, et al. Factors influencing inpatients' satisfaction with hospitalization service in public hospitals in Shanghai, People's Republic of China. Patient Prefer Adherence 2016;10:469-477 [FREE Full text] [doi: 10.2147/PPA.S98095] [Medline: 27110101]
51. Priebe S, Miglietta E. Assessment and determinants of patient satisfaction with mental health care. World Psychiatry 2019;18(1):30-31 [FREE Full text] [doi: 10.1002/wps.20586] [Medline: 30600613]
52. Harvill LM. Standard error of measurement. Educ Meas Issues Pract 1991;10(2):33-41. [doi: 10.1111/j.1745-3992.1991.tb00195.x]
53. Choi SW, Swartz RJ. Comparison of CAT item selection criteria for polytomous items. Appl Psychol Meas 2009;33(6):419-440 [FREE Full text] [doi: 10.1177/0146621608327801] [Medline: 20011456]
54. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16(3):297-334. [doi: 10.1007/BF02310555]
55. Browne MW, Cudeck R. Alternative ways of assessing model fit. Sociol Methods Res 1992;21(2):230-258. [doi: 10.1177/0049124192021002005]
56. Kline RB. Principles and Practice of Structural Equation Modeling, 2nd Edition. New York, NY: Guilford Press; 2005.
57. Bjorner JB, Kosinski M, Ware JE. Calibration of an item pool for assessing the burden of headaches: an application of item response theory to the Headache Impact Test (HIT). Qual Life Res 2003;12(8):913-933. [doi: 10.1023/a:1026163113446] [Medline: 14651412]
58. Fliege H, Becker J, Walter OB, Bjorner JB, Klapp BF, Rose M. Development of a Computer-Adaptive Test for depression (D-CAT). Qual Life Res 2005;14(10):2277-2291. [doi: 10.1007/s11136-005-6651-9] [Medline: 16328907]
59. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Contr 1974;19(6):716-723. [doi: 10.1109/TAC.1974.1100705]
60. Schwarz G. Estimating the dimension of a model. Ann Statist 1978;6(2):461-464 [FREE Full text] [doi: 10.1214/aos/1176344136]
61. Ware JE, Bjorner JB, Kosinski M. Practical implications of item response theory and computerized adaptive testing: a brief summary of ongoing studies of widely used headache impact scales. Med Care 2000;38(Suppl 9):II73-I182. [Medline: 10982092]
62. Baker FB. The Basics of Item Response Theory, 2nd Edition. Washington, DC: ERIC Clearinghouse on Assessment and Evaluation; 2001.
63. Chang HH, Ying Z. A global information approach to computerized adaptive testing. Appl Psychol Meas 1996;20(3):213-229. [doi: 10.1177/014662169602000303]
64. Bond TG, Fox CM. Applying the Rasch Model: Fundamental Measurement in the Human Sciences, 3rd Edition. New York, NY: Lawrence Erlbaum Associates; 2015.
65. Wright B, Linacre JM. Reasonable mean-square fit values. Rasch Meas Trans 1994;8:370-371 [FREE Full text]
66. Zumbo BD. A handbook on the theory and methods of Differential Item Functioning (DIF): logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores. Directorate of Human Resources Research and Evaluation, Department of National Defense. Ottawa, ON; 1999. URL: https://faculty.educ.ubc.ca/zumbo/DIF/handbook.pdf [accessed 2023-02-20]
67. Choi SW, Reise SP, Pilkonis PA, Hays RD, Cella D. Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. Qual Life Res 2010;19(1):125-136 [FREE Full text] [doi: 10.1007/s11136-009-9560-5] [Medline: 19941077]
68. Michel P, Auquier P, Baumstarck K, Pelletier J, Loundou A, Ghattas B, et al. Development of a cross-cultural item bank for measuring quality of life related to mental health in multiple sclerosis patients. Qual Life Res 2015;24(9):2261-2271. [doi: 10.1007/s11136-015-0948-0] [Medline: 25712324]
69. Rose M, Bjorner JB, Becker J, Fries JF, Ware JE. Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). J Clin Epidemiol 2008;61(1):17-33. [doi: 10.1016/j.jclinepi.2006.06.025] [Medline: 18083459]
70. Ferketich S. Focus on psychometrics. Aspects of item analysis. Res Nurs Health 1991;14(2):165-168. [doi: 10.1002/nur.4770140211] [Medline: 2047538]
71. Chalmers RP. Mirt: a Multidimensional Item Response Theory package for the R Environment. J Stat Softw 2012;48(6):1-29 [FREE Full text] [doi: 10.18637/jss.v048.i06]
72. Choi SW, Gibbons LE, Crane PK. Lordif: an R Package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. J Stat Softw 2011;39(8):1-30 [FREE Full text] [doi: 10.18637/jss.v039.i08] [Medline: 21572908]
73. Dueber D. Package 'BifactorIndicesCalculator': a package for computing statistical indices relevant to bifactor measurement models. CRAN. 2022. URL: https://cran.r-project.org/web/packages/BifactorIndicesCalculator/BifactorIndicesCalculator.pdf [accessed 2023-03-16]
74. Chalmers RP. Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. J Stat Softw 2016;71(5):1-38 [FREE Full text] [doi: 10.18637/jss.v071.i05]

75. Elf M, Nordin S, Wijk H, Mckee KJ. A systematic review of the psychometric properties of instruments for assessing the quality of the physical environment in healthcare. J Adv Nurs 2017;73(12):2796-2816 [FREE Full text] [doi: 10.1111/jan.13281] [Medline: 28207946]
76. Reise SP, Scheines R, Widaman KF, Haviland MG. Multidimensionality and structural coefficient bias in structural equation modeling: a bifactor perspective. Educ Psychol Meas 2012;73(1):5-26. [doi: 10.1177/0013164412449831]
77. Cook KF, Kallen MA, Amtmann D. Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. Qual Life Res 2009;18(4):447-460 [FREE Full text] [doi: 10.1007/s11136-009-9464-4] [Medline: 19294529]
78. Kwan YH, Uy EJ, Bautista DC, Xin X, Xiao Y, Lee GL, et al. Development and calibration of a novel positive mindset item bank to measure Health-Related Quality of Life (HRQoL) in Singapore. PLoS One 2019;14(7):e0220293 [FREE Full text] [doi: 10.1371/journal.pone.0220293] [Medline: 31365554]
79. Haley SM, Fragala-Pinkham MA, Dumas HM, Ni P, Gorton GE, Watson K, et al. Evaluation of an item bank for a computerized adaptive test of activity in children with cerebral palsy. Phys Ther 2009;89(6):589-600 [FREE Full text] [doi: 10.2522/ptj.20090007] [Medline: 19423642]
80. Crins MHP, Terwee CB, Klausch T, Smits N, de Vet HCW, Westhovens R, et al. The Dutch-Flemish PROMIS physical function item bank exhibited strong psychometric properties in patients with chronic pain. J Clin Epidemiol 2017;87:47-58. [doi: 10.1016/j.jclinepi.2017.03.011] [Medline: 28363734]
81. Pilkonis PA, Yu L, Dodds NE, Johnston KL, Lawrence SM, Hilton TF, et al. An item bank for abuse of prescription pain medication from the Patient-Reported Outcomes Measurement Information System (PROMIS®). Pain Med 2017;18(8):1516-1527 [FREE Full text] [doi: 10.1093/pm/pnw233] [Medline: 28339555]
82. Rindestig FC, Wiberg M, Chaplin JE, Henje E, Dennhag I. Psychometrics of three Swedish physical pediatric item banks from the Patient-Reported Outcomes Measurement Information System (PROMIS)®: pain interference, fatigue, and physical activity. J Patient Rep Outcomes 2021;5(1):105 [FREE Full text] [doi: 10.1186/s41687-021-00382-2] [Medline: 34637029]
83. Terwee CB, Crins MHP, Boers M, de Vet HCW, Roorda LD. Validation of two PROMIS item banks for measuring social participation in the Dutch general population. Qual Life Res 2019;28(1):211-220 [FREE Full text] [doi: 10.1007/s11136-018-1995-0] [Medline: 30203302]
84. Kandel H, Khadka J, Goggin M, Pesudovs K. Patient-reported outcomes for assessment of quality of life in refractive error: a systematic review. Optom Vis Sci 2017;94(12):1102-1119. [doi: 10.1097/OPX.0000000000001143] [Medline: 29095758]
85. Olusina AK, Ohaeri JU, Olatawura MO. Patient and staff satisfaction with the quality of in-patient psychiatric care in a Nigerian general hospital. Soc Psychiatry Psychiatr Epidemiol 2002;37(6):283-288. [doi: 10.1007/s00127-002-0548-5] [Medline: 12111034]
86. Lundqvist LO, Ahlström G, Wilde-Larsson B, Schröder A. The patient's view of quality in psychiatric outpatient care: patients' view of psychiatric care. J Psychiatr Ment Health Nurs 2012;19(7):629-637. [doi: 10.1111/j.1365-2850.2012.01899.x] [Medline: 22385022]
87. Beckers T, Koekkoek B, Tiemens B, Jaeqx-van Tienen L, Hutschemaekers G. Substituting specialist care for patients with severe mental illness with primary healthcare. Experiences in a mixed methods study. J Psychiatr Ment Health Nurs 2019;26(1-2):1-10. [doi: 10.1111/jpm.12499] [Medline: 30270481]
88. Zendjidjian XY, Auquier P, Lançon C, Loundou A, Parola N, Faugère M, et al. Determinants of patient satisfaction with hospital health care in psychiatry: results based on the SATISPSY-22 questionnaire. Patient Prefer Adherence 2014;8:1457-1464 [FREE Full text] [doi: 10.2147/PPA.S67641] [Medline: 25368515]
89. Roche E, Madigan K, Lyne JP, Feeney L, O'Donoghue B. The therapeutic relationship after psychiatric admission. J Nerv Ment Dis 2014;202(3):186-192. [doi: 10.1097/NMD.0000000000000102] [Medline: 24566503]
90. Wyder M, Bland R, Blythe A, Matarasso B, Crompton D. Therapeutic relationships and involuntary treatment orders: service users' interactions with health-care professionals on the ward. Int J Ment Health Nurs 2015;24(2):181-189. [doi: 10.1111/inm.12121] [Medline: 25628260]
91. Fond G, Young AH, Godin O, Messiaen M, Lançon C, Auquier P, et al. Improving diet for psychiatric patients: high potential benefits and evidence for safety. J Affect Disord 2020;265:567-569. [doi: 10.1016/j.jad.2019.11.092] [Medline: 31757621]
92. Gill R, Tyndall SF, Vora D, Hasan R, Megna JL, Leontieva L. Diet quality and mental health amongst acute inpatient psychiatric patients. Cureus 2021;13(1):e12434 [FREE Full text] [doi: 10.7759/cureus.12434] [Medline: 33552754]
93. Messina G, Fenucci R, Vencia F, Niccolini F, Quercioli C, Nante N. Patients' evaluation of hospital foodservice quality in Italy: what do patients really value? Public Health Nutr 2013;16(4):730-737 [FREE Full text] [doi: 10.1017/S1368980012003333] [Medline: 22874795]
94. Scott NW, Fayers PM, Aaronson NK, Bottomley A, de Graeff A, Groenvold M, et al. Differential Item Functioning (DIF) analyses of health-related quality of life instruments using logistic regression. Health Qual Life Outcomes 2010;8:81 [FREE Full text] [doi: 10.1186/1477-7525-8-81] [Medline: 20684767]
95. Hagan TL, Belcher SM, Donovan HS. Mind the mode: differences in paper vs. web-based survey modes among women with cancer. J Pain Symptom Manage 2017;54(3):368-375 [FREE Full text] [doi: 10.1016/j.jpainsymman.2017.07.005] [Medline: 28711752]

Abbreviations

CAT: computerized adaptive test
CFA: confirmatory factor analysis
DIF: differential item functioning
GPCM: generalized partial credit model
IRT: item response theory
MDD: major depressive disorder
PREMIUM: Patient-Reported Experience Measure for Improving Quality of Care in Mental Health
PREMIUM-CE: Patient-Reported Experience Measure for Improving Quality of Care in Mental Health for Care Environment
QoL: quality of life
RMSEA: root mean square error of approximation
SEM: standard error of measurement
SF-12: 12-item Short Form
SMI: severe mental illness
VAS: visual analog scale

Edited by J Torous; submitted 13.06.23; peer-reviewed by M Kamouchi, D Dinh, L Hua; comments to author 11.08.23; revised version received 15.09.23; accepted 21.01.24; published 16.05.24.

Please cite as:
Fernandes S, Brousse Y, Zendjidjian X, Cano D, Riedberger J, Llorca PM, Samalin L, Dassa D, Trichard C, Laprevote V, Sauvaget A, Abbar M, Misdrahi D, Berna F, Lancon C, Coulon N, El-Hage W, Rozier PE, Benoit M, Giordana B, Caqueo-Urízar A, Yon DK, Tran B, Auquier P, Fond G, Boyer L
Psychometric Assessment of an Item Bank for Adaptive Testing on Patient-Reported Experience of Care Environment for Severe Mental Illness: Validation Study
JMIR Ment Health 2024;11:e49916
URL: https://mental.jmir.org/2024/1/e49916
doi:10.2196/49916
PMID:38753416

Original Paper

Reliability and Validity of Ecological Momentary Assessment Response Time–Based Measures of Emotional Clarity: Secondary Data Analysis

Raymond Hernandez[1,2], OTR/L, PhD; Claire Hoogendoorn[3,4], PhD; Jeffrey S Gonzalez[3,4], PhD; Elizabeth A Pyatak[2], OTR/L, PhD; Gladys Crespo-Ramos[4], PhD; Stefan Schneider[1,5], PhD

[1]Center for Economic and Social Research, University of Southern California, Los Angeles, CA, United States
[2]Chan Division of Occupational Science and Occupational Therapy, University of Southern California, Los Angeles, CA, United States
[3]Ferkauf Graduate School of Psychology, Yeshiva University, Bronx, NY, United States
[4]Department of Medicine (Endocrinology), Albert Einstein College of Medicine, Bronx, NY, United States
[5]Department of Psychology, University of Southern California, Los Angeles, CA, United States

Corresponding Author:
Raymond Hernandez, OTR/L, PhD
Center for Economic and Social Research
University of Southern California
635 Downey Way
Los Angeles, CA, 90089
United States
Phone: 1 2138211899
Email: hern939@usc.edu

Abstract

Background: Emotional clarity has often been assessed with self-report measures, but efforts have also been made to measure it passively, which has advantages such as avoiding potential inaccuracy in responses stemming from social desirability bias or poor insight into emotional clarity. Response times (RTs) to emotion items administered in ecological momentary assessments (EMAs) may be an indirect indicator of emotional clarity. Another proposed indicator is the *drift rate* parameter, which assumes that, aside from how fast a person responds to emotion items, the measurement of emotional clarity also requires the consideration of how careful participants were in providing responses.

Objective: This paper aims to examine the reliability and validity of RTs and drift rate parameters from EMA emotion items as indicators of individual differences in emotional clarity.

Methods: Secondary data analysis was conducted on data from 196 adults with type 1 diabetes who completed a 2-week EMA study involving the completion of 5 to 6 surveys daily. If lower RTs and higher drift rates (from EMA emotion items) were indicators of emotional clarity, we hypothesized that greater levels (ie, higher clarity) should be associated with greater life satisfaction; lower levels of neuroticism, depression, anxiety, and diabetes distress; and fewer difficulties with emotion regulation. Because prior literature suggested emotional clarity could be valence specific, EMA items for negative affect (NA) and positive affect were examined separately.

Results: Reliability of the proposed indicators of emotional clarity was acceptable with a small number of EMA prompts (ie, 4 to 7 prompts in total or 1 to 2 days of EMA surveys). Consistent with expectations, the average drift rate of NA items across multiple EMAs had expected associations with other measures, such as correlations of r=–0.27 (P<.001) with depression symptoms, r=–0.27 (P=.001) with anxiety symptoms, r=–0.15 (P=.03) with emotion regulation difficulties, and r=0.63 (P<.001) with RTs to NA items. People with a higher NA drift rate responded faster to NA emotion items, had greater subjective well-being (eg, fewer depression symptoms), and had fewer difficulties with overall emotion regulation, which are all aligned with the expectation for an emotional clarity measure. Contrary to expectations, the validities of average RTs to NA items, the drift rate of positive affect items, and RTs to positive affect items were not strongly supported by our results.

Conclusions: Study findings provided initial support for the validity of NA drift rate as an indicator of emotional clarity but not for that of other RT-based clarity measures. Evidence was preliminary because the sample size was not sufficient to detect small but potentially meaningful correlations, as the sample size of the diabetes EMA study was chosen for other more primary research questions. Further research on passive emotional clarity measures is needed.

(JMIR Ment Health 2024;11:e58352) doi:10.2196/58352

KEYWORDS

digital mental health; drift-diffusion model; ecological momentary assessment; emotional clarity; emotional health; emotion regulation; response time; positive affect; negative affect; type 1 diabetes; mobile phone

Introduction

Background

Emotion regulation is highly relevant to subjective well-being from both hedonic and eudaimonic perspectives. According to the hedonic perspective, well-being is the experience of happiness or the occurrence of positive affect (PA) and absence of negative affect (NA) [1]. In the eudaimonic perspective, well-being arises when individuals live with a sense of growth, meaningfulness, and purpose [2]. People often engage in emotion regulation to obtain hedonic benefits (eg, feel more PA and reduce NA) [3], but eudaimonic motivations have been found to be important as well [4] (eg, downregulation of negative emotions to achieve a sense of growth in one's ability to handle daily stressors).

An important aspect of emotion regulation is emotional clarity, a person's ability to lucidly identify the type of emotion they are experiencing [5]. Emotional clarity is highly relevant to the James Gross Model of Emotion Regulation [6]. The model consists of emotion regulation strategies that are either "antecedent-focused" or "response-focused," which refers to whether the strategy is used before or after an emotional response fully develops. In the updated emotion regulation model of Gross, understanding and identifying one's emotions accurately (ie, emotional clarity) are precursors to both these types of emotion regulation strategies [7]. Therefore, individuals with low emotional clarity are less likely to use emotion regulation strategies (as they are failing to identify the need for them), which can negatively impact well-being. Lower emotional clarity has often been associated with reduced mental health [8-10], although there are exceptions. For instance, prior research has suggested that higher emotional clarity may be adaptive primarily for individuals who do not have very frequent and strong experiences of negative emotions but maladaptive for those who frequently have strong feelings of NA [11].

Both direct and indirect measures of emotional clarity have been developed [12,13]. Direct assessments involve the metacognitive task of reflecting on one's emotional clarity level, while indirect assessments measure the performance of a task relevant to emotional clarity (ie, answering emotion items) and do not require self-insight [12]. Emotional clarity is commonly directly assessed with cross-sectional self-report measures such as the clarity subscales of the Trait Meta-Mood Scale [14] and the Difficulties in Emotion Regulation Scale (DERS) [15], but indirect measures have been argued to have potential advantages over self-report assessments [12,13]. For instance, they could help avoid possible issues with subjective reports, including poor insight into emotional clarity and the possibility of social desirability bias (ie, participants not wanting to report uncertainty about their feelings) [12,16]. However, for the potential utility to be realized, further investigations of the validity of indirect measures of emotional clarity are needed.

Assessing Emotional Clarity With Item Response Times

Response times (RTs) to emotion questions in ecological momentary assessments (EMAs) have been argued to be indirect measures of emotional clarity at the momentary (within-person) level [12]. Theoretically, the greater an individual's momentary affective clarity, the less time should be needed to provide a rating of momentary affect [12], whereas longer RTs to affect items should be indicative of lower emotional clarity. Evidence supporting this theory has been found, such as shorter RTs to affect items being associated with better momentary emotion regulation and mood [12]. However, emotional clarity may be confounded with emotional intensity [9], and evidence suggests the validity of RTs as a measure of affective clarity is enhanced by controlling for emotional intensity at the within-person (and not between-person) level [17].

There was no strong evidence that the study-long aggregates of emotion item RTs (between-person level) could act as indicators of trait emotional clarity [12,17]. These aggregates were found to not have significant relationships with global measures of emotional clarity and have inconsistent relationships with global emotion regulation measures [12,17]. Suggested reasons for the lack of a relationship between global measures of emotional clarity and the aggregate of emotion item RTs include a modality difference (ie, self-report vs indirect behavior-based assessment), low conceptual correspondence (ie, different forms of emotional clarity are being assessed), and difference in assessment timing (ie, 1 time vs repeated EMA measurements) [12,17]. The relationships between the study-long aggregate of emotion item RTs and subjective well-being variables of relevance to emotional clarity (eg, depression and anxiety) [8,18] were not examined, which could have served as useful additional convergent validity tests. Finally, the possibility that processing speed had a confounding effect on correlations between study-long aggregates of emotion item RTs and other variables was not investigated. In prior research, RTs to emotion EMA items have been found to have moderate correlations with processing speed measures [19], suggesting that individual differences in emotion item RTs may at least in part be attributable to processing speed.

Assessing Emotional Clarity With the Drift Rate Parameter

Another indirect measure of emotional clarity was recently proposed, drift rate, which is computed using the drift-diffusion model [20]. This model, which is often used in cognitive psychology, was explicitly developed to disentangle different components of RTs [21]. The drift-diffusion model proposes that decisions (eg, choosing responses on EMA items) are made through the accumulation of information until a threshold of sufficient information (as determined by the individual) is reached [21]. Decisions can be fast if the speed of information accumulation (ie, drift rate) is fast, if the threshold for a decision (ie, boundary separation or response caution) is low, or both. The D-diffusion item response theory (IRT) model is an IRT version of the drift-diffusion model that was specifically designed for the analysis of self-report ratings, such as emotion ratings [22]. In the context of answering mood questions, the D-diffusion IRT model additionally considers that RTs may also be fast if extremes of the emotion are experienced (ie, very high or low levels), making the provision of ratings more straightforward [20]. When the D-diffusion model is applied to EMA responses and their RTs, it can take into account all the aforementioned influences on RTs and output drift rate. In the EMA context, drift rate can be interpreted as the speed of access to information relevant to the question being asked [20,21], which here is information regarding emotions. A more detailed description of how the drift rate parameter was computed can be found in the *Emotional Clarity Measure 2: Drift Rate From the Drift-Diffusion Model* subsection in the *Methods* section.

A person would be considered to have high emotional clarity when responding both fast and carefully to emotion items, which differs from how emotional clarity assessed via RTs considers only speed. As the absolute difference between a person's NA level and the level of NA captured by an item increases, the D-diffusion model assumes that an individual would be expected to have faster RTs as a result of the so-called "distance-difficulty" principle, a well-established phenomenon in the RT literature [23], whereby the more items contrast with a person's state, the easier they are. For instance, an individual with a very low NA would be expected to quickly respond to an item asking about being scared (an item often associated with a high NA) [24], whereas the same individual would be expected to require a longer RT for an item asking about irritability (which has been associated with a low NA) [24]. According to the D-diffusion model, response caution is low when a person responds similarly quickly (or slowly) to all items regardless of their content (an indication that the person did not answer carefully). Assuming a particular observed item RT, when response caution is lower (ie, less information is gathered), drift rate (ie, the rate of information accumulation or information divided by time) would also be lower. Conversely, for a person with a higher response caution, the estimated drift rate (and emotional clarity) would be greater for the same RT.

Our initial evidence suggested that the drift rate parameter derived from EMA NA items may be an indicator of emotional clarity of low versus high NA, including expected relationships (ie, negative associations) at the between-person level with neuroticism and depression [20], but we did not account for factors aside from emotional clarity that could impact drift rate. For instance, drift rate could be impacted by individual differences in the general speed of responding to questionnaires (eg, due to reading speed, motor behavior, and familiarity with computer use). Studies on emotional clarity measured by survey item RTs had accounted for individual differences in this baseline speed of responding by adjusting for it [12], but this was not done in our prior study [20]. Cognitive processing speed was hypothesized to also potentially impact drift rate but was also not adjusted for, as the measure was not available in the prior study [20].

Objectives

Trait Emotional Clarity

There has been substantial prior research and interest in examining emotional clarity as a trait [12,14,15], but further research on the indirect (instead of direct) measurement of individual differences in emotional clarity via EMA data is needed. Compared to traditional cross-sectional self-report measures, indirect measurement of individual differences in emotional clarity potentially has the advantage of less susceptibility to self-report biases stemming from causes such as social desirability and poor insight [12,16]. Furthermore, if an emotion is being assessed via EMA, *indirect assessment of emotional clarity via emotion items affords the possibility of capturing individual differences in emotional clarity without burdening participants with additional items*. The use of EMA methodology is ubiquitous in a broad range of fields [25-27], and the use of EMA emotion items is commonplace [28]. Indirect assessment of trait emotional clarity would, therefore, make the investigation of emotional clarity possible for a large number of EMA data sets without additional emotional clarity items.

The focus of this paper was to examine the reliability and validity of 2 RT-based indirect indicators of individual differences in emotional clarity among adults with type 1 diabetes (T1D). One indicator was the average of repeated measures of RTs to emotion items, and the other was the average of repeated measures of drift rate (ie, the speed of accessing information about one's current affect) derived from emotion items. In validity testing, in contrast to prior work [20], *individual differences in the baseline speed of responding and processing speed were controlled for*. We examined data from an EMA study of adults with T1D [29]. EMA surveys (which included emotion items) were completed 5 to 6 times a day (depending on the participants' sleep schedule) for 2 weeks, and the RTs for each item were recorded. Notably, RT-based metrics can vary within people, and their study-long averages can vary across people. Multilevel modeling was used to account for both these potential sources of variance.

Emotional clarity may be of particular relevance for adults with T1D. Compared to the general population, adults with T1D may be exposed to stressors more frequently, specifically emotional distress related to the daily care of diabetes. A prior study estimated that performing all recommended diabetes self-management activities would require >2 hours daily [30]. Given such burden and associated diabetes distress (emotional distress specific to the daily care of diabetes), individuals with diabetes are more likely to experience lower subjective well-being, such as more symptoms of depression, compared to healthy populations [31]. Diabetes distress has been found to be negatively associated with emotion regulation ability [32]. Therefore, emotional clarity, a possible precursor to emotion regulation [10], may be integral for adults with T1D to cope with the burden of the condition. Poor coping with diabetes can lead to the neglect of diabetes self-management behaviors, which can amplify the health consequences of diabetes. Greater diabetes distress has been associated with lower adherence to insulin regimens [33,34], while greater depressive symptoms have been associated with lower adherence to diet, exercise, and glucose testing recommendations [35].

Reliability

The reliability of the proposed measures for emotional clarity was investigated by examining the test-retest stability of the measures across EMA measurement occasions.

Validity

The validity of the RT-based indicators of emotional clarity was tested by examining their associations with well-validated measures of relevance to emotional clarity. In forming our hypotheses (summarized in Table 1), we made a distinction between the clarity of PA and the clarity of NA because of prior research suggesting that the latter had associations with mental health while the former did not [9]. Therefore, we speculated that the awareness of NA is a more direct precursor to the application of coping strategies and successful coping than the awareness of PA.

Table 1. Hypothesized associations of the negative affect (NA) and positive affect (PA) clarity indicators with subjective well-being and emotion regulation.

	Emotional clarity indicator			
	NA drift rate	PA drift rate	RT[a] for NA items[b]	RT for PA items
Subjective well-being				
Satisfaction with life	Positive	Null	Positive	Null
Neuroticism	Negative	Null	Negative	Null
Depression	Negative	Null	Negative	Null
Anxiety	Negative	Null	Negative	Null
Diabetes distress	Negative	Null	Negative	Null
Emotion regulation and its 6 components				
Emotion regulation difficulties (overall)	Negative	Null	Negative	Null
Limited strategies	Negative	Null	Negative	Null
Nonacceptance	Negative	Null	Negative	Null
Impulse control difficulties	Negative	Null	Negative	Null
Difficulties with goal directedness	Negative	Null	Negative	Null
Lack of awareness	Negative	Negative	Negative	Negative
Lack of emotional clarity	Negative	Negative	Negative	Negative

[a]RT: response time.

[b]Multiplied by –1 so that higher values indicate greater emotional clarity.

Hypothesized Associations With Subjective Well-Being

Past literature has found that increased emotional clarity and emotion regulation ability, as assessed by questionnaires or RTs to emotion items, were associated with greater subjective well-being, including greater life satisfaction [36], lower neuroticism [9], lower depression [18], fewer anxiety symptoms [8], and less diabetes distress [32]. However, following the results of the study by Thompson et al [9], we hypothesized that these relationships should hold only for indicators of NA clarity, not PA clarity.

Hypothesized Associations With Emotion Regulation

We considered the valence of emotions in the generation of our hypotheses regarding associations between passively collected indicators of emotional clarity and emotion regulation, given previous work showing that the clarity of positive emotions had different associations with other measures (eg, neuroticism and depression) compared to the clarity of negative emotions [9]. The DERS [15] has 6 subscales representative of emotion regulation components (listed in Table 1), as well as an overall score. Of the 6 DERS subscales, 4 (67%) specifically assess problems with regulating NA (ie, limited access to emotion regulation strategies when upset, nonacceptance of negative emotional responses, impulse control difficulties when upset, and difficulties with goal-directed behaviors when upset); we hypothesized that greater difficulties indicated on each of these component scales would be associated with lower NA clarity on the proposed RT-based measures [10], whereas we did not expect them to be associated with PA clarity on the RT-based measures (Table 1). That is, we expected NA clarity to precede and hence be more relevant to NA regulation than PA clarity, consistent with prior findings of associations between neuroticism and NA clarity, but not PA clarity [9]. The remaining 2 (33%) components of emotion regulation (ie, a lack of emotional awareness and lack of emotional clarity) are not specific to the regulation of NA [15]; therefore, we hypothesized that they would be associated with the clarity of both NA and

PA using the RT-based measures. Notably, 1 of these 2 components assesses self-reported emotional clarity; we expected that this component would have associations of a greater magnitude with RT-based measures of NA and PA clarity compared to other emotion regulation subscales. Finally, we hypothesized an association between the DERS total score and indicators of NA clarity but not indicators of PA clarity because 4 (67%) of the 6 DERS subscales were relevant to emotion regulation when experiencing NA.

Adjustment by Individual Differences in RT and Processing Speed

We examined whether adjusting the RT-based emotional clarity indicators by individual differences in RT and participant processing speed would affect the results of the convergent or divergent validity tests. Both RTs to survey items and drift rate were expected to not be purely indicators of emotional clarity but likely also be impacted by processing speed [19,20,37] and differences in the baseline speed of responding [12]. Therefore, we examined the robustness of the results of validity testing when statistically adjusting for both these variables.

Methods

Overview

The analyzed data were from an EMA study focused on investigating the relationships among momentary blood glucose level, emotional state, and functioning in adults with T1D [29]. Participants were recruited from 3 clinical sites through health care provider referrals, mailings, flyers, and emails. Inclusion criteria were having a diagnosis of T1D, being able to speak and read English or Spanish, and the ability and willingness to carry out study procedures (eg, completion of EMAs and cognitive tests on smartphones). Study procedures included the completion of baseline surveys, training in the use of study devices, 2 weeks of 5 to 6 EMAs and ambulatory cognitive tests daily, and follow-up surveys. EMA surveys began at participants' selected wake-up time each day and were administered at 3-hour intervals after that until sleep time. If a participant reported that they would likely be sleeping by the time of the sixth survey (ie, 15 hours after the first survey), then they were given the option to complete 5 surveys daily instead of 6. To encourage EMA compliance, 3 brief check-in emails or calls were scheduled with study staff. In addition, if EMA survey completion was low, then study staff would contact the participant to query if any support was required. The study procedures are described in greater depth in a protocol paper [29].

Ethical Considerations

All study procedures were approved by the University of Southern California's Institutional Review Board (proposal #HS-18-01014). Informed consent was provided before study participation. Participants were compensated US $200 via a reloadable debit card for completion of all study procedures. Study data were deidentified before analysis.

Measures

RT-Based Measures of Emotional Clarity

The clarity of PA was assessed with RTs to EMA items about how happy, content, enthusiastic, or excited participants felt "right now," while the clarity of NA was assessed with RTs to EMA items asking how tense, upset, sad, or disappointed participants felt in the moment. These emotion adjectives were taken from the "Stress and Working Memory" study [38] and were chosen because they mapped neatly onto the circumplex model of affect [39]. That is, there were 2 items in each of the 4 circumplex dimensions (ie, unpleasant and activated, unpleasant and deactivated, pleasant and activated, and pleasant and deactivated), thereby ensuring that a range of emotion types was represented. The responses were all given on slider scales from 0 (not at all) to 100 (extremely). These emotion items were administered at every EMA prompt using the mobile EMA app (ilumivu [40]). Items were presented one at a time on study-provided smartphones. For each item, RTs were recorded in milliseconds. RTs that were deemed too fast (ie, <0.2 seconds) or too slow (ie, >30 seconds) were set to missing for analyses (5979/461,896, 1.29% of the observations) because such outliers could be indicative of careless responding or distractions during survey completion [20,41].

Emotional Clarity Measure 1: Median RTs

NA and PA clarity were computed as the median RT of the 4 NA items and the 4 PA items at each EMA prompt [12,17] (RTs were multiplied by –1 such that higher values indicate greater clarity). In this paper, median RT for NA items will be referred to as NA RT and median RT for PA items as PA RT.

Following prior research, the median RTs were adjusted for the baseline speed of responding to partial out individual differences in RT stemming from factors such as reading and screen tapping speed [12,17]. We assessed participants' baseline speed by taking the median RT across all EMA occasions on a multiple-choice question asking what they were doing immediately before the survey.

Emotional Clarity Measure 2: Drift Rate From the Drift-Diffusion Model

Drift rates for both NA and PA items were calculated for each EMA following procedures described in detail in a prior study [20] with software code available [42]. In brief, we estimated the drift rates for each person and EMA measurement occasion using the IRT based-variant of the drift-diffusion model that was specifically developed for use with self-report (eg, EMA) items. Because the IRT model requires binary variables, responses to PA and NA items were converted into dichotomous variables such that responses below the midpoint of the scale were coded as 0, while responses at or above the midpoint of the scale were coded as 1. Previous analyses have shown that drift rate measures derived from continuous items that were dichotomized demonstrate convergent validity with items that were already presented to respondents in a binary response format [20]. Next, we examined whether the dichotomized PA and NA item sets were unidimensional, a condition necessary for the calculation of drift rate; we conducted a confirmatory factor analysis in Mplus (version 8.8; Muthén & Muthén) [43] using the weighted least square mean and variance adjusted estimator, using cluster-robust SEs to account for the nesting of multiple EMA measurement occasions within individuals. We examined whether fit indices were within the traditional ranges for acceptable model fit, including a root mean square error of approximation (RMSEA) of at least <0.08, comparative fit index (CFI) and Tucker-Lewis Index (TLI) of at least >0.90, and standardized root mean square residual (SRMR) of <0.08 [44]. Within and between-person reliability (McDonald omega) coefficients were also computed for the dichotomized and nondichotomized NA and PA items [45].

A drift-diffusion model was then applied to the RTs and dichotomized response values for PA and NA items using the *diffIRT* package in R (R Foundation for Statistical Computing) [22], where drift rate and RTs were modeled as latent factors. Factor score estimates for the drift rate parameter were then calculated for each EMA occasion, separately for NA and PA. To examine the fit of the drift-diffusion model on the PA and NA item sets, we examined the level of consistency between the observed and diffusion model–predicted RT distributions with histograms and density plots.

The drift rate parameters used in our primary analyses are a processed version of the drift rate parameters from the diffusion IRT models referred to as the *absolute drift rate*, computed as the mean absolute difference between the drift rate parameter factor scores and the item difficulty levels. The drift rate parameter is an estimation of a person's tendency to report high NA (or PA) in a moment after taking into account both a person's responses and item RTs. However, emotional clarity should be indicated by the speed in carefully accessing one's mood regardless of its valence (eg, high or low NA). Therefore, following the distance-difficulty hypothesis-informed formula for the speed of information accumulation in the D-diffusion model [22], we found the absolute value of the difference between the drift rate parameter and average item difficulty and operationalized this absolute drift rate as emotional clarity. These absolute drift rates were then log transformed to normalize their distributions. When used in analyses, drift rates were also adjusted by baseline speed.

Other Measures

The measures used for convergent or divergent validity testing were completed before ("baseline") or after ("follow-up") the EMA study period. Life satisfaction was assessed with the Satisfaction with Life Scale [46], neuroticism with the Ten-Item Personality Inventory [47], depression with the Patient Health Questionnaire [48], anxiety with the Generalized Anxiety Disorder Scale [49], diabetes distress with the Problem Areas in Diabetes Scale short form [50], and emotion regulation with the DERS short form (DERS-SF) [15]. The Ten-Item Personality Inventory, Patient Health Questionnaire, Generalized Anxiety Disorder Scale, and DERS-SF were administered at baseline, while the Satisfaction with Life Scale and Problem Areas in Diabetes Scale short form were completed at follow-up [29].

Processing speed was assessed with the Symbol Search task [51], an ambulatory cognitive test administered as part of every EMA prompt [29]. The Symbol Search task captured perceptual speed, which is a component of processing speed [51,52]. Participants were presented with 2 cards at the top and 2 cards at the bottom of the phone screen, each with 2 symbols. As quickly as they could, they were asked to choose the card at the bottom of the screen that matched with one of the cards on top. The task consisted of 20 trials, and processing speed was measured as the median RT in accurate trials, only for sessions with at least 70% (14/20) matching accuracy [51]. Symbol Search RTs were calculated such that higher values indicate faster processing speed.

Statistical Analyses

Reliability

Reliability was assessed for each of the emotional clarity indicators: NA drift rate, PA drift rate, NA RT, and PA RT. It was calculated using the following formula: between-person reliability=variance (between-person)/(variance [between-person] + variance [within-person]/n) [53], where between-person variance is the variance in the average of scores across measurement occasions (ie, EMA prompts), within-person variance is the variance of scores across measurement occasions within a person, and n is the number of measurement occasions. Between and within-person variance were calculated with multilevel models, with EMA prompts nested in individuals, where the measure of interest (eg, the NA drift rate) was specified at both levels 1 and 2. To examine how many measurement occasions would be needed to obtain acceptable reliability (≥0.7) for each emotional clarity indicator, we estimated how reliability changed as a function of the number of measurement occasions, moving from 2 EMA prompts to a maximum of 70 prompts. Mplus (version 8.10) [43] was used for reliability analyses via the package *MplusAutomation* [54] in the statistical software R (The R Foundation) [55].

Validity

Validity testing was performed for average NA drift rate, PA drift rate, NA RT, and PA RT. To account for the nested data structure, with multiple EMA prompts nested in individuals, we estimated correlation coefficients between the emotional clarity indicators and other measures using multilevel structural equation models (MSEMs). Multilevel variables with both within- and between-person variance (NA drift rate, PA drift rate, NA RT, PA RT, and processing speed) were specified at both level 1 (within-person) and level 2 (between-person), with latent means estimated at level 2 of the MSEM. As all cross-sectional measures only contained between-person variance, they were entered into the MSEM at level 2 and allowed to correlate with the NA drift rate, PA drift rate, NA RT, and PA RT variables. To adjust for individual differences in baseline speed, NA drift rate, PA drift rate, NA RT, and PA RT were regressed on baseline speed at level 2 of the MSEM. In addition, covariances were specified between baseline speed and all other level-2 variables. Prior research indicated that RTs to tasks were affected by the time of day [19,56]. Therefore, at level 1, all the RT-based metrics were adjusted for (ie, regressed on) the time of day, coded as a categorical variable, where a participant's first survey per day was categorized as taking place in the morning, their final scheduled survey of the day as taking place in the evening, and all their surveys in between as taking place midday. The reference group was “midday,” meaning that at level 2, the latent means for the RT-based metrics were for midday surveys.

Various sensitivity tests were also conducted. Separate multilevel regression models explored whether adjustment for emotional intensity impacted the relationships between RT and each cross-sectional measure, such that NA RT and PA RT were additionally regressed on linear and quadratic terms of overall NA and PA ratings (ie, average rating across NA or PA items within an EMA survey), respectively. In supplemental analyses in which all the proposed emotional clarity indicators were controlled for processing speed, they were all regressed on processing speed at level 2. Finally, we tested whether the association between NA and PA drift rate and other measures would differ if the drift metrics were computed from mood items that were dichotomized at each person’s mean NA and PA response rather than the scale midpoint. For all analyses, data from all participants were included regardless of completion rates because MSEMs estimate latent averages of variables at level 1 that account for potential unreliability stemming from sparse participant data [57]. All validity analyses were conducted in M*plus* (version 8.10) [43] using maximum likelihood with robust SEs. Code for both the reliability and validity analyses is provided [58].

Results

Descriptive Statistics

Characteristics of the study sample, 196 adults with T1D, are shown in Table 2. The median EMA completion rate over the 2-week study period was 92% (IQR 11%).

Descriptive statistics for the EMA variables are presented in Table 3. Distributions for RTs to individual NA and PA items are shown in Figures S1 and S2 in Multimedia Appendix 1, respectively. Tables S1 and S2 in Multimedia Appendix 1 show the between-person correlations between (unadjusted) study measures.

A unidimensional model was found to fit both the (dichotomized) 4 NA and 4 PA items acceptably, justifying the calculation of drift rates for both types of items. For NA, χ^2_2=10.6, *P*=.005; CFI=0.998; TLI=0.994; RMSEA=0.018; and SRMR=0.018, while for PA, χ^2_2=53.3, *P*<.001; CFI=0.989; TLI=0.966; RMSEA=0.044; and SRMR=0.031. All these values were within commonly suggested ranges for acceptable model fit [44]. The within-person omega estimate for the 4 dichotomized NA items was 0.710, while the between-person omega estimate was 0.938. For the 4 dichotomized PA items, the within-person omega estimate was 0.674, and the between-person omega estimate was 0.939. The within-person omega estimate for the 4 nondichotomized NA items was 0.793, while the between-person omega estimate was 0.955. For the 4 nondichotomized PA items, the within-person omega estimate was 0.773, and the between-person omega estimate was 0.937. For all PA and NA items, the observed RT distributions (histograms) were consistent with the drift-diffusion model–predicted RT distributions as per density plots of these RTs (Figures S1 and S2 in Multimedia Appendix 1), indicating a good fit overall for the D-diffusion IRT model [22].

Table 2. Participant characteristics (N=196).

Characteristic	Values
Age (y), mean (SD; range)	39.6 (14.3; 18-75)
Sex n (%)	
Male	88 (44.9)
Female	108 (55.1)
Other	0 (0)
Ethnicity, n (%)	
African American	29 (14.8)
Asian	7 (3.6)
Latino	80 (40.8)
White	56 (28.6)
Multiethnic	14 (7.1)
Other	6 (3.1)
Not reported	4 (2)
Employment status, n (%)	
Full time	69 (35.2)
Part time	23 (11.7)
Full-time homemaker	9 (4.6)
Student	18 (9.2)
Unemployed	27 (13.8)
Retired	15 (7.7)
Disabled	23 (11.7)
Other	8 (4.1)
Not reported	4 (2)
Annual household income (US $), n (%)	
<25,000	47 (24)
25,000-49,999	43 (21.9)
50,000-74,999	15 (7.7)
≥75,000	40 (20.4)
Not provided	51 (26)
Mental health[a]	
SWLS[b] score, mean (SD; range)	22.0 (7.5; 5-35)
TIPI[c] neuroticism score, mean (SD; range)	3.2 (1.3; 1-7)
PHQ[d] score, mean (SD; range)	5.4 (4.3; 0-24)
PHQ score≥9 (moderate or higher depression), n (%)	30 (15.3)
GAD[e] score, mean (SD; range)	4.6 (3.8; 0-21)
GAD score≥9 (moderate or higher anxiety), n (%)	19 (9.7)
PAID[f] score, mean (SD; range)	8.0 (5.5; 0-20)
DERS[g] score, mean (SD; range)	1.9 (0.6; 1-5)

[a]Possible score ranges for the surveys were listed. Only for the Generalized Anxiety Disorder Scale, the observed score range (0-19) was different from the possible range.

[b]SWLS: Satisfaction with Life Scale.

[c]TIPI: Ten-Item Personality Inventory.

[d]PHQ: Patient Health Questionnaire.

[e]GAD: Generalized Anxiety Disorder Scale.

[f]PAID: Problem Areas in Diabetes Scale assessing diabetes distress.

[g]DERS: Difficulties in Emotion Regulation Scale.

Table 3. Summary statistics for ecological momentary assessment (EMA) variables.

EMA variable	All observations, mean (SD; range)	Between-person variance	Average within-person variance	ICC[a]	Between-person variance of log within-person variances
NA[b] drift rate	–0.05 (0.84; –2.43 to 2.31)	0.19	0.54	0.27	0.27
PA[c] drift rate	–0.53 (0.83; –1.68 to 2.05)	0.19	0.51	0.28	0.18
NA RT[d]	2.22 (1.70; 0.31 to 24.20)	0.70	1.95	0.26	1.00
PA RT	2.39 (1.83; 0.22 to 28.24)	0.84	2.02	0.29	0.89
Sum of 4 NA items	19.69 (20.26; 0 to 100)	247.10	243.15	0.50	1.60
Sum of 4 PA items	49.78 (23.87; 0 to 100)	338.85	327.01	0.51	1.40
Sum of dichotomized NA items	0.71 (1.21; 0 to 4)	0.75	1.95	0.28	3.27
Sum of dichotomized PA items	2.43 (1.50; 0 to 4)	1.05	1.88	0.36	1.64

[a]ICC: intraclass correlation coefficient.

[b]NA: negative affect.

[c]PA: positive affect.

[d]RT: response time.

Reliability

Figure S3 in Multimedia Appendix 1 shows the between-person reliabilities as a function of the number of EMA prompts completed for each proposed emotional clarity indicator. For all indicators, reliability increased with more prompts, but each of the proposed emotional clarity indicators demonstrated acceptable reliability (ie, ≥0.70) with a relatively small number of EMA occasions. NA RT required 5 EMA prompts for acceptable reliability, and PA RT required 4 prompts. Both average NA and PA drift rates showed 0.70 reliability when 7 EMA prompts were completed.

Validity

Relationships between NA RT and other measures were not consistent with our hypotheses. No significant associations were found with subjective well-being or emotion regulation measures (Table 4; *P* values ranging from .08 to .98). PA RT was not significantly associated with any well-being or emotion regulation variable (Table 4; *P* values ranging from 25 to 93), except, unexpectedly, for diabetes distress (r=0.17; P=.009). After adjustment for processing speed, the associations between both NA RT and PA RT and diabetes distress were unexpectedly significant in a positive direction (r=0.14; P=.03 and r=0.20; P=.002, respectively). Adjustment for emotional intensity changed effects very minimally at the between-person level, consistent with the findings of a prior study [17], so the results from that model were not reported here. Both NA RT and PA RT had significant associations with processing speed (r=0.25; P<.001 and r=0.21; P=.001, respectively).

The associations between NA drift rate and other measures were consistent with our hypotheses overall. Unexpectedly, NA drift rate was not related to the lack of emotional clarity (P=.13) or 3 of the other DERS subscales (Table 4; *P* values from .18 to .76). It was also not associated with life satisfaction (P=.17). However, a greater NA drift rate was significantly associated with lower neuroticism (r=–0.18; P=.01), depression (r=–0.27; P<.001), anxiety (r=–0.27; P<.001), and diabetes distress (r=–0.17; P=.005); lower overall difficulties with emotion regulation (total DERS score; r=–0.15; P=.03); less limited emotion regulation strategies (r=–0.15; P=.048); and lower lack of emotion awareness (r=–0.18; P=.009). Adjustment by processing speed resulted in minimal changes. When NA and PA drift rates were computed from mood items that were dichotomized at each person's mean NA and PA ratings (rather than at the scale midpoint), the results showed mostly nonsignificant correlations with other measures (refer to Table S3 in Multimedia Appendix 1).

Table 4. Between-person correlations between emotional clarity indices and other measures. All correlations were adjusted for the baseline speed of responding and time of day, but only columns with adjusted values additionally had processing speed as a control variable.

	NA[a] drift rate	NA drift rate, adjusted[b]	PA[c] drift rate	PA drift rate, adjusted[b]	NA RT[d,e]	NA RT, adjusted[b,e]	PA RT[e]	PA RT, adjusted[b,e]
Subjective well-being								
Satisfaction with life								
r	0.1	0.1	–0.1	–0.1	–0.03	–0.04	–0.04	–0.05
P value	.17	.18	.09	.09	.67	.57	.51	.45
Neuroticism								
r	–0.18	–0.18	–0.08	–0.08	–0.06	–0.06	–0.01	–0.01
P value	.01[f]	.01[f]	.29	.29	.51	.44	.91	.85
Depression								
r	–0.27	–0.27	0.03	0.03	0	0	0.06	0.06
P value	<.001[f]	<.001[f]	.73	.73	.98	.99	.47	.45
Anxiety								
r	–0.27	–0.27	–0.09	–0.09	–0.09	–0.09	–0.03	–0.03
P value	<.001[f]	<.001[f]	.21	.22	.34	.28	.77	.74
Diabetes distress								
r	–0.17	–0.16	0.17	0.17	0.09	0.14	0.17	0.2
P value	.005[f]	.008[f]	.009[f]	.008[f]	.14	.03[f]	.009[f]	.002[f]
Emotion regulation								
DERS[g] (total)								
r	–0.15	–0.16	0.03	0.03	0.03	0.01	0.05	0.03
P value	.03[f]	.01[f]	.78	.78	.64	.92	.39	.58
Limited strategies								
r	–0.15	–0.15	0.03	0.03	–0.04	–0.04	–0.03	–0.03
P value	.048[f]	.04[f]	.76	.76	.59	.59	.66	.67
Nonacceptance								
r	–0.1	–0.11	0.04	0.03	–0.01	–0.03	0	–0.01
P value	.18	.13	.64	.66	.87	.61	.93	.85
Impulse control difficulties								
r	0.02	0.03	0.13	0.13	0.02	0.04	0.02	0.05
P value	.76	.65	.18	.18	.79	.39	.67	.33
Difficulties with goal directedness								
r	–0.07	–0.08	–0.11	–0.11	0.11	0.09	0.08	0.06
P value	.32	.26	.20	.18	.08	.13	.25	.36
Lack of awareness								
r	–0.18	–0.18	–0.09	–0.09	–0.07	–0.1	–0.01	–0.03
P value	.009[f]	.005[f]	.20	.18	.35	.19	.84	.64
Lack of emotional clarity								
r	–0.11	–0.12	0.04	0.04	0	–0.03	0.03	0
P value	.13	.09	.59	.59	.95	.64	.73	.95
Processing speed (control variable)								

	NA[a] drift rate	NA drift rate, adjusted[b]	PA[c] drift rate	PA drift rate, adjusted[b]	NA RT[d,e]	NA RT, adjusted[b,e]	PA RT[e]	PA RT, adjusted[b,e]
r	0.09	—[h]	0.02	—	0.25	—	0.21	—
P value	.18	—	.78	—	<.001[f]	—	<.001[f]	—

[a]NA: negative affect.

[b]Adjusted for processing speed.

[c]PA: positive affect.

[d]RT: response time.

[e]Multiplied by –1 so that higher values indicate greater emotional clarity.

[f]*P*<.05.

[g]DERS: Difficulties in Emotion Regulation Scale.

[h]Not applicable.

Consistent with expectations, PA drift rate was not significantly associated with most subjective well-being or NA-specific emotion regulation measures (Table 4; *P* values from .09 to .76). Unexpectedly, PA drift rate was not significantly associated with emotional clarity (*P*=.59) or emotional awareness (*P*=.20) before or after adjustment for processing speed. Furthermore, contrary to expectations, a greater PA drift rate was associated with greater diabetes distress (*r*=0.17; *P*=.009). Neither NA drift rate nor PA drift rate had significant associations with processing speed (*P*=.18 and *P*=.78, respectively). Without adjustment for baseline speed, a few of the associations differed, such as the relationship between NA drift rate and life satisfaction (*r*=0.16; *P*=.02; Table 5).

Table 5. Between-person correlations between emotional clarity indices and other measures. Correlations were adjusted for the time of day but not for the baseline speed of responding, and only columns with adjusted values had processing speed as a control variable.

	NA[a] drift rate	NA drift rate, adjusted[b]	PA[c] drift rate	PA drift rate, adjusted[b]	NA RT[d,e]	NA RT, adjusted[b,e]	PA RT[e]	PA RT, adjusted[b,e]
Subjective well-being								
Satisfaction with life								
r	0.16	0.14	0	–0.03	0.1	0.06	0.09	0.05
P value	.02[f]	.05	.96	.67	.15	.40	.15	.42
Neuroticism								
r	–0.14	–0.16	–0.05	–0.06	–0.01	–0.04	0.02	0
P value	.04[f]	.02[f]	.48	.38	.87	.63	.74	.94
Depression								
r	–0.27	–0.28	–0.01	0	–0.05	–0.04	–0.01	0
P value	<.001[f]	<.001[f]	.92	1.00	.43	.57	.85	.95
Anxiety								
r	–0.25	–0.26	–0.08	–0.09	–0.07	–0.08	–0.03	–0.03
P value	<.001[f]	<.001[f]	.24	.21	.34	.24	.69	.64
Diabetes distress								
r	–0.25	–0.2	0.04	0.1	–0.08	0.03	–0.04	0.07
P value	<.001[f]	.003[f]	.53	.17	.23	.67	.59	.33
Emotion regulation								
DERS[g] (total)								
r	–0.15	–0.17	0	–0.01	–0.01	–0.04	0	–0.02
P value	.045[f]	.009[f]	.97	.92	.91	.51	.95	.70
Limited strategies								
r	–0.12	–0.13	0.03	0.04	–0.02	–0.02	–0.01	–0.01
P value	.12	.07	.70	.67	.81	.78	.90	.90
Nonacceptance								
r	–0.06	–0.09	0.06	0.04	0.04	–0.01	0.05	0
P value	.41	.18	.42	.61	.58	.83	.44	.95
Impulse control difficulties								
r	0	0.03	0.09	0.12	–0.02	0.04	–0.02	0.04
P value	.99	.67	.36	.21	.79	.48	.82	.47
Difficulties with goal directedness								
r	–0.07	–0.09	–0.1	–0.12	0.07	0.04	0.04	0.01
P value	.35	.21	.23	.14	.27	.46	.47	.79
Lack of awareness								
r	–0.19	–0.21	–0.12	–0.13	–0.1	–0.14	–0.06	–0.09
P value	.01[f]	.002[f]	.13	.07	.13	.02[f]	.32	.12
Lack of emotional clarity								
r	–0.14	–0.16	–0.01	–0.02	–0.07	–0.1	–0.05	–0.07
P value	.07	.04[f]	.92	.81	.38	.17	.54	.32
Processing speed (control variable)								

	NA[a] drift rate	NA drift rate, adjusted[b]	PA[c] drift rate	PA drift rate, adjusted[b]	NA RT[d,e]	NA RT, adjusted[b,e]	PA RT[e]	PA RT, adjusted[b,e]
r	0.28	—[h]	0.25	—	0.5	—	0.48	—
P value	<.001[f]	—	.005[f]	—	<.001[f]	—	<.001[f]	—

[a]NA: negative affect.

[b]Adjusted for processing speed.

[c]PA: positive affect.

[d]RT: response time.

[e]Multiplied by –1 so that higher values indicate greater emotional clarity.

[f]*P*<.05.

[g]DERS: Difficulties in Emotion Regulation Scale.

[h]Not applicable.

Discussion

Principal Findings

The most notable finding from this study was that the average NA drift rate, a proposed indicator of typical emotional clarity, had expected associations overall with validated measures of subjective well-being and emotion regulation, both before and after adjustment for processing speed and emotional intensity. By contrast, NA RT, another proposed indicator of typical emotional clarity, did not have the anticipated associations with the validated measures. Relative to NA drift rate, NA RT may be confounded by a greater number of factors aside from emotional clarity.

Nevertheless, there is no sufficient evidence to conclude that the average NA drift rate is a valid indicator of NA clarity, while the average NA RT is not a valid indicator of NA clarity due to sample size constraints. In post hoc power analyses, with our sample size of 196 participants, there was 80% power to detect a between-person correlation of 0.20. Therefore, the study may have been underpowered to detect small correlations that might still be meaningful. Furthermore, the sample size was not chosen a priori to be sufficiently powered to detect between-person relationships after adjustment for multiplicity of testing [59] because the sample size was conditioned on research questions that were more primary for the diabetes EMA study.

While we cannot conclude that NA drift rate is a valid indicator of NA clarity, our results suggest that researchers in future studies should continue to investigate NA drift rate as an implicit measure of emotional clarity. We applied stringent validity tests by adjusting for both individual differences in the baseline speed of responding and processing speed and found significant associations between NA drift rate and 4 (80%) of the 5 subjective well-being ratings. Correlations with 4 (67%) of the 6 emotion regulation subscales were in the expected directions. Finally, it is notable that NA drift rate had a correlation of 0.63 (*P*<.001) with NA RT. People with a higher NA drift rate responded faster to the NA emotion items, had greater subjective well-being (eg, fewer depression symptoms), and had fewer difficulties with overall emotion regulation, which are all aligned with the expectation for an emotional clarity measure. Collectively, study results suggest that NA drift rate deserves further attention in future research.

Some of the magnitudes of correlations between NA drift rate and other study measures were comparable to sizes of correlations between formal (self-report) assessments of emotional clarity and other measures found in prior studies. One study found that in a group with generalized anxiety disorder, a self-report assessment of emotional clarity had correlations of –0.29 and –0.33 with depression and anxiety, respectively [60], which are similar to the correlations found in this study. In the same study, the association between emotional clarity and depression or anxiety was not significant in the healthy control group [60], suggesting that the mental health status of the sample may affect the magnitude of the observed relationships. Other correlations found between emotional clarity and depression were –0.24 in a clinical sample [61] and –0.29 in elementary school–aged children [62]. Correlations between self-reported emotional clarity and neuroticism ranged between –0.31 in a sample of adolescents [63] and –0.37 in a sample of college students [64], and correlations between emotional clarity and life satisfaction ranged between 0.31 in adolescents [65] and 0.35 in undergraduate students [36], which are larger than the correlations observed in this study.

When NA drift rate was computed from mood items that were dichotomized at each person's mean response rather than the scale midpoint, no significant associations were observed with other study measures. The distance-difficulty hypothesis underlying the D-diffusion model that was used to generate the NA drift parameter assumes that people respond faster to items that contrast more with their current state [23]. According to this hypothesis, an individual with very low NA (regardless of their average level of NA) was expected to quickly report not being scared (an item often associated with high NA). Dichotomizing mood at each person's own midpoint created a variable representing whether their mood was higher or lower relative to their personal average and not higher or lower in absolute terms (which could be approximated by dichotomizing at the scale midpoint). Perhaps, the former was less relevant to the distance-difficulty hypothesis compared to the latter because it captured relative mood and not actual mood, leading to the creation of NA drift parameters with no associations with other study measures.

The reliabilities for the average NA drift rate, PA drift rate, NA RT, and PA RT were all acceptable with a small number of EMA prompts (ie, 4 to 7 EMA prompts or 1 to 2 days of EMA surveys). Therefore, reliable measurement of these proposed emotional clarity indicators would likely be feasible in most EMA studies where affect items are administered and RTs are recorded.

Secondary Findings

We found preliminary evidence supporting the argument that emotional clarity deficits are valence specific [9]. NA drift rate and PA drift rate had differential associations with self-report measures. Furthermore, they were moderately correlated with each other (r=0.38; refer to Table S1 in Multimedia Appendix 1). Had NA and PA drift rates been redundant with one another, a high correlation would have been expected.

Greater PA drift rate and PA RT were unexpectedly found to be associated with higher diabetes distress and not associated with the awareness or clarity subscales of the DERS. It is unclear why people with greater diabetes distress would have greater clarity of positive emotions. Perhaps, when overwhelmed with burden from diabetes, people had a greater appreciation of positive emotional states and hence greater clarity of PA. One possible reason why PA drift rate and PA RT had far from significant associations with self-reported emotional clarity may have been because, given that other items in the DERS-SF asked questions relevant to the NA context, participants were primed to answer the emotional clarity questions with reference to feeling NA. More assessments of the validity of PA drift rate and PA RT are needed.

Future Directions

More conclusive evidence of the validity of RT-based measures of emotional clarity may come from studies where NA clarity can be manipulated, and the proposed emotional clarity indicators can be compared for sensitivity to these changes. For instance, people who undergo a mindfulness intervention may be expected to have higher NA clarity in the period following the intervention, and this effect should be reflected in changes in drift rates or RTs to NA EMA items.

EMA mood item RT–based measures of emotional clarity have a great potential utility. They can serve as indices of emotional clarity that do not require burdening participants with additional emotional clarity items. Furthermore, they can help avoid possible issues with subjective reports, including poor insight into emotional clarity and the possibility of social desirability bias [12]. Although the validity of the RT-based emotional clarity measures at the within-person level was not investigated in this study, such validity would allow for the investigation of changes in emotional clarity within and across days and the situational factors that contribute to them. For the potential utility of EMA mood item RT–based measures of emotional clarity to be realized, further investigations of the validity of EMA RT–based measures of emotional clarity (ie, at both the between-person and within-person levels) are needed.

Limitations

We had decided not to adjust for multiple comparisons, but to also acknowledge that any results would require replication by future studies. It has been argued that the need for adjustment for multiple comparisons should be evaluated on a case-by-case basis and P value adjustment should not be used for all analyses [66]. For instance, adjustment for multiple comparisons comes with not only the benefit of lowering type 1 error but also the disadvantage of increasing the chance of type 2 error. Therefore, one factor to consider when deciding whether to adjust for multiple comparisons is the relative cost of type 1 and type 2 errors for a particular research question [67]. In confirmatory studies with results that have implications for changes in clinical practice or the use of a treatment, the cost of a type 1 error may be higher than that of a type 2 error; hence, P value adjustment for multiple comparisons would be sensible [66]. When performing post hoc analyses on existing data as part of theory building and testing (and without direct treatment implications), the relative cost of a type 2 error may be higher; hence, there may be a stronger argument for not using multiple comparisons adjustment [66]. That is, a type 2 error could cause researchers to not detect potentially important findings [68]. If adjustment is not used, there would need to be an acknowledgment that, to account for the possibility of a type 1 error, further research is needed to examine whether results can be replicated [66].

Nevertheless, we still tested the effect of false discovery rate adjustment on the P values of correlations for the different groups of hypotheses that were tested (eg, association between NA drift rate and subjective well-being measures). Adjustments for multiple comparisons are often applied separately for distinct families of hypotheses [69]. Tables S4 and S5 in Multimedia Appendix 1 show the false discovery rate–adjusted P values associated with Tables 4 and 5, respectively. The biggest differences to note were that several of the associations between NA drift rate and emotional regulation measures were no longer significant.

Using the drift-diffusion model had the advantage of reducing the impact of individual differences in response caution (and potential careless responding) from emotion item RTs but had the disadvantage of assuming a 2-choice task (eg, a high vs low NA) underlying people’s emotion ratings. Because the drift-diffusion model required making the continuous PA and NA items dichotomous for data analysis, granular differences in emotional clarity may have been missed with the drift rate parameter.

We were unable to examine the within-person validity of RT-based clarity measures because EMA measures key to such testing (eg, self-reported emotional clarity and mood regulation success) [12] were not administered in this study. Future studies are needed to examine the within-person validity of the drift rate parameter as a within-person indicator of emotional clarity.

The context in which RTs for emotion items were calculated was similar but not identical to that in prior work. For instance, in this study, median RTs for EMA were calculated based on RTs to 4 items. In prior studies, the median RTs for 5 to 8 items were computed [12,17]. In the original paper examining the validity of RTs to emotion items as indicators of emotional clarity, bipolar mood items were used (eg, items with options from “very unhappy” to “very happy”) [12], whereas our study analyzed unipolar mood items. The results of this study may have been impacted to an extent by differences in EMA administration, such as variations in the type of emotion items used.

Additional evidence is needed that study results generalize beyond adults with T1D. Other populations with an increased likelihood of experiencing lower subjective well-being (eg, individuals with various chronic conditions) [70] may be appropriate targets for future emotional clarity studies.

Conclusions

A measure of NA drift rate derived from RTs to momentary NA items had expected associations with validated measures of relevance to emotional clarity, providing initial evidence supporting its validity as an indicator of individual differences in the clarity of negative emotions. The validities of NA RT, PA RT, and PA drift rate were not strongly supported by our results. More studies are needed to investigate the validities of NA and PA drift rate and NA and PA RT with larger sample sizes. The development of passive measures of emotional clarity would help create minimally burdensome measures of emotional clarity that are less vulnerable to possible issues from subjective self-reports, such as poor clarity insight and social desirability bias. Such measures may be useful in investigations relevant to the role of emotional clarity in people's experience of well-being.

Acknowledgments

This work was supported by the National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, under grant number NIH/NIDDK 1R01DK121298-01, and the National Institute of Aging, National Institutes of Health, under grant number NIH/NIA R01 AG068190. The Southern California Clinical and Translational Science Institute (SC CTSI) supported the use of REDCap (Research Electronic Data Capture; Vanderbilt University) in this study (NIH/NCATS UL1TR001855 and UL1TR000130). Additional support was provided by the New York Regional Center for Diabetes Translation Research (NIH/NIDDK P30 DK111022). The Symbol Search cognitive test was developed in a project supported by the National Institute on Aging at the National Institutes of Health (NIH/NIA #1U2CAG060408-01; principal investigator: Martin Sliwinski).

Data Availability

The data set analyzed for this study is available from the corresponding author, RH, on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Plots of between person reliability as a function of number of ecological momentary assessments and unadjusted between person correlations between study variables.
[DOCX File , 807 KB - mental_v11i1e58352_app1.docx]

References

1. Kahneman D, Diener E, Schwarz N. Well-Being: Foundations of Hedonic Psychology. New York, NY: Russell Sage Foundation; 1999.
2. Stone AA, Mackie C. Subjective Well-Being: Measuring Happiness, Suffering, and Other Dimensions of Experience. Washington, DC: National Academies Press; 2013.
3. Tamir M, Chiu CY, Gross JJ. Business or pleasure? Utilitarian versus hedonic considerations in emotion regulation. Emotion 2007 Aug;7(3):546-554. [doi: 10.1037/1528-3542.7.3.546] [Medline: 17683211]
4. Ortner CN, Corno D, Fung TY, Rapinda K. The roles of hedonic and eudaimonic motives in emotion regulation. Pers Individ Differ 2018 Jan;120:209-212. [doi: 10.1016/j.paid.2017.09.006]
5. Coffey E, Berenbaum H, Kerns J. The dimensions of emotional intelligence, alexithymia, and mood awareness: associations with personality and performance on an emotional stroop task. Cogn Emot 2003 Jul 18;17(4):671-679. [doi: 10.1080/02699930302304]
6. Gross JJ. The emerging field of emotion regulation: an integrative review. Rev Gen Psychol 1998 Sep;2(3):271-299. [doi: 10.1037/1089-2680.2.3.271]
7. Gross JJ. Emotion regulation: current status and future prospects. Psychol Inq 2015 Mar 09;26(1):1-26. [doi: 10.1080/1047840X.2014.940781]
8. Sendzik L, Ö Schäfer J, C Samson A, Naumann E, Tuschen-Caffier B. Emotional awareness in depressive and anxiety symptoms in youth: a meta-analytic review. J Youth Adolesc 2017 Apr;46(4):687-700. [doi: 10.1007/s10964-017-0629-0] [Medline: 28101745]
9. Thompson RJ, Kuppens P, Mata J, Jaeggi SM, Buschkuehl M, Jonides J, et al. Emotional clarity as a function of neuroticism and major depressive disorder. Emotion 2015 Oct;15(5):615-624 [FREE Full text] [doi: 10.1037/emo0000067] [Medline: 25844973]

10. Vine V, Aldao A. Impaired emotional clarity and psychopathology: a transdiagnostic deficit with symptom-specific pathways through emotion regulation. J Soc Clin Psychol 2014 Apr;33(4):319-342. [doi: 10.1521/jscp.2014.33.4.319]
11. Park J, Naragon-Gainey K. Is more emotional clarity always better? An examination of curvilinear and moderated associations between emotional clarity and internalising symptoms. Cogn Emot 2020 Mar;34(2):273-287 [FREE Full text] [doi: 10.1080/02699931.2019.1621803] [Medline: 31122138]
12. Lischetzke T, Angelova R, Eid M. Validating an indirect measure of clarity of feelings: evidence from laboratory and naturalistic settings. Psychol Assess 2011 Jun;23(2):447-455. [doi: 10.1037/a0022211] [Medline: 21319904]
13. Eckland NS, Thompson RJ. State emotional clarity is an indicator of fluid emotional intelligence ability. J Intell 2023 Oct 10;11(10):196 [FREE Full text] [doi: 10.3390/jintelligence11100196] [Medline: 37888428]
14. Salovey P, Mayer JD, Goldman SL, Turvey C, Palfai TP. Emotional attention, clarity, and repair: exploring emotional intelligence using the Trait Meta-Mood Scale. Am Psychol Assoc 1995:125-154. [doi: 10.1037/10182-006]
15. Kaufman EA, Xia M, Fosco G, Yaptangco M, Skidmore CR, Crowell SE. The difficulties in emotion regulation scale short form (DERS-SF): validation and replication in adolescent and adult samples. J Psychopathol Behav Assess 2015 Nov 23;38(3):443-455. [doi: 10.1007/s10862-015-9529-3]
16. Lucas RE, Baird BM. Global self-assessment. In: Eid M, Diener E, editors. Handbook of Multimethod Measurement in Psychology. Washington, DC: American Psychological Association; 2006.
17. Arndt C, Lischetzke T, Crayen C, Eid M. The assessment of emotional clarity via response times to emotion items: shedding light on the response process and its relation to emotion regulation strategies. Cogn Emot 2018 May;32(3):530-548. [doi: 10.1080/02699931.2017.1322039] [Medline: 28482749]
18. Blöte AW, Westenberg PM. The temporal association between emotional clarity and depression symptoms in adolescents. J Adolesc 2019 Feb;71:110-118. [doi: 10.1016/j.adolescence.2019.01.005] [Medline: 30685515]
19. Hernandez R, Hoogendoorn C, Gonzalez JS, Jin H, Pyatak EA, Spruijt-Metz D, et al. Reliability and validity of noncognitive ecological momentary assessment survey response times as an indicator of cognitive processing speed in people's natural environment: intensive longitudinal study. JMIR Mhealth Uhealth 2023 May 30;11:e45203 [FREE Full text] [doi: 10.2196/45203] [Medline: 37252787]
20. Schneider S, Hernandez R, Junghaenel DU, Orriens B, Lee PJ, Stone AA. Response times in Ecological Momentary Assessment (EMA): shedding light on the response process with a drift diffusion model. Curr Psychol 2023 May 27;43(7):5868-5886. [doi: 10.1007/s12144-023-04773-0]
21. Ratcliff R, McKoon G. The diffusion decision model: theory and data for two-choice decision tasks. Neural Comput 2008 Apr;20(4):873-922 [FREE Full text] [doi: 10.1162/neco.2008.12-06-420] [Medline: 18085991]
22. Molenaar D, Tuerlinckx F, Maas HL. Fitting diffusion item response theory models for responses and response times using the R package diffIRT. J Stat Softw 2015;66(4):1-34. [doi: 10.18637/jss.v066.i04]
23. Ferrando PJ. Person-item distance and response time: an empirical study in personality measurement. Psicológica 2006 Jan;27(1):137-148 [FREE Full text]
24. Medvedev ON, Roemer A, Krägeloh CU, Sandham MH, Siegert RJ. Enhancing the precision of the Positive and Negative Affect Schedule (PANAS) using Rasch analysis. Curr Psychol 2021 Mar 05;42(2):1554-1563. [doi: 10.1007/s12144-021-01556-3]
25. Mascarenhas Fonseca L, Strong RW, Singh S, Bulger JD, Cleveland M, Grinspoon E, et al. Glycemic variability and fluctuations in cognitive status in adults with type 1 diabetes (GluCog): observational study using ecological momentary assessment of cognition. JMIR Diabetes 2023 Jan 05;8:e39750 [FREE Full text] [doi: 10.2196/39750] [Medline: 36602848]
26. Kwasnicka D, Kale D, Schneider V, Keller J, Yeboah-Asiamah Asare B, Powell D, et al. Systematic review of ecological momentary assessment (EMA) studies of five public health-related behaviours: review protocol. BMJ Open 2021 Jul 16;11(7):e046435 [FREE Full text] [doi: 10.1136/bmjopen-2020-046435] [Medline: 34272218]
27. Wrzus C, Neubauer AB. Ecological momentary assessment: a meta-analysis on designs, samples, and compliance across research fields. Assessment 2023 Apr;30(3):825-846 [FREE Full text] [doi: 10.1177/10731911211067538] [Medline: 35016567]
28. Scott SB, Sliwinski MJ, Zawadzki M, Stawski RS, Kim J, Marcusson-Clavertz D, et al. A coordinated analysis of variance in affect in daily life. Assessment 2020 Dec;27(8):1683-1698 [FREE Full text] [doi: 10.1177/1073191118799460] [Medline: 30198310]
29. Pyatak EA, Hernandez R, Pham LT, Mehdiyeva K, Schneider S, Peters A, et al. Function and emotion in everyday life with type 1 diabetes (FEEL-T1D): protocol for a fully remote intensive longitudinal study. JMIR Res Protoc 2021 Oct 22;10(10):e30901 [FREE Full text] [doi: 10.2196/30901] [Medline: 34463626]
30. Russell LB, Suh DC, Safford MA. Time requirements for diabetes self-management: too much for many? J Fam Pract 2005 Jan;54(1):52-56. [Medline: 15623407]
31. Gendelman N, Snell-Bergeon JK, McFann K, Kinney G, Paul Wadwa R, Bishop F, et al. Prevalence and correlates of depression in individuals with and without type 1 diabetes. Diabetes Care 2009 Apr;32(4):575-579 [FREE Full text] [doi: 10.2337/dc08-1835] [Medline: 19171719]

32. Coccaro EF, Lazarus S, Joseph J, Wyne K, Drossos T, Phillipson L, et al. Emotional regulation and diabetes distress in adults with type 1 and type 2 diabetes. Diabetes Care 2021 Jan;44(1):20-25 [FREE Full text] [doi: 10.2337/dc20-1059] [Medline: 33444157]
33. Kumar N, Unnikrishnan B, Thapar R, Mithra P, Kulkarni V, Holla R, et al. Distress and its effect on adherence to antidiabetic medications among type 2 diabetes patients in coastal South India. J Nat Sci Biol Med 2017;8(2):216-220 [FREE Full text] [doi: 10.4103/0976-9668.210008] [Medline: 28781491]
34. Halepian L, Saleh MB, Hallit S, Khabbaz LR. Adherence to insulin, emotional distress, and trust in physician among patients with diabetes: a cross-sectional study. Diabetes Ther 2018 Apr;9(2):713-726 [FREE Full text] [doi: 10.1007/s13300-018-0389-1] [Medline: 29520741]
35. Aikens JE. Prospective associations between emotional distress and poor outcomes in type 2 diabetes. Diabetes Care 2012 Dec;35(12):2472-2478 [FREE Full text] [doi: 10.2337/dc12-0181] [Medline: 23033244]
36. Extremera N, Fernández-Berrocal P. Perceived emotional intelligence and life satisfaction: predictive and incremental validity using the Trait Meta-Mood Scale. Pers Individ Differ 2005 Oct;39(5):937-948. [doi: 10.1016/j.paid.2005.03.012]
37. Roque N, Sliwinski M, Ram N. Questionnaire-based everyday reaction time: reliable, valid, and unobtrusive measure of cognition. Innov Aging 2020 Dec 16;4(Supplement_1):599. [doi: 10.1093/geroni/igaa057.2015]
38. Scott SB, Ram N, Smyth JM, Almeida DM, Sliwinski MJ. Age differences in negative emotional responses to daily stressors depend on time since event. Dev Psychol 2017 Jan;53(1):177-190 [FREE Full text] [doi: 10.1037/dev0000257] [Medline: 28026195]
39. Posner J, Russell JA, Peterson BS. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev Psychopathol 2005;17(3):715-734 [FREE Full text] [doi: 10.1017/S0954579405050340] [Medline: 16262989]
40. ilumivu homepage. ilumivu. URL: https://ilumivu.com/ [accessed 2024-07-02]
41. Jaso BA, Kraus NI, Heller AS. Identification of careless responding in ecological momentary assessment research: from posthoc analyses to real-time data monitoring. Psychol Methods 2022 Dec;27(6):958-981. [doi: 10.1037/met0000312] [Medline: 34582244]
42. Drift diffusion model on EMA items. OSF Home. 2022 Aug 1. URL: https://osf.io/r82hn/ [accessed 2024-07-02]
43. Muthén LK, Muthén B. Mplus User's Guide. Los Angeles, CA: Muthén & Muthén; 2017.
44. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Modeling Multidiscip J 1999;6(1):1-55. [doi: 10.1080/10705519909540118]
45. Geldhof GJ, Preacher KJ, Zyphur MJ. Reliability estimation in a multilevel confirmatory factor analysis framework. Psychol Methods 2014 Mar;19(1):72-91. [doi: 10.1037/a0032138] [Medline: 23646988]
46. Diener E, Emmons RA, Larsen RJ, Griffin S. The satisfaction with life scale. J Pers Assess 1985 Feb;49(1):71-75. [doi: 10.1207/s15327752jpa4901_13] [Medline: 16367493]
47. Gosling SD, Rentfrow PJ, Swann WBJ. A very brief measure of the Big-Five personality domains. J Res Personal 2003 Dec;37(6):504-528. [doi: 10.1016/S0092-6566(03)00046-1]
48. Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
49. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
50. McGuire BE, Morrison TG, Hermanns N, Skovlund S, Eldrup E, Gagliardino J, et al. Short-form measures of diabetes-related emotional distress: the Problem Areas in Diabetes Scale (PAID)-5 and PAID-1. Diabetologia 2010 Jan;53(1):66-69. [doi: 10.1007/s00125-009-1559-5] [Medline: 19841892]
51. Sliwinski MJ, Mogle JA, Hyun J, Munoz E, Smyth JM, Lipton RB. Reliability and validity of ambulatory cognitive assessments. Assessment 2018 Jan;25(1):14-30 [FREE Full text] [doi: 10.1177/1073191116643164] [Medline: 27084835]
52. Salthouse TA. Aging and measures of processing speed. Biol Psychol 2000 Oct;54(1-3):35-54. [doi: 10.1016/s0301-0511(00)00052-1] [Medline: 11035219]
53. Raykov T, Marcoulides GA. On multilevel model reliability estimation from the perspective of structural equation modeling. Struct Equ Modeling Multidiscip J 2006;13(1):130-141. [doi: 10.1207/s15328007sem1301_7]
54. Hallquist MN, Wiley JF. MplusAutomation: an R package for facilitating large-scale latent variable analyses in Mplus. Struct Equ Modeling 2018;25(4):621-638 [FREE Full text] [doi: 10.1080/10705511.2017.1402334] [Medline: 30083048]
55. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2019. URL: https://www.R-project.org/ [accessed 2021-01-14]
56. Hawks ZW, Strong R, Jung L, Beck ED, Passell EJ, Grinspoon E, et al. Accurate prediction of momentary cognition from intensive longitudinal data. Biol Psychiatry Cogn Neurosci Neuroimaging 2023 Aug;8(8):841-851. [doi: 10.1016/j.bpsc.2022.12.002] [Medline: 36922302]
57. Asparouhov T, Muthén B. Latent variable centering of predictors and mediators in multilevel and time-series models. Struct Equ Modeling Multidiscip J 2018 Sep 28;26(1):119-142. [doi: 10.1080/10705511.2018.1511375]
58. Examining the reliability and validity of ecological momentary assessment response time based measures of emotional clarity. OSF Home. 2023 Nov 9. URL: https://osf.io/gc7ez/ [accessed 2024-07-02]

59. Leon AC. Multiplicity-adjusted sample size requirements: a strategy to maintain statistical power with Bonferroni adjustments. J Clin Psychiatry 2004 Nov;65(11):1511-1514. [Medline: 15554764]
60. Lizeretti NP, Extremera N. Emotional intelligence and clinical symptoms in outpatients with generalized anxiety disorder (GAD). Psychiatr Q 2011 Sep;82(3):253-260. [doi: 10.1007/s11126-011-9167-1] [Medline: 21279686]
61. Vine V, Marroquín B. Affect intensity moderates the association of emotional clarity with emotion regulation and depressive symptoms in unselected and treatment-seeking samples. Cognit Ther Res 2018 Feb;42(1):1-15 [FREE Full text] [doi: 10.1007/s10608-017-9870-9] [Medline: 29657347]
62. Flynn M, Rudolph KD. A prospective examination of emotional clarity, stress responses, and depressive symptoms during early adolescence. J Early Adolesc 2014;34(7):923-939 [FREE Full text] [doi: 10.1177/0272431613513959] [Medline: 28936023]
63. Oropesa Ruiz NF, Mercader Rubio I, Gutiérrez Ángel N, Pérez García MA. Neuroticism and emotional intelligence in adolescence: a mediation model moderate by negative affect and self-esteem. Behav Sci (Basel) 2022 Jul 19;12(7):241 [FREE Full text] [doi: 10.3390/bs12070241] [Medline: 35877311]
64. Abeyta AA, Routledge C, Juhl J, Robinson MD. Finding meaning through emotional understanding: emotional clarity predicts meaning in life and adjustment to existential threat. Motiv Emot 2015 Jun 4;39(6):973-983. [doi: 10.1007/s11031-015-9500-3]
65. Ramos-Díaz E, Rodríguez-Fernández A, Axpe I, Ferrara M. Perceived emotional intelligence and life satisfaction among adolescent students: the mediating role of resilience. J Happiness Stud 2018 Nov 24;20(8):2489-2506. [doi: 10.1007/s10902-018-0058-0]
66. Althouse AD. Adjust for multiple comparisons? It's not that simple. Ann Thorac Surg 2016 May;101(5):1644-1645. [doi: 10.1016/j.athoracsur.2015.11.024] [Medline: 27106412]
67. Rothman KJ. Six persistent research misconceptions. J Gen Intern Med 2014 Jul;29(7):1060-1064 [FREE Full text] [doi: 10.1007/s11606-013-2755-z] [Medline: 24452418]
68. Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology 1990 Jan;1(1):43-46. [Medline: 2081237]
69. Veazie PJ. When to combine hypotheses and adjust for multiple tests. Health Serv Res 2006 Jun;41(3 Pt 1):804-818 [FREE Full text] [doi: 10.1111/j.1475-6773.2006.00512.x] [Medline: 16704513]
70. Lotfaliany M, Bowe SJ, Kowal P, Orellana L, Berk M, Mohebbi M. Depression and chronic diseases: co-occurrence and communality of risk factors. J Affect Disord 2018 Dec 01;241:461-468. [doi: 10.1016/j.jad.2018.08.011] [Medline: 30149333]

Abbreviations

CFI: comparative fit index
DERS: Difficulties in Emotion Regulation Scale
DERS-SF: Difficulties in Emotion Regulation Scale short form
EMA: ecological momentary assessment
IRT: item response theory
MSEM: multilevel structural equation model
NA: negative affect
PA: positive affect
RMSEA: root mean square error of approximation
RT: response time
SRMR: standardized root mean square residual
T1D: type 1 diabetes
TLI: Tucker-Lewis Index

Edited by J Torous; submitted 13.03.24; peer-reviewed by LMM Kivelae, MR Frumkin; comments to author 27.04.24; revised version received 17.05.24; accepted 05.06.24; published 18.07.24.

Please cite as:
Hernandez R, Hoogendoorn C, Gonzalez JS, Pyatak EA, Crespo-Ramos G, Schneider S
Reliability and Validity of Ecological Momentary Assessment Response Time–Based Measures of Emotional Clarity: Secondary Data Analysis
JMIR Ment Health 2024;11:e58352
URL: https://mental.jmir.org/2024/1/e58352
doi:10.2196/58352
PMID:39024004

Media Use and Its Associations With Paranoia in Schizophrenia and Bipolar Disorder: Ecological Momentary Assessment

Vincent Paquin[1,2,3], MD; Robert A Ackerman[4], PhD; Colin A Depp[5], PhD; Raeanne C Moore[5], PhD; Philip D Harvey[6], PhD; Amy E Pinkham[4], PhD

1
2
3
4
5
6

Corresponding Author:
Vincent Paquin, MD

Abstract

Background: Paranoia is a spectrum of fear-related experiences that spans diagnostic categories and is influenced by social and cognitive factors. The extent to which social media and other types of media use are associated with paranoia remains unclear.

Objective: We aimed to examine associations between media use and paranoia at the within- and between-person levels.

Methods: Participants were 409 individuals diagnosed with schizophrenia spectrum or bipolar disorder. Measures included sociodemographic and clinical characteristics at baseline, followed by ecological momentary assessments (EMAs) collected 3 times daily over 30 days. EMA evaluated paranoia and 5 types of media use: social media, television, music, reading or writing, and other internet or computer use. Generalized linear mixed models were used to examine paranoia as a function of each type of media use and vice versa at the within- and between-person levels.

Results: Of the 409 participants, the following subgroups reported at least 1 instance of media use: 261 (63.8%) for using social media, 385 (94.1%) for watching TV, 292 (71.4%) for listening to music, 191 (46.7%) for reading or writing, and 280 (68.5%) for other internet or computer use. Gender, ethnoracial groups, educational attainment, and diagnosis of schizophrenia versus bipolar disorder were differentially associated with the likelihood of media use. There was a within-person association between social media use and paranoia: using social media was associated with a subsequent decrease of 5.5% (fold-change 0.945, 95% CI 0.904-0.987) in paranoia. The reverse association, from paranoia to subsequent changes in social media use, was not statistically significant. Other types of media use were not significantly associated with paranoia.

Conclusions: This study shows that social media use was associated with a modest decrease in paranoia, perhaps reflecting the clinical benefits of social connection. However, structural disadvantage and individual factors may hamper the accessibility of media activities, and the mental health correlates of media use may further vary as a function of contents and contexts of use.

(JMIR Ment Health 2024;11:e59198) doi:10.2196/59198

KEYWORDS

paranoia; social media; digital media; technology; psychosis; schizophrenia; schizoaffective; bipolar disorder; ecological momentary assessment; spectrum; sociodemographic; linear mixed model; media use; mental health; digital intervention; adult; adults; medical center; mental health clinic; psychiatry; psychiatrist

Introduction

Background

Paranoia, defined as the "unfounded fear that others intend to cause you harm" [1], is the most commonly identified delusion in schizophrenia spectrum conditions [2] and is also experienced in bipolar disorder, unipolar depression, borderline personality disorder, and other conditions [3-5]. Paranoia is a spectrum of beliefs and experiences spanning mild (eg, people talk about you behind your back) to severe ideas of threat (eg, people are trying to harm you) [6]. Paranoid experiences along this hierarchy are associated with social functioning problems including poorer interpersonal relationships, reduced marital satisfaction, and difficulties with peers [7-12]. Paranoia can fluctuate over the course of days as a function of cognitive and interpersonal factors [13-15]. Factors that have been associated with increased paranoia include rumination, loneliness, and social exclusion, whereas distraction and being in the company of familiar individuals have been associated with decreases in paranoia [13,14,16,17]. Many of these social and cognitive experiences may be influenced by day-to-day engagement with media [18], but it is unclear whether media use influences paranoia in schizophrenia and bipolar disorder.

Possible Effects of Media on Paranoia

Media are vehicles for accessing or sharing information, including for leisure, social, and occupational purposes [19]. This definition encompasses both digital (eg, social media) and traditional (eg, books) forms of media. Similar to other environmental exposures, media may potentially influence a person's paranoia by informing the degree of perceived environmental threat. Better situating the naturalistic effects of media on paranoia may help guide lifestyle counseling and the development of media-based interventions for this symptom dimension.

Multiple mechanisms might contribute to the effects of media on paranoia. First, media activities can provide a distraction that helps a person move away from persecutory thoughts. Social media may help decrease paranoia by facilitating access to social support and reducing loneliness. However, both traditional and digital media activities may also promote paranoia if they perpetuate harmful avoidance behaviors (eg, if a socially withdrawn person watches TV instead of going out and seeing friends) or directly amplify the perception of environmental threat (eg, through consumption of catastrophic news or conspiracist discourses). Additionally, social media has several features that distinguish it from in-person communication; some of these features, such as asynchronicity and lower availability of nonverbal cues, may hypothetically increase uncertainty about social relationships [20], which might build on a person's paranoia.

There is currently limited evidence regarding the relationship between media use and paranoia. Recent studies with the general population (in Canada and the United Kingdom) have found cross-sectional associations between greater use of digital media and higher levels of psychotic-like experiences [21,22], including paranoia [23], but 2 studies with prospective data did not find robust associations between media use and the subsequent risk of psychotic-like experiences [22,24]. Rather than reflecting causal effects of media on paranoia, the concurrent association between media use and paranoia in the general population may be due to confounders, such as preexisting mental health problems and social adversity [24]. However, these findings may not apply to individuals with psychotic or mood disorders, in whom there is generally a greater propensity for paranoia, and thus potentially a greater sensitivity to media effects that are otherwise not apparent in non-clinical populations. As such, surveys indicate that 25% - 35% of individuals with psychotic or mood disorders believe that digital devices can increase their paranoia [25,26].

To our knowledge, only 1 previous study has examined the associations of media use and paranoia in a clinical sample [27]. This study by Berry et al [27] included 44 participants: 19 individuals with schizophrenia spectrum or bipolar disorder and 25 nonclinical individuals. Participants completed ecological momentary assessments (EMAs) of social media use and paranoia 6 times daily over 6 days. Posting about feelings, venting on social media, viewing profiles of people who were not "friends," and lower perceived social rank were associated with higher paranoia at the next time point. The results therefore supported a possible contribution of specific social media experiences to paranoia. However, the results may have been affected by unmeasured confounders, such as individual differences in social adversity or mental health that predisposed individuals to certain patterns of social media behaviors. In addition, because the usage of other types of media was not assessed, it remains unclear to what extent the risk is specific to social media. Another potential gap is whether day-to-day fluctuations in paranoia reciprocally make people more or less drawn to media use [28]—a reverse association that has yet to be examined.

Population Differences in Media Use and Paranoia

An important consideration in investigating the relationship of media activities with paranoia is that media-related behaviors vary between populations. Some studies have found, for example, that people with schizophrenia are less likely than those with bipolar disorder to use digital media [29,30], perhaps reflecting the impact of socioeconomic disadvantage on media access and digital literacy. Similarly, in the general population, there is evidence of unequal access to media technologies as a result of socioeconomic disadvantage [31,32]. Gender is another factor linked to differences in media use, with meta-analytic evidence of higher prevalence of problematic social media use in women versus men [33].

The risk of paranoia may also vary between populations. Although there is limited comparative data, paranoia appears to be common in both schizophrenia and mood disorders and to have similar psychosocial and neurobiological correlates across diagnostic groups [2,4,34,35]. One study found no difference in patterns of attributional bias between schizophrenia and bipolar disorder, suggesting shared cognitive mechanisms [34]. However, paranoia appears to be relatively less common in unipolar depression with psychotic features than in schizophrenia and bipolar disorder, indicating different levels of risk between diagnoses [36]. Likewise, in the general population, levels of paranoia may be higher among men than among women [8]. Whether these epidemiological features might translate into diagnosis- or gender-specific effects of media on paranoia remains unknown.

This Study

To advance knowledge about the relationship between media use and paranoia, this study draws from an EMA sample comprising adults with schizophrenia and those with bipolar disorder. EMA provides dense, repeated measures that are well suited for analyzing bidirectional relationships between media use and paranoia. With EMA data, confounders from individual traits can be removed using within-person analyses, where each person is their own comparator over time [37]. We thus aimed to examine media use and its bidirectional associations with paranoia in people with schizophrenia and those with bipolar disorder, both at the between- and within-person levels. To identify whether specific types of media activities are uniquely associated with paranoia, we considered social media and 4 other types of media use: watching TV, listening to music, reading or writing, and internet or computer use. We first evaluated associations of sociodemographic and clinical characteristics with any versus no use of each type of media across the study period. Then, we examined the associations between media use and paranoia at the between- and within-person levels. We explored moderating effects of gender and clinical group on within-person associations. We hypothesized that social media use would be associated with subsequent increases in paranoia, and that this association would be greater in women than in men. We did not formulate hypotheses for other types of media.

Methods

Participants

Data were obtained from a study designed to assess the introspective accuracy of self-reported cognition and functioning in the context of mental illness [38]. Participants were adults aged 18 - 65 years meeting diagnostic criteria for schizophrenia (or schizoaffective disorder) and bipolar disorder (type 1 or 2). Recruitment sites included The University of Texas at Dallas, Miller School of Medicine–University of Miami, and University of California San Diego. Participants were recruited via medical centers, public mental health and local community clinics, nonprofit organizations, direct contact by service providers, follow-up with research participants from previous projects, study flyers, and internet-based advertisements.

Diagnostic information was collected by trained interviewers using the Mini International Neuropsychiatric Interview [39] and the psychosis module of the Structured Clinical Interview for *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]) [40]. Final diagnoses were generated through a local consensus procedure. Inclusion criteria were clinical stability (ie, no hospitalization or extended emergency department visit) for a minimum of 6 weeks and no significant (>20%) medication dose changes in the past 2 weeks. Participants with bipolar disorder had to have at least 1 mood episode recurrence or incomplete remission from a first episode, indicating stage 3 severity or higher according to the classification of Frank et al [41]. Exclusion criteria were nonproficiency in English; a current or past medical or neurological disorder that may affect brain functioning (eg, brain tumors and seizures); intellectual disability or pervasive developmental disorder; active substance use of moderate severity or higher; and any visual or hearing impairment limiting assessments.

Baseline Assessments

During the initial interview, participants reported their age, gender (male, female, or other), race (Asian, Black or African American, White, or other), ethnicity (Hispanic or non-Hispanic), educational attainment, and relationship status. Trained raters assessed symptom severity the day before the first EMA survey using the Positive and Negative Syndrome Scale (PANSS) [42] for schizophrenia-related symptoms, the Montgomery-Åsberg Depression Rating Scale for depressive symptoms [43], and the Young Mania Rating Scale [44] for mania-related symptoms.

The PANSS consists of 30 items, each scored on a scale from 1=“Absent” to 7=“Extreme.” Its positive symptom subscale (7 items, total score range 7 - 49) includes 1 item that measures the severity of suspiciousness and paranoia (P6). Negative symptoms were assessed with the PANSS using the 2-factor model by developed Khan et al [45], which identifies dimensions of expressive and experiential deficits. The items for reduced emotional experience include emotional withdrawal (N2), passive or apathetic social withdrawal (N4), and active social avoidance (G16). The items for reduced emotional expression include blunted affect (N1), poor rapport (N3), lack of spontaneity and flow of conversation (N6), and motor retardation (G7).

EMAs

Participants completed EMA surveys via the NeuroUX platform (Playpower Labs Inc) with either their own smartphone or a smartphone provided by the study investigators. The initial protocol planned that all participants would be provided smartphones, but participants were allowed to use their own smartphones following the onset of the COVID-19 pandemic to facilitate more flexible participation. In the context of this rapid modification of the protocol, data on smartphone ownership were not collected systematically and thus could not be analyzed.

Participants received SMS text messages prompting them to complete internet-based surveys 3 times daily for 30 days. Data were instantly uploaded to cloud-based servers. Text messages were sent at stratified random intervals within, on average, 2-hour windows and between 9 AM and 9 PM. The first and last daily assessment times were adapted to each participant’s typical sleep and wake schedules. Survey responses were only allowed within 1-hour periods following deployment of the SMS text messages. Participants had the option of silencing the alarms for 30-minute intervals (eg, if driving).

Media use was measured using checkbox questions about activities performed since the previous survey. Response options included six items: (1) using “social media (eg, Facebook, Twitter)”; (2) “watching TV”; (3) “listening to music”; (4) “reading, writing, or journaling”; (5) “shopping online”; and (6) “other internet, computer, or tablet use.” Due to low endorsement rates for shopping on the web, this activity was merged with other internet, computer, or tablet uses. Other types of activities, such as working, excising, eating, etc, were also measured as part of this questionnaire [46] but were not analyzed in this study. Participants could report multiple activities per assessment. For each type of media, these checkbox questions produced dichotomous indices of use at each survey (0=did not use and 1=used).

Paranoia was measured using the following item: “Since the past alarm, how much have you had thoughts that others might want to harm you or that people are untrustworthy?” Response choices ranged from 1 to 7, with higher values indicating greater severity or frequency. In support of the convergent validity of this item, we found that individual mean scores for paranoia across the 30 days were significantly correlated with scores on the PANSS suspiciousness item (Spearman ρ=0.41, P<.001). Both media use and paranoia were measured 3 times daily.

Statistical Analysis

Analyses were conducted in R (version 4.1.2; The R Foundation). Codes are available on the web [47]. To minimize the impact of potentially invalid entries, we excluded participants who completed less than one-third of EMA surveys, following a common rule of thumb in EMA research [48]. Characteristics of included and excluded participants were compared to consider attrition effects; in line with recommendations for descriptive statistics [49], we used effect sizes instead of *P* values for these comparisons. Descriptive analyses such as the root-mean-square of successive differences (RMSSD) and the intraclass correlation coefficient were used to evaluate the variability of paranoia EMAs [50].

To examine sociodemographic and clinical correlates of media use, we dichotomized each type of media use as any versus no use over the follow-up period. We regressed these dichotomous variables on the predictors using logistic regressions adjusted for age and gender. To account for pairwise evaluations of 16 predictors with each media type, we report nominal statistical significance after adjustment for the false discovery rate (using the Benjamini and Hochberg [51] method) in addition to uncorrected 95% CIs.

Next, to examine associations between ecological momentary reports of media use and paranoia (between and within persons), we used generalized mixed models with observations nested in individuals [52]. Models were estimated using maximum likelihood with Laplace approximation. We considered 2 directions of association: media use as a function of paranoia, and paranoia as a function of media use. For models of media use as a function of paranoia, we applied the binomial distribution (logistic regression) given that media use assessments were dichotomous. For models of paranoia as a function of media use, we applied a gamma distribution with log-link function given the skewed distribution of paranoia. All models included random intercepts, mean levels of the predictor, and lagged mean-centered values of the predictor. Random slopes of the lagged mean-centered predictor were also added if they improved model fit ($P<.05$) on the likelihood ratio test [53]. Values of 95% CI not overlapping the null were considered significant.

For each model of media use and paranoia, we subsequently explored potential moderating effects of gender (0=male, 1=female) and clinical group (0=schizophrenia, 1=bipolar) on within-person associations. Interactions were considered statistically significant at $P<.05$. Recognizing, however, that interaction analyses are frequently underpowered due to weaker effects [54], we also probed interactions if the *P* value ranged between .05 and .10 and reported them as "tentative." To probe interactions, we estimated marginal slopes of lagged mean-centered predictors as a function of the moderator [55].

In sensitivity analyses, we considered whether including all participants (ie, not removing participants who completed less than one-third of EMA surveys) impacted the primary results. We also considered the impact of auto-correlation by controlling for the lagged dependent variables.

Ethical Considerations

All participants provided written informed consent, and the institutional review board of each university approved the study (The University of Texas at Dallas, #18 - 93; Miller School of Medicine–University of Miami, #20180352; and University of California San Diego, #180716).

Results

Of 446 participants, 409 completed at least one-third of the EMA surveys (Table 1). Completion of at least one-third of assessments was associated with the site of recruitment, higher educational attainment, and non-Hispanic ethnicity (effect sizes≥0.100; see Table S1 in Multimedia Appendix 1). Table 2 provides descriptive statistics on paranoia EMAs, indicating higher mean ratings and greater intraindividual variability (higher SD and RMSSD) of paranoia in the schizophrenia versus bipolar disorder group.

Table . Characteristics of participants in the schizophrenia and bipolar disorder groups. Positive symptoms were measured with the positive symptom subscale of the Positive and Negative Syndrome Scale (PANSS; score range 7-49). Reduced emotional experience (range 3-21) and reduced emotional expression (range 4-28) were measured with negative symptom and general psychopathology items of the PANSS. Depressive symptoms were measured with the Montgomery-Åsberg Depression Rating Scale (range 0-60). Mania-related symptoms were measured with the Young Mania Rating Scale (range 0-60).

		Schizophrenia (n=189)	Bipolar disorder (n=220)
Site of recruitment, n (%)			
	The University of Texas at Dallas	86 (45.5)	87 (39.5)
	University of Miami	61 (32.3)	54 (24.5)
	University of California San Diego	42 (22.2)	79 (35.9)
Age (years), median (IQR)		41 (32-52)	39 (30-50)
Gender, n (%)			
	Male	93 (49.2)	69 (31.4)
	Female	95 (50.3)	150 (68.2)
	Other	<5 (<1)	<5 (<1)
Race, n (%)			
	White	72 (38.1)	140 (63.6)
	Black or African American	96 (50.8)	42 (19.1)
	Asian	5 (2.65)	15 (6.82)
	Other	16 (8.47)	23 (10.5)
Ethnicity, n (%)			
	Hispanic	40 (21.2)	56 (25.6)
	Non-Hispanic	149 (78.8)	163 (74.4)
Educational level, n (%)			
	High school diploma or less	83 (43.9)	40 (18.2)
	Some college	71 (37.6)	77 (35.0)
	College degree or higher	35 (18.5)	103 (46.8)
Relationship status, n (%)			
	Not in a relationship	115 (60.8)	101 (45.9)
	In a relationship	74 (39.2)	119 (54.1)
Positive symptoms, median (IQR)		16 (13-19)	11 (9-14)
Reduced emotional experience, median (IQR)		6 (3-8)	3 (3-6)
Reduced emotional expression, median (IQR)		6 (4-9)	4 (4-6)
Depressive symptoms, median (IQR)		6 (0-16)	11 (0-20)
Mania-related symptoms, median (IQR)		0 (0-0)	0 (0-5)

Table . Descriptive statistics of paranoia ecological momentary assessments. Person-level statistics were calculated for each participant then averaged (with SD values) across individuals. *P* values indicate group differences in these statistics based on *t* tests. The intraclass correlation coefficient (ICC) corresponds to the proportion of between-person variance; the proportion of within-person variance is 1–ICC.

Descriptor		Diagnostic group		*P* value
		Schizophrenia (n=189)	Bipolar disorder (n=220)	
Number of surveys across participants, n				
	No paranoia (score=1)	6685	12,630	—[a]
	Some paranoia (score>1)	6445	2684	—
	Missing	3880	4486	—
	Total	17,010	19,800	—
Some paranoia (score>1) on at least 1 assessment, n (%)		164 (86.6)	123 (55.9)	—
Person-level mean of paranoia, mean (SD)		2.45 (1.57)	1.57 (1.18)	<.001
Person-level SD of paranoia, mean (SD)		0.91 (0.65)	0.47 (0.66)	<.001
Person-level RMSSD[b] of paranoia, mean (SD)		0.93 (0.69)	0.48 (0.68)	<.001
ICC[c] (95% CI)		0.66 (0.61-0.71)	0.69 (0.65-0.73)	—

[a]Not applicable.

[b]RMSSD: root-mean-square of successive differences.

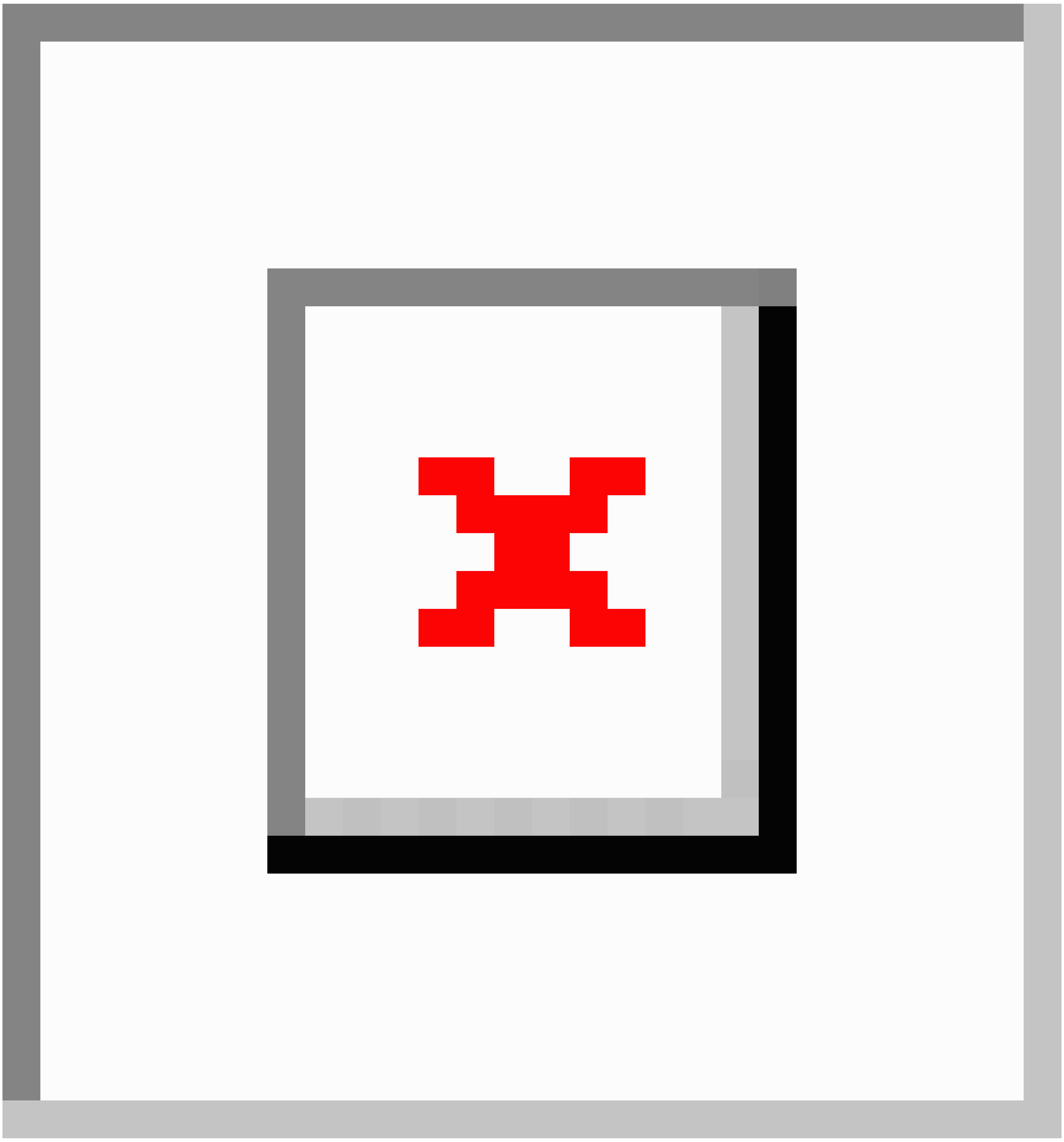
[c]ICC: intraclass correlation coefficient.

Of the 409 participants, the following subgroups reported at least 1 instance of media use: 261 (63.8%) for using social media, 385 (94.1%) for watching TV, 292 (71.4%) for listening to music, 191 (46.7%) for reading or writing, and 280 (68.5%) for other internet or computer use. Within each of these subgroups, there was a total of 1982 reports of social media over 18,478 available surveys (10.7%), 6779 reports of watching TV over 26,911 available surveys (25.2%), 1997 reports of listening to music over 20,778 available surveys (9.6%), 864 reports of reading or writing over 13,478 available surveys (6.4%), and 2500 reports of other internet or computer use over 19,649 available surveys (12.7%).

Factors Associated With Any Versus No Media Use

Figure 1 presents factors associated with any versus no use of each type of media. After adjustment for the false discovery rate, female versus male gender was associated with higher odds of social media use. Black or African American individuals (versus White individuals) had lower odds of other internet or computer use. Hispanic versus non-Hispanic ethnicity was associated with lower odds of listening to music. Higher educational attainment was associated with higher odds of social media use, reading or writing, and other internet or computer use. Diagnosis of bipolar disorder versus schizophrenia was associated with higher odds of social media and other internet or computer use. Higher levels of positive symptoms (but not paranoia) were associated with lower odds of other internet or computer use.

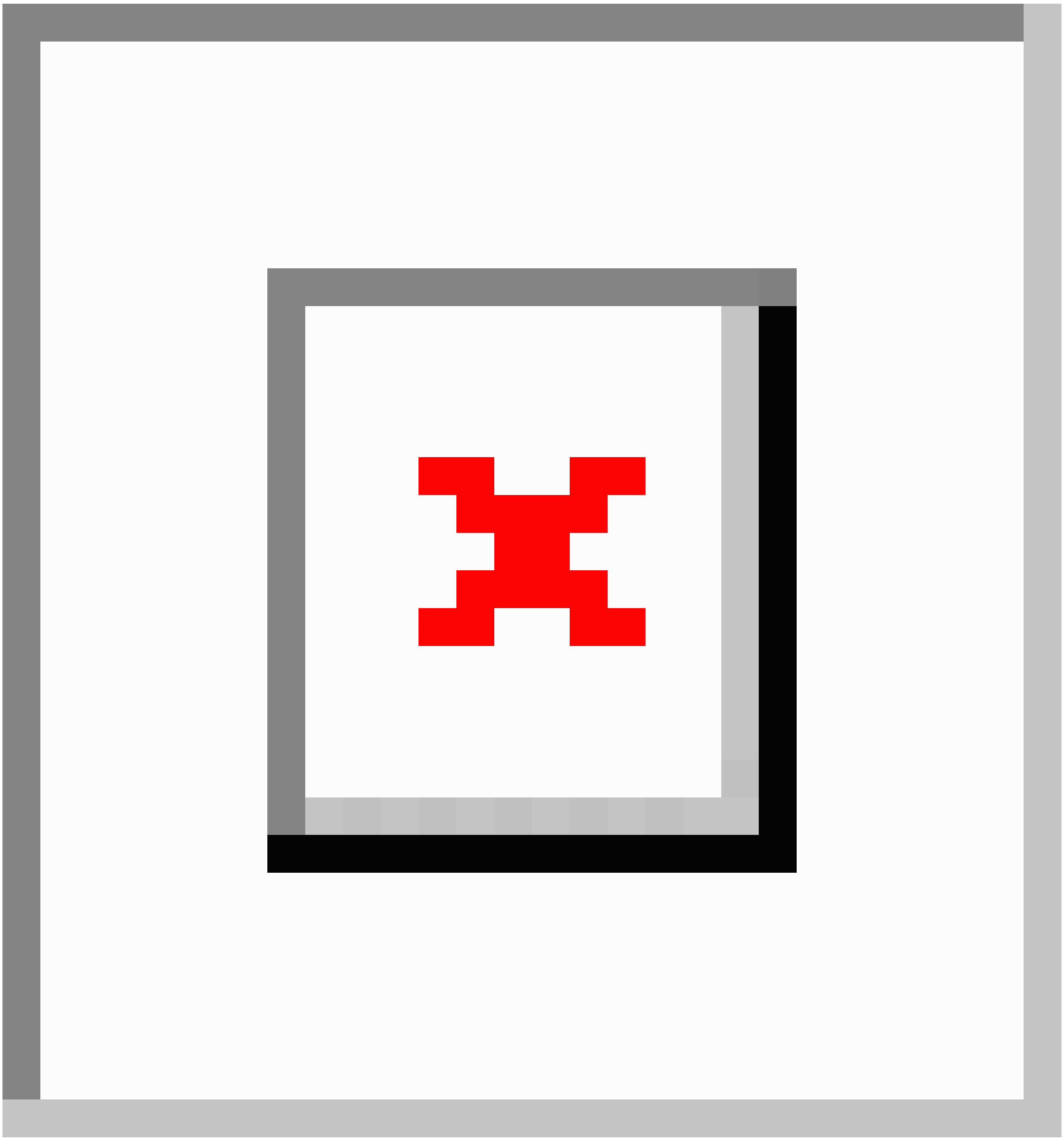
Figure 1. Odds ratios (95% CIs) of any versus no use of each type of media as a function of participant characteristics. Logistic regression models were adjusted for age and gender (n=409). Statistical significance after adjustment for the false discovery rate across columns (adjusted *P*<.05) is indicated in boldface. Comparator category for female gender: male gender; for Black or African American ("Black"), Asian, and other racialized groups: White; for Hispanic ethnicity: non-Hispanic ethnicity; for some college education and college degree or higher: high school or lower; and for the bipolar group: the schizophrenia group. Paranoia is the mean rating for paranoia across ecological momentary assessments (range 1 - 7).



Paranoia as a Function of Media Use

Figure 2 presents the mean ratio (between persons) and fold-change (within persons) in paranoia as a function of media use. Analyses are in the subsets of participants reporting that type of media (ie, subgroups ranging from n=191 for reading or writing to n=385 for watching TV). Across types of media, associations were not significant at the between-person level: participants' average levels of media use were not associated with their average levels of paranoia. At the within-person level, social media use above a person's average was associated with a ~5.5% reduction in their level of paranoia on the subsequent survey (fold-change 0.945; 95% CI 0.904-0.987). Other within-person associations were not significant.

Figure 2. Paranoia as a function of preceding (lagged) media use in gamma mixed models with observations nested within individuals. Predictors included mean levels of media use (for between-person associations) and lagged mean-centered media use (for within-person associations). All models include random slopes of lagged mean-centered media use.

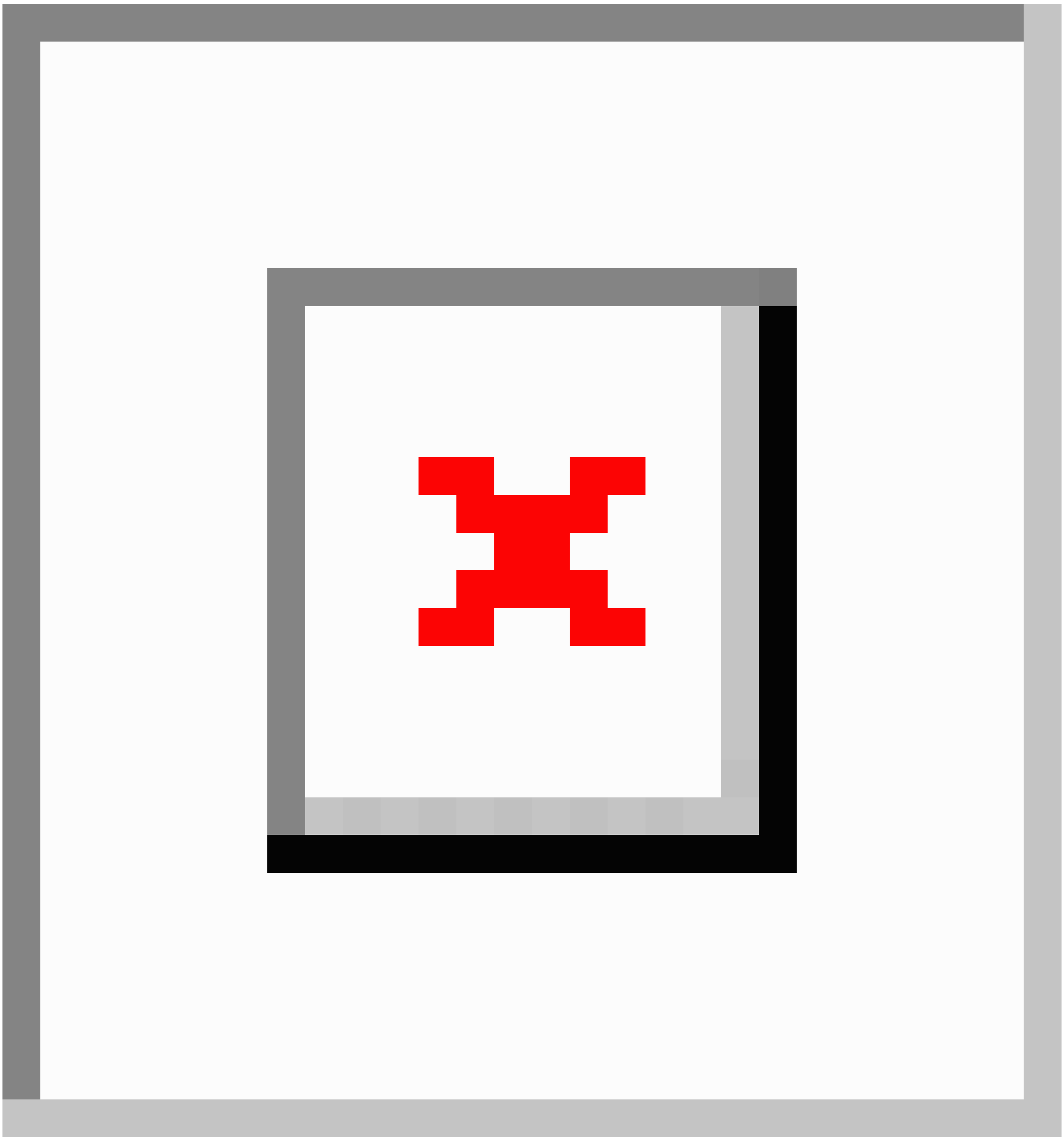


There was a tentative moderating effect of the clinical group on the within-person association of social media use with paranoia (coefficient, lagged social media use×bipolar disorder versus schizophrenia=1.084; *P*=.07). According to this interaction, there was a within-person association between social media use and subsequently reduced paranoia in the schizophrenia group (fold-change 0.900, 95% CI 0.840-0.965). In the bipolar disorder group, however, the same association was not statistically significant (fold-change 0.975, 95% CI 0.923-1.031). There was also a tentative moderating effect of the clinical group on the association between listening to music and paranoia (coefficient, lagged music×bipolar disorder versus schizophrenia=0.939; *P*=.08). According to this interaction, there was a within-person association between listening to music and subsequently decreased paranoia in the bipolar disorder group (fold-change 0.949, 95% CI 0.906-0.995). This association was not significant in the schizophrenia group (fold-change 1.011, 95% CI 0.961-1.063). Other moderating effects of gender or clinical group did not pass the *P*<.10 threshold and were not probed further.

Media Use as a Function of Paranoia

Figure 3 presents odds ratios of each type of media use as a function of paranoia. Associations were not significant at the between-person level: participants' average levels of paranoia were not associated with the odds of media use. There were also no significant associations at the within-person level: participants' variations in paranoia relative to their personal average did not predict variations in their odds of media use over time. The moderating effects of gender or clinical group over within-person associations did not pass the threshold of *P*<.10 and were not probed further.

Figure 3. Media use as a function of preceding (lagged) paranoia in generalized logistic mixed models with observations nested in individuals. Predictors include mean levels of paranoia (for between-person associations) and lagged mean-centered paranoia (for within-person associations). Models of social media use and TV include random slopes of lagged mean-centered paranoia. Other models include random intercepts only. Lagged mean-centered paranoia was removed from models of reading or writing due to lack of convergence (indicated by "NA" in the bottom panel).



Additional Analyses

Within- and between-person associations of media use and paranoia were consistent in the full sample, that is, after lifting the exclusion rule for participants with less than one-third of EMA surveys (see Figures S1 and S2 in Multimedia Appendix 1). Results were also consistent in models adjusted for lagged values of the dependent variables, suggesting no impact of auto-correlated effects on the association between lagged social media use and lower paranoia (see Figures S3 and S4 in Multimedia Appendix 1). Post hoc, we explored associations between media use and paranoia on concurrent surveys, instead of across lagged surveys. None of the concurrent associations were significant (see Figures S5 and S6 in Multimedia Appendix 1).

Discussion

Principal Findings

In a large cohort of individuals with schizophrenia or bipolar disorder, we examined the usage of social media and 4 other types of media, and their associations with paranoia over 30 days of EMA surveys. Sociodemographic and clinical characteristics were associated with the likelihood of any versus no use of media types. In individuals who did use those media over the follow-up period, media use was either associated with no significant change in paranoia, or some decrease, depending on the type of media and diagnostic group.

We did not find evidence that media use predicted increased paranoia overall, or that paranoia predicted subsequent media use. Contrary to our hypothesis, there was instead an association between social media use and a subsequent decrease in paranoia. The reduction was approximately 5.5% in the total sample and up to 10% in the schizophrenia group. Although there are no established benchmarks for interpreting the clinical significance of paranoia reductions on the present scale, it should be noted that these reductions are smaller than 1-point changes relative to the 7-point range of the scale, leading us to interpret them as modest. The reverse associations (between paranoia and subsequent social media use) and concurrent associations (between paranoia and concurrent social media use) were not significant, supporting a temporal sequence between social media use and subsequent decreases in paranoia.

Although social media may have negative effects on mental health through negative social comparisons, increased uncertainty in social communication, and other mechanisms [20,27,56], there is evidence that people with psychosis [57] and other mental health conditions [58] primarily use social media to connect with friends or family. Other studies have found that (physical) exposure to familiar social company and lower levels of loneliness are associated with decreased paranoia [14,16]. Building on this literature, this finding may reflect the effect of web-based social company through social media on reducing paranoia. The fact that this association was stronger in the schizophrenia versus bipolar disorder group is of unclear significance. In the literature, loneliness has been reported as a risk factor for paranoia regardless of whether people have a psychotic disorder [16], suggesting that increased social company through social media could hypothetically be protective across diagnostic groups. We believe that the lack of an association between social media and paranoia in the bipolar disorder group may be a consequence of the group's lower intraindividual variation in paranoia, which constrained the ability to detect effects of social media.

To our knowledge, the only previous investigation of longitudinal associations between social media use and paranoia is the one conducted by Berry et al [27]. While they did not directly assess the perception of web-based social company, they found that viewing social media newsfeeds predicted lower paranoia, whereas a perceived low social rank when using social media predicted higher paranoia. Their results, in line with community-based research on other aspects of well-being [59,60], indicate that social media can have both positive and negative associations with mental health depending on the contents of use. Our study measured social media use as a whole and did not distinguish among browsing, posting, or private messaging on social media. Hence, the results may be best understood as average correlates of media behaviors, spanning reductions in paranoia to increases therein, with differences in the direction of association that may arise from unmeasured individual, environmental, and media-related factors [61].

After probing a tentative (not statistically significant) interaction, we also found an association between listening to music and a subsequent decrease in paranoia in people with bipolar disorder. Potentially beneficial effects of music have been described in an internet-based sample of 457 adults with schizophrenia in the United States, where 42% of them reported that listening to music or audio files helped with managing auditory hallucinations [25]. The association in this study was modest in size, but it could be argued that in some individuals, listening to music promotes distraction from paranoid thoughts [17].

Differences in the likelihood of media use as a function of participant characteristics indicate the potential impacts of socioenvironmental adversity. Consistent with prior research [29,30], participants diagnosed with schizophrenia rather than bipolar disorder were less likely to report any versus no social media use and any versus no other forms of internet and computer use, a gap that may stem from lower socioeconomic status and functional impairments associated with schizophrenia. Black versus White participants and those who did not go to college were also less likely to report any versus no internet or computer use. Prior research from clinical samples in the United States and other western countries similarly show that lower educational attainment was associated with a lower likelihood of using digital media [26,32,62], and in a US sample of 322 people receiving psychiatric care, participants of racialized minority groups were less likely than White participants to use social media [29]. These associations may reflect the impact of systemic inequities on material deprivation and lower media-related literacy, which are some of the most common barriers to digital media use in clinical and nonclinical populations [31,32].

Strengths and Limitations

This study is novel in its examination of media use in a large sample of participants with psychiatric diagnoses, who altogether contributed thousands of EMA surveys. By modeling associations between media use and paranoia at the within-person level, we were able to remove potential influences from time-invariant confounders such as sociodemographic features [63], thereby lending greater credence to possible causal effects of media use on paranoia. However, this method is not exempt from other sources of bias, including those from time-varying confounders: for example, a person's use of social media could be triggered by a momentary increase in the availability of social support, which could be driving subsequent improvements in paranoia, thus acting as a time-varying confounder of the association between social media use and decreased paranoia. The validity of the media use questionnaire could not be assessed, but previous research has found moderate correlations between self-reported digital media use and app-or device-based activity logs [64]. The types of media evaluated here were not exhaustive and they aggregated heterogeneous media contents that may be differentially associated with paranoia. Each type of media was assessed with a single item, which was open to the respondent's interpretation: for example, social media may have been understood by some participants and not others as including private messaging. Dichotomous assessments of media use did not provide information on the duration and contents of use (eg, browsing newsfeeds or messaging individuals). Paranoia EMA was evaluated with a single item, which had a significant, moderate correlation with clinician-rated trait paranoia but has not been validated in external samples. Our findings were also bound by a specific time frame of 3 assessments per day that produced intervals of hours between measures, potentially underestimating the frequency of media use. Effects of media use on paranoia may unfold over longer intervals than those assessed here (eg, after weeks of intense exposure). Lastly, while lower incidence of paranoia in the bipolar group might have constrained statistical power, the results raise the possibility of diagnosis-specific associations with media use, which may help better tailor media-based interventions.

Future research that provides more detailed assessments of media use over various time frames will be needed to examine these questions. Inclusion of putative mechanisms (eg, decreased loneliness) and concurrent symptoms (eg, negative symptoms) will also help elucidate the pathways between media use and paranoia. Another direction for future research is to consider associations of media use with paranoia in additional populations, notably adolescents and individuals with other mental health conditions such as borderline personality disorder and depression.

Conclusions

This study found that social media use was associated with modest decreases in paranoia in a sample of individuals with schizophrenia or bipolar disorder. A better understanding of the social and cognitive possibilities of media technologies may help guide lifestyle counseling and media-based interventions for this population. However, systemic inequities and individual factors may hamper the accessibility of media, and the mental health correlates of media use may further vary as a function of specific contents and contexts of use.

Acknowledgments

This research was supported by a National Institute of Mental Health (NIMH) grant (RO1MH112620) awarded to AEP. VP is supported by an award from the Fonds de recherche du Québec – Santé and the Ministère de la santé et des services sociaux (FRQS-MSSS). The funders had no role in the preparation of this paper.

Data Availability

The data in this study are being deposited in the NIMH's Research Domain Criteria (RDoC) repository. They will be available for public access 6 months after data lock. In the interim, the data that support the findings of this study are available from AEP upon reasonable request.

Authors' Contributions

VP designed the research questions, reviewed the literature, conducted the analyses, and drafted the manuscript. AEP contributed to writing the manuscript and was the lead for funding acquisition and data collection. CAD, RCM, and PDH contributed to funding acquisition and data collection. RAA contributed to formulating the analytic plan. All authors contributed to manuscript editing and data interpretation and approved the final manuscript.

Conflicts of Interest

RCM is a cofounder of KeyWise AI, Inc, and NeuroUX, Inc. The terms of these arrangements have been reviewed and approved by the University of California San Diego in accordance with its conflict of interests policies. PDH has received consulting fees or travel reimbursements from Alkermes, Bio Excel, Boehringer Ingelheim, Intra-Cellular Therapies, Minerva Pharma, Otsuka America, Regeneron, Roche Pharma, and Sunovion Pharma. He receives royalties from the Brief Assessment of Cognition in Schizophrenia and the MATRICS Consensus Battery. He is chief scientific officer of i-Function, Inc. The other authors declare no potential conflicts of interest.

Multimedia Appendix 1
Additional information.
[DOCX File, 582 KB - mental_v11i1e59198_app1.docx]

References

1. Freeman D, Pugh K, Antley A, et al. Virtual reality study of paranoid thinking in the general population. Br J Psychiatry 2008 Apr;192(4):258-263. [doi: 10.1192/bjp.bp.107.044677] [Medline: 18378984]

2. Bentall RP, Rowse G, Shryane N, et al. The cognitive and affective structure of paranoid delusions: a transdiagnostic investigation of patients with schizophrenia spectrum disorders and depression. Arch Gen Psychiatry 2009 Mar;66(3):236-247. [doi: 10.1001/archgenpsychiatry.2009.1] [Medline: 19255373]
3. Goodwin FK, Jamison KR. Manic-Depressive Illness: Bipolar Disorders and Recurrent Depression, 2nd edition: Oxford University Press; 2007.
4. Picardi A, Fonzi L, Pallagrosi M, Gigantesco A, Biondi M. Delusional themes across affective and non-affective psychoses. Front Psychiatry 2018;9:132. [doi: 10.3389/fpsyt.2018.00132] [Medline: 29674982]
5. Kingdon DG, Ashcroft K, Bhandari B, et al. Schizophrenia and borderline personality disorder: similarities and differences in the experience of auditory hallucinations, paranoia, and childhood trauma. J Nerv Ment Dis 2010 Jun;198(6):399-403. [doi: 10.1097/NMD.0b013e3181e08c27] [Medline: 20531117]
6. Freeman D. Suspicious minds: the psychology of persecutory delusions. Clin Psychol Rev 2007 May;27(4):425-457. [doi: 10.1016/j.cpr.2006.10.004] [Medline: 17258852]
7. Bird JC, Fergusson EC, Kirkham M, et al. Paranoia in patients attending child and adolescent mental health services. Aust N Z J Psychiatry 2021 Dec;55(12):1166-1177. [doi: 10.1177/0004867420981416] [Medline: 33423520]
8. Freeman D, McManus S, Brugha T, Meltzer H, Jenkins R, Bebbington P. Concomitants of paranoia in the general population. Psychol Med 2011 May;41(5):923-936. [doi: 10.1017/S0033291710001546] [Medline: 20735884]
9. Hajdúk M, Klein HS, Harvey PD, Penn DL, Pinkham AE. Paranoia and interpersonal functioning across the continuum from healthy to pathological - network analysis. Br J Clin Psychol 2019 Mar;58(1):19-34. [doi: 10.1111/bjc.12199] [Medline: 30028025]
10. Pinkham AE, Harvey PD, Penn DL. Paranoid individuals with schizophrenia show greater social cognitive bias and worse social functioning than non-paranoid individuals with schizophrenia. Schizophr Res Cogn 2016 Mar;3:33-38. [doi: 10.1016/j.scog.2015.11.002] [Medline: 27990352]
11. South SC. Personality pathology and daily aspects of marital functioning. Personal Disord 2014 Apr;5(2):195-203. [doi: 10.1037/per0000039] [Medline: 24364502]
12. Springfield CR, Ackerman RA, Pinkham AE. The dyadic effects of subclinical paranoia on relationship satisfaction in roommate relationships and college adjustment. J Abnorm Psychol 2021 Oct;130(7):797-805. [doi: 10.1037/abn0000689] [Medline: 34398625]
13. Fett AKJ, Hanssen E, Eemers M, Peters E, Shergill SS. Social isolation and psychosis: an investigation of social interactions and paranoia in daily life. Eur Arch Psychiatry Clin Neurosci 2022 Feb;272(1):119-127. [doi: 10.1007/s00406-021-01278-4] [Medline: 34129115]
14. Collip D, Oorschot M, Thewissen V, Van Os J, Bentall R, Myin-Germeys I. Social world interactions: how company connects to paranoia. Psychol Med 2011 May;41(5):911-921. [doi: 10.1017/S0033291710001558] [Medline: 20735885]
15. Buck B, Hallgren KA, Scherer E, et al. Capturing behavioral indicators of persecutory ideation using mobile technology. J Psychiatr Res 2019 Sep;116:112-117. [doi: 10.1016/j.jpsychires.2019.06.002] [Medline: 31226579]
16. Bell V, Velthorst E, Almansa J, Myin-Germeys I, Shergill S, Fett AK. Do loneliness and social exclusion breed paranoia? An experience sampling investigation across the psychosis continuum. Schizophr Res Cogn 2023 Sep;33:100282. [doi: 10.1016/j.scog.2023.100282] [Medline: 37006705]
17. Martinelli C, Cavanagh K, Dudley REJ. The impact of rumination on state paranoid ideation in a nonclinical sample. Behav Ther 2013 Sep;44(3):385-394. [doi: 10.1016/j.beth.2013.02.002] [Medline: 23768666]
18. Paquin V, Ferrari M, Sekhon H, Rej S. Time to think “meta”: a critical viewpoint on the risks and benefits of virtual worlds for mental health. JMIR Serious Games 2023 Feb 7;11:e43388. [doi: 10.2196/43388] [Medline: 36661284]
19. McLuhan M. Understanding Media: The Extensions of Man: University of Toronto Press; 1964.
20. Ferguson AM, Turner G, Orben A. Social uncertainty in the digital world. Trends Cogn Sci 2024 Apr;28(4):286-289. [doi: 10.1016/j.tics.2024.02.005] [Medline: 38448356]
21. Lin BD, Pries LK, Sarac HS, et al. Nongenetic factors associated with psychotic experiences among UK Biobank participants: exposome-wide analysis and mendelian randomization analysis. JAMA Psychiatry 2022 Sep 1;79(9):857-868. [doi: 10.1001/jamapsychiatry.2022.1655] [Medline: 35857297]
22. Paquin V, Philippe FL, Shannon H, Guimond S, Ouellet-Morin I, Geoffroy MC. Associations between digital media use and psychotic experiences in young adults of Quebec, Canada: a longitudinal study. Soc Psychiatry Psychiatr Epidemiol 2024 Jan;59(1):65-75. [doi: 10.1007/s00127-023-02537-6] [Medline: 37516683]
23. Bird JC, Evans R, Waite F, Loe BS, Freeman D. Adolescent paranoia: prevalence, structure, and causal mechanisms. Schizophr Bull 2019 Sep 11;45(5):1134-1142. [doi: 10.1093/schbul/sby180] [Medline: 30534970]
24. Paquin V, Ferrari M, Rej S, et al. Trajectories of adolescent media use and their associations with psychotic experiences. JAMA Psychiatry 2024 Apr 10:e240384. [doi: 10.1001/jamapsychiatry.2024.0384] [Medline: 38598210]
25. Gay K, Torous J, Joseph A, Pandya A, Duckworth K. Digital technology use among individuals with schizophrenia: results of an online survey. JMIR Ment Health 2016 May 4;3(2):e15. [doi: 10.2196/mental.5379] [Medline: 27146094]
26. Miller BJ, Stewart A, Schrimsher J, Peeples D, Buckley PF. How connected are people with schizophrenia? Cell phone, computer, email, and social media use. Psychiatry Res 2015 Feb 28;225(3):458-463. [doi: 10.1016/j.psychres.2014.11.067] [Medline: 25563669]

27. Berry N, Emsley R, Lobban F, Bucci S. Social media and its relationship with mood, self-esteem and paranoia in psychosis. Acta Psychiatr Scand 2018 Dec;138(6):558-570. [doi: 10.1111/acps.12953] [Medline: 30203454]
28. Eisner E, Berry N, Bucci S. Digital tools to support mental health: a survey study in psychosis. BMC Psychiatry 2023 Oct 7;23(1):726. [doi: 10.1186/s12888-023-05114-y] [Medline: 37803367]
29. Beard C, Silverman AL, Forgeard M, Wilmer MT, Torous J, Björgvinsson T. Smartphone, social media, and mental health app use in an acute transdiagnostic psychiatric sample. JMIR Mhealth Uhealth 2019 Jun 7;7(6):e13364. [doi: 10.2196/13364] [Medline: 31199338]
30. Young AS, Cohen AN, Niv N, et al. Mobile phone and smartphone use by people with serious mental illness. Psychiatr Serv 2020 Mar 1;71(3):280-283. [doi: 10.1176/appi.ps.201900203] [Medline: 31744429]
31. Robinson L, Schulz J, Blank G, et al. Digital inequalities 2.0: legacy inequalities in the information age. First Monday 2020;25(7) [FREE Full text] [doi: 10.5210/fm.v25i7.10842]
32. Trefflich F, Kalckreuth S, Mergl R, Rummel-Kluge C. Psychiatric patients' internet use corresponds to the internet use of the general public. Psychiatry Res 2015 Mar 30;226(1):136-141. [doi: 10.1016/j.psychres.2014.12.037] [Medline: 25623020]
33. Su W, Han X, Yu H, Wu Y, Potenza MN. Do men become addicted to internet gaming and women to social media? A meta-analysis examining gender-related differences in specific internet addiction. Comp Human Behav 2020 Dec;113:106480. [doi: 10.1016/j.chb.2020.106480]
34. Sanford N, Woodward TS. Symptom-related attributional biases in schizophrenia and bipolar disorder. Cogn Neuropsychiatry 2017 Jul;22(4):263-279. [doi: 10.1080/13546805.2017.1314957] [Medline: 28421867]
35. Fan L, Klein H, Bass E, Springfield C, Pinkham A. Amygdala hyperconnectivity in the paranoid state: a transdiagnostic study. J Psychiatr Res 2021 Jun;138:117-124. [doi: 10.1016/j.jpsychires.2021.03.049] [Medline: 33848967]
36. Rosen C, Marvin R, Reilly JL, et al. Phenomenology of first-episode psychosis in schizophrenia, bipolar disorder, and unipolar depression: a comparative analysis. Clin Schizophr Relat Psychoses 2012 Oct;6(3):145-151. [doi: 10.3371/CSRP.6.3.6] [Medline: 23006239]
37. Myin-Germeys I, Kuppens P, editors. The Open Handbook of Experience Sampling Methodology: A Step-by-Step Guide to Designing, Conducting, and Analyzing ESM Studies: The Center for Research on Experience Sampling and Ambulatory Methods (REAL); 2022.
38. Morgan O, Strassnig MT, Moore RC, et al. Accuracy of immediate self-assessment of neurocognitive test performance: associations with psychiatric diagnosis and longitudinal psychotic symptoms. J Psychiatr Res 2022 Dec;156:594-601. [doi: 10.1016/j.jpsychires.2022.10.069] [Medline: 36372002]
39. Sheehan DV, Lecrubier Y, Sheehan KH, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 1998;59 Suppl 20:22-33. [Medline: 9881538]
40. First MB, Williams JBW, Karg RS, Spitzer RL. Structured Clinical Interview for DSM-5® Disorders—Clinician Version (SCID-5-CV): American Psychiatric Association; 2016.
41. Frank E, Nimgaonkar VL, Phillips ML, Kupfer DJ. All the world's a (clinical) stage: rethinking bipolar disorder from a longitudinal perspective. Mol Psychiatry 2015 Feb;20(1):23-31. [doi: 10.1038/mp.2014.71] [Medline: 25048003]
42. Kay SR, Fiszbein A, Opler LA. The positive and negative syndrome scale (PANSS) for schizophrenia. Schizophr Bull 1987;13(2):261-276. [doi: 10.1093/schbul/13.2.261] [Medline: 3616518]
43. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry 1979 Apr;134:382-389. [doi: 10.1192/bjp.134.4.382] [Medline: 444788]
44. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. Br J Psychiatry 1978 Nov;133:429-435. [doi: 10.1192/bjp.133.5.429] [Medline: 728692]
45. Khan A, Liharska L, Harvey PD, Atkins A, Ulshen D, Keefe RSE. Negative symptom dimensions of the Positive and Negative Syndrome Scale across geographical regions: implications for social, linguistic, and cultural consistency. Innov Clin Neurosci 2017 Dec 1;14(11–12):30-40. [Medline: 29410935]
46. Strassnig MT, Miller ML, Moore R, Depp CA, Pinkham AE, Harvey PD. Evidence for avolition in bipolar disorder? A 30-day ecological momentary assessment comparison of daily activities in bipolar disorder and schizophrenia. Psychiatry Res 2021 Jun;300:113924. [doi: 10.1016/j.psychres.2021.113924] [Medline: 33848963]
47. Paquin V. Media use and paranoia. Open Science Framework. 2023. URL: https://osf.io/cxvq4/ [accessed 2024-05-17]
48. Palmier-Claus JE, Myin-Germeys I, Barkus E, et al. Experience sampling research in individuals with mental illness: reflections and guidance. Acta Psychiatr Scand 2011 Jan;123(1):12-20. [doi: 10.1111/j.1600-0447.2010.01596.x] [Medline: 20712828]
49. Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Ann Intern Med 2007 Oct 16;147(8):1500-1524. [doi: 10.7326/0003-4819-147-8-200710160-00010-w1] [Medline: 17938389]
50. Siepe BS, Rieble C, Tutunji R, et al. Understanding EMA data: a tutorial on exploring item performance in ecological momentary assessment data. PsyArXiv. Preprint posted online on Jan 26, 2024. [doi: 10.31234/osf.io/dvj8g]
51. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 1995 Jan 1;57(1):289-300. [doi: 10.1111/j.2517-6161.1995.tb02031.x]

52. Johannes N, Dienlin T, Bakhshi H, Przybylski AK. No effect of different types of media on well-being. Sci Rep 2022 Jan 6;12(1):61. [doi: 10.1038/s41598-021-03218-7] [Medline: 34992220]
53. Lafit G. Non-normal, higher-level, and Var(1) models for the analysis of ESM data. In: Myin-Germeys I, Kuppens P, editors. The Open Handbook of Experience Sampling Methodology: A Step-by-Step Guide to Designing, Conducting, and Analyzing ESM Studies: The Center for Research on Experience Sampling and Ambulatory Methods (REAL); 2021:185-214.
54. Brysbaert M. How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. J Cogn 2019 Jul 19;2(1):16. [doi: 10.5334/joc.72] [Medline: 31517234]
55. Lenth R. emmeans: Estimated Marginal Means, aka Least-Squares Means. The Comprehensive R Archive Network. URL: https://CRAN.R-project.org/package=emmeans [accessed 2024-06-21]
56. Khetawat D, Steele RG. Examining the association between digital stress components and psychological wellbeing: a meta-analysis. Clin Child Fam Psychol Rev 2023 Dec;26(4):957-974. [doi: 10.1007/s10567-023-00440-9] [Medline: 37432506]
57. Highton-Williamson E, Priebe S, Giacco D. Online social networking in people with psychosis: a systematic review. Int J Soc Psychiatry 2015 Feb;61(1):92-101. [doi: 10.1177/0020764014556392] [Medline: 25381145]
58. Naslund JA, Aschbrenner KA, Bartels SJ. How people with serious mental illness use smartphones, mobile apps, and social media. Psychiatr Rehabil J 2016 Dec;39(4):364-367. [doi: 10.1037/prj0000207] [Medline: 27845533]
59. Beyens I, Pouwels JL, van Driel II, Keijsers L, Valkenburg PM. The effect of social media on well-being differs from adolescent to adolescent. Sci Rep 2020 Jul 1;10(1):10763. [doi: 10.1038/s41598-020-67727-7] [Medline: 32612108]
60. Popat A, Tarrant C. Exploring adolescents' perspectives on social media and mental health and well-being - a qualitative literature review. Clin Child Psychol Psychiatry 2023 Jan;28(1):323-337. [doi: 10.1177/13591045221092884] [Medline: 35670473]
61. Valkenburg PM, Peter J. The differential susceptibility to media effects model: differential susceptibility to media effects model. J Commun 2013 Apr;63(2):221-243. [doi: 10.1111/jcom.12024]
62. Abdel-Baki A, Lal S, D-Charron O, Stip E, Kara N. Understanding access and use of technology among youth with first-episode psychosis to inform the development of technology-enabled therapeutic interventions. Early Interv Psychiatry 2017 Feb;11(1):72-76. [doi: 10.1111/eip.12250] [Medline: 26011657]
63. Murayama K, Gfrörer T. Thinking clearly about time-invariant confounders in cross-lagged panel models: a guide for choosing a statistical model from a causal inference perspective. PsyArXiv. Preprint posted online on Oct 15, 2022. [doi: 10.31234/osf.io/bt9xr]
64. Parry DA, Davidson BI, Sewall CJR, Fisher JT, Mieczkowski H, Quintana DS. A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. Nat Hum Behav 2021 Nov;5(11):1535-1547. [doi: 10.1038/s41562-021-01117-5] [Medline: 34002052]

ABBREVIATIONS

DSM-5**:** *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition)
EMA: ecological momentary assessment
PANSS: Positive and Negative Syndrome Scale
RMSSD: root-mean-square of successive differences

Edited by J Torous; submitted 05.04.24; peer-reviewed by A Polillo, AK Chun Chau, J Lee; revised version received 22.05.24; accepted 26.05.24; published 03.07.24.

Please cite as:
Paquin V, Ackerman RA, Depp CA, Moore RC, Harvey PD, Pinkham AE
Media Use and Its Associations With Paranoia in Schizophrenia and Bipolar Disorder: Ecological Momentary Assessment
JMIR Ment Health 2024;11:e59198
URL: https://mental.jmir.org/2024/1/e59198
doi:10.2196/59198

Long-Term Effects of Internet-Based Cognitive Behavioral Therapy on Depression Prevention Among University Students: Randomized Controlled Factorial Trial

Yukako Nakagami[1,2,*], MD, PhD; Teruhisa Uwatoko[2,3,*], MD, PhD; Tomonari Shimamoto[1], PhD; Masatsugu Sakata[4,5], PhD; Rie Toyomoto[4], MSc; Kazufumi Yoshida[4], MD, DrPH; Yan Luo[4], MD, PhD; Nao Shiraishi[6], MD, PhD, MPH; Aran Tajika[4], MD, PhD; Ethan Sahker[4,7], PhD; Masaru Horikoshi[8], PhD; Hisashi Noma[9], PhD; Taku Iwami[1], MD, PhD; Toshi A Furukawa[4,10], MD, PhD

1
2
3
4
5
6
7
8
9
10

*these authors contributed equally

Corresponding Author:
Teruhisa Uwatoko, MD, PhD

Abstract

Background: Internet-based cognitive behavioral therapy (iCBT) shows promise in the prevention of depression. However, the specific iCBT components that contribute to its effectiveness remain unclear.

Objective: We aim to evaluate the effects of iCBT components in preventing depression among university students.

Methods: Using a smartphone cognitive behavioral therapy (CBT) app, we randomly allocated university students to the presence or absence of 5 different iCBT components: self-monitoring, behavioral activation, cognitive restructuring, assertiveness training, and problem-solving. The active intervention lasted 8 weeks but the app remained accessible through the follow-up. The primary outcome was the onset of a major depressive episode (MDE) between baseline and the follow-up after 52 weeks, as assessed with the computerized World Health Organization Composite International Diagnostic Interview. Secondary outcomes included changes in the 9-item Patient Health Questionnaire, 7-item General Anxiety Disorder, and CBT Skills Scale.

Results: During the 12-month follow-up, 133 of 1301 (10.22%) participants reported the onset of an MDE. There were no significant differences in the incidence of MDEs between the groups with or without each component (hazard ratios ranged from 0.85, 95% CI 0.60 - 1.20, for assertiveness training to 1.26, 95% CI 0.88 - 1.79, for self-monitoring). Furthermore, there were no significant differences in the changes on the 9-item Patient Health Questionnaire, 7-item General Anxiety Disorder, or for CBT Skills Scale between component allocation groups. However, significant reductions in depression and anxiety symptoms were observed among all participants at the 52-week follow-up.

Conclusions: In this study, we could not identify any specific iCBT components that were effective in preventing depression or the acquisition of CBT skills over the 12-month follow-up period, but all participants with and without intervention of each iCBT component demonstrated significant improvements in depressive and anxiety symptoms. Further research is needed to explore the potential impact of frequency of psychological assessments, nonspecific intervention effects, natural change in the mental state, and the baseline depression level.

Trial Registration: UMINCTR UMIN000031307; https://center6.umin.ac.jp/cgi-open-bin/ctr_e/ctr_view.cgi?recptno=R000035735

International Registered Report Identifier (IRRID): RR2-10.1186/s13063-018-2719-z

(JMIR Ment Health 2024;11:e56691) doi:10.2196/56691

KEYWORDS

iCBT; depression prevention; student mental health; factorial randomized controlled trial; mobile phone

Introduction

Major depressive disorder (MDD) is one of the most common mental disorders among university students with an estimated 12-month prevalence of 18.5% [1]. MDD can disrupt academic, interpersonal, and social functioning and can lead to suicide [2]. The suicide risk among people with MDD is more than 20 times greater than in the general population [2]. Worldwide, the World Health Organization (WHO) reports that suicide is the fourth leading cause of death among 15-29 year olds [3]. According to Ministry of Health, Labor and Welfare Suicide Countermeasures Promotion Office, in Japan, between 2020 and 2022, over 1000 students per year lost their lives to suicide, and suicide was the leading cause of death among those aged 15 to 34 years [4]. To improve functioning and to reduce suicide risk, preventing major depressive episodes (MDEs) among university students is especially important.

Several psychological interventions have shown promise for preventing depression [5] and cognitive behavioral therapy (CBT)–based strategies make up the bulk of the evidence [6,7]. Further, 1 recent systematic review showed that psychological interventions, mostly CBT-based, reduce the chances of depression incidence by 19% [8]. Furthermore, CBT-based depression prevention interventions are proven efficacious among child and adolescent populations [7]. Additionally, internet-based cognitive behavioral therapy (iCBT) has demonstrated consistent efficaciousness in treating depression [9,10]. iCBT is highly desirable because of its ease of accessibility and lower costs than face-to-face interventions [11].

However, the efficaciousness of iCBT has been examined as a package of several components [9]. In these packages, participants can receive specific modules from a menu of CBT skills, including psychoeducation, self-monitoring (SM), behavioral activation (BA), cognitive restructuring (CR), assertiveness training (AT), and problem-solving (PS). Hence, the efficacy of individual components for preventing depression remains unknown. By focusing on the effective components, iCBT could be conducted more efficiently, further boosting the scalability among university students.

In this large factorial trial, we aimed to identify the efficacious components of iCBT for preventing depression in the long-term. Our previous analysis of acute phase effects on depressive symptoms indicated an improvement in symptoms regardless of iCBT component allocation [12]. However, the long-term effects of each iCBT component on MDE incidence is still unclear. If reducing the occurrence of MDEs is due to specific iCBT components learned over the course of a year, this would indicate the necessary intervention ingredients needed to reduce long-term MDD progression.

Methods

Design

We conducted the Healthy Campus Trial, a fully factorial trial of 5 iCBT components: SM, BA, CR, AT, and PS. This study was designed as a fully factorial randomized controlled trial. The original protocol [13] explains the finer details of the Healthy Campus Trial.

Participants

Undergraduate or graduate students from 5 universities in Japan were recruited between September 2018 and May 2021. Students aged between 18 and 39 years, possessing a smartphone, were introduced to online screening with the 9-item Patient Health Questionnaire (PHQ-9). To examine the effect of smartphone CBT on those with and without mild mental symptoms, based on the screening PHQ-9 scores, we included a random tenth of students scoring 4 or less on the PHQ-9 and all of the subthreshold depressive students with PHQ-9 scores between 5 and 9, or between 10 and 14 with suicidal ideation for less than half of the days. We were able to include only 1 in 10 of the students with no to very mild symptoms because including them all would have made the total sample size 10 times greater, and our budget and manpower could not afford this. Since the focus of the acute phase intervention [12,13] was on those with subthreshold depression, we prioritized their inclusion, while for the long-term outcome, we also wished to examine the effect of the smartphone CBT on those with or without mild symptoms. We excluded students under treatment by any mental health professionals. No student was diagnosed with an MDE at baseline.

Interventions

All participants received a psychoeducation module. With psychoeducation, participants learned the importance of monitoring their own mental state through the trial duration using the smartphone app "Resilience Training," which was developed for this study. After psychoeducation and baseline PHQ-9 assessment, iCBT combinations were randomly allocated for each participant among 32 combinations (with or without each of the 5 iCBT components, hence 2^5=32 combinations, see Multimedia Appendix 1). Participants could be either assigned to or not assigned to each one individually. As a result, some participants might end up with zero iCBT components after the psychoeducation, while others could receive all 5. We used block randomization stratified by participant university and baseline PHQ-9 score (4 or less versus 5 or more). Each component was supposed to take 1 week. All participants, regardless of the number of the allocated components, were expected to finish the app in 8 weeks. Participants work on each component for 7 - 10 days. They were able to review any assigned components as often as they liked during the 8 weeks and fill in as many additional worksheets as they wished. The participants had to complete (read all the lessons and complete at least one worksheet) each component before they could advance to the next component. Participants received digital gift cards worth US $9 for completing self-checks in weeks 4 and 24 and US $18 in weeks 8 and 52. In addition, research staff sent regular emails to encourage all participants to move forward with the program and reminder emails to fill in questionnaires if required. These emails did not include any specific guidance on the content of iCBT.

The contents of the five iCBT components were as follows:

1. SM: Participants monitored their own reactions to different situations (ie, feelings, thoughts, behaviors, and physical reactions) on mind maps and learned the association between each reaction.
2. BA: Participants recorded their activities with facilitation of new challenges through gamification in the "Action Marathon."
3. CR: Participants recorded automatic thoughts and situations in a mind map. They were challenged to broaden their thoughts using several items, including telephone calls from imagined friends, a thought credibility percentage calculator, and a "Serenity Prayer"–modeled self-talk script.
4. AT: Participants learned assertive communication methods and recorded the content of real conversations, the situations, and the people involved.
5. PS: Participants set goals regarding difficult situations, raised ideas for solving associated problems, and evaluated the merits and demerits of each idea.

All CBT component lessons were delivered through the "Resilience Training" app.

Measurements

The primary outcome was time to the first MDE by 52 weeks post randomization. The occurrence of an MDE was evaluated by the computerized version of the WHO Composite International Diagnostic Interview (CIDI) version 3.0 depression section, which has demonstrated good reliability [14,15]. The secondary outcomes included depressive symptoms, anxiety symptoms, and 5 CBT skills, as measured by the PHQ-9 [16,17], 7-item General Anxiety Disorder (GAD-7) [18,19], and CBT Skills Scale [20], respectively. The reliability of each questionnaire is well supported by previous evidence, including our study in which the CBT Skills Scale was developed and validated [20]. The timing for each questionnaire is detailed in the original protocol [13]. For the current analysis, the PHQ-9 was measured at baseline and at weeks 1 through 8, followed by every 4 weeks thereafter, up to week 52. The GAD-7 and the CBT Skills Scale were measured at baseline, and at the fourth, eighth, 24th, and 52nd weeks of this study. The CIDI results were obtained at week 52. To ensure participant safety, those with high levels of depression and suicidal ideation measured by the PHQ-9 were monitored and advised via email to seek professional care. All questionnaires were administered using the Japanese version.

Statistical Analysis

We used SAS (version 9.4, SAS Institute) and R (version 4.2.1, R Foundation for Statistical Computing) for all analyses. We performed Cox regression analyses to investigate whether the presence or absence of each CBT component was related to the occurrence of an MDE. The model involves 5 iCBT components, baseline PHQ-9 scores, university, age, and sex as explanatory variables. Allocation of each iCBT component was defined based on an intention to treat (ITT). We plotted Kaplan-Meier survival curves for a graphical representation of MDE-free survival for each iCBT component. Interval censorings were addressed with Anderson-Bergman adjustment methods [21,22]. In addition, changes in PHQ-9, GAD-7, and CBT Skills Scale were repeatedly measured up to 52 weeks and analyzed with mixed models for repeated measures [23]. The correlation matrix of repeated outcomes was assumed to be unstructured. For explanatory variables, 5 iCBT components, week, iCBT component by week interaction, university, age, and baseline scores were modeled. The between-group effect size was estimated by dividing the estimated mean difference by the observed SD of week 52 scores, which is the most widely used definition of standardized difference in clinical trials of psychiatry [24]. In addition, the within-group effect size was estimated by dividing the estimated mean change by the baseline SD.

Ethical Considerations

This study involves human participants and has been approved by the Ethics Committee of the Kyoto University School of Medicine (C1357). All data obtained from participants were deidentified, and participants provided informed consent before participating in this study.

Results

Demographics

Figure 1 presents the CONSORT diagram. Of the 5063 students assessed for eligibility, 1627 students provided informed consent and were randomly allocated to 1 of the 32 iCBT combinations (Multimedia Appendix 1). Finally, 1626 students were included in this study, after 1 student withdrew consent and refused the use of their data for analysis. Table 1 tabulates the demographic characteristics of the participants. The mean age was 21.5 (SD 3.0) years, and 933 (57.38%) were women. The mean PHQ-9 and GAD-7 scores were 6.4 (SD 3.4) and 5.5 (SD 3.4), respectively. Participants were allocated to each module (SM: n=808; BA: n=817; CR: n=811; AT: n=814; PS: n=811). The characteristics of the participants per component are described in Multimedia Appendix 2, and the number of participants allocated to each of the 32 combinations is described in Multimedia Appendix 1.

Figure 1. CONSORT diagram. CIDI: Composite International Diagnostic Interview; CONSORT: Consolidated Standards of Reporting Trials; IC: informed consent; PHQ-9: 9-item Patient Health Questionnaire.

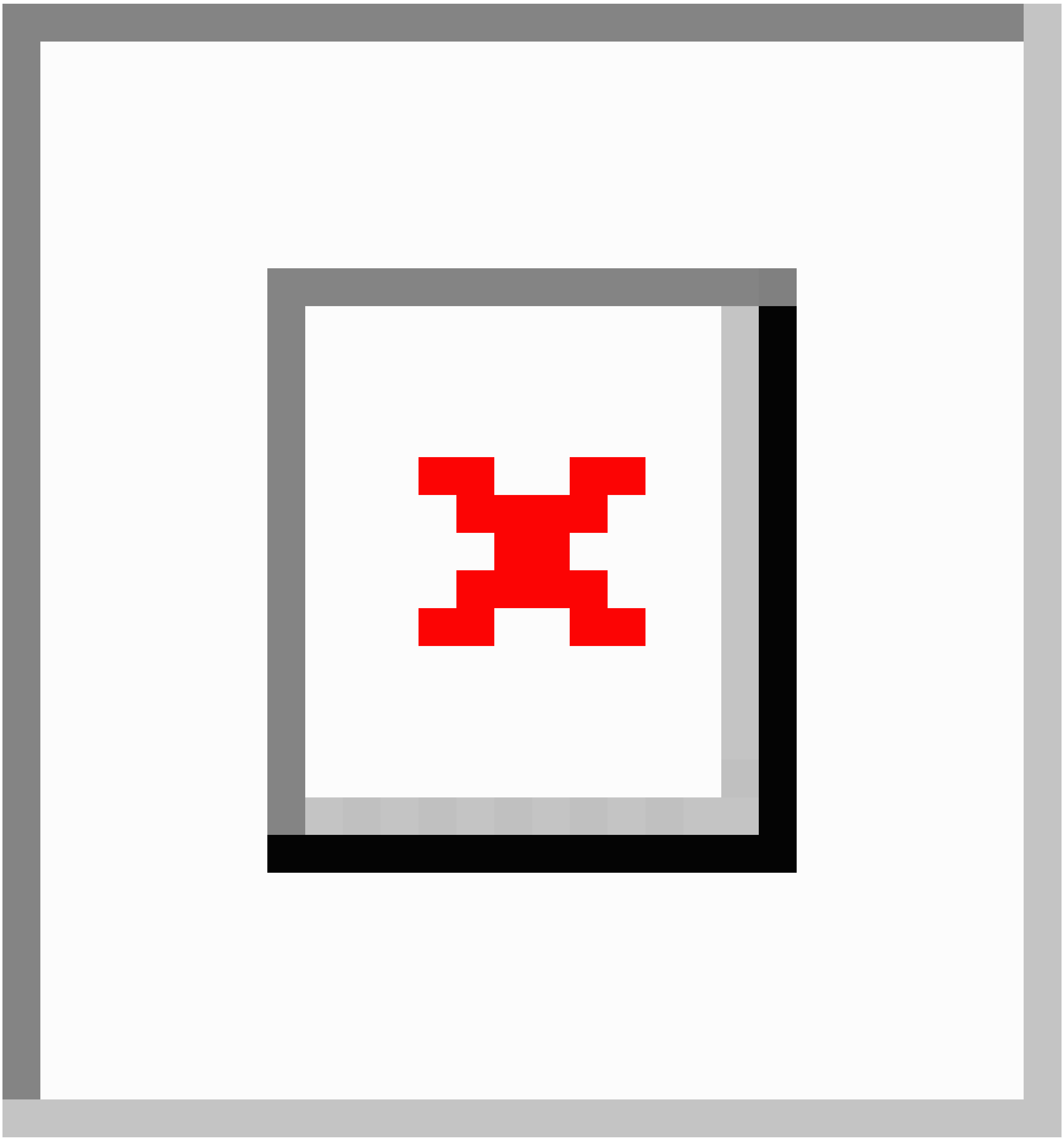


Table . Baseline characteristics of total participants for 1-year analysis (N=1626).

	Participants
Demographics	
Age (years), mean (SD)	21.5 (2.99)
Sex (female), n (%)	933 (57.38)
Undergraduate, n (%)	1250 (76.88)
Married, n (%)	33 (2.03)
Living alone (apart from family), n (%)	1002 (61.62)
Part-time employment, n (%)	1261 (77.55)
Smoking experience, n (%)	94 (5.78)
Drinking alcohol regularly, n (%)	665 (40.90)
Exercise opportunity, n (%)	1014 (62.36)
History of psychiatric or psychological treatment, n (%)	184 (11.32)
History of major depressive episode in past year (CIDI[a]), n (%)	151 (9.29)
Cognitive and behavioral skills, mean (SD)	
SM[b] skills	8.03 (3.26)
BA[c] skills	10.53 (4.33)
CR[d] skills	8.71 (3.44)
AT[e] skills	9.60 (3.55)
PS[f] skills	10.6 (3.09)
Clinical characteristic, mean (SD)	
PHQ-9[g]	6.38 (3.40)
GAD-7[h]	5.47 (3.35)
Function, mean (SD)	
WHO-HPQ[i] Presenteeism	6.30 (1.58)

[a]CIDI: Composite International Diagnostic Interview.
[b]SM: self-monitoring.
[c]BA: behavioral activation.
[d]CR: cognitive restructuring.
[e]AT: assertiveness training.
[f]PS: problem-solving.
[g]PHQ-9: 9-item Patient Health Questionnaire.
[h]GAD-7: 7-item General Anxiety Disorder.
[i]WHO-HPQ: World Health Organization Health and Work Performance Questionnaire.

Primary Analysis

Of the 1626 study participants, 1301 (80.01%) responded to the 52-week follow-up survey. Among the responders, 133 (10.22%) reported incident MDE by week 52. The number of completions and their percentages for each assigned module were recorded (SM: 743/808, 92.6%; BA: 703/817, 86%; CR: 688/811, 84.8%; AT: 680/814, 83.8%; PS: 666/811, 82.1%). Table 2 summarizes the results of the multivariable Cox regression analysis of each CBT component on MDE occurrence. The hazard ratios (95% CI) of presence over absence were 1.26 (0.88 - 1.79) for SM, 1.09 (0.77 - 1.54) for BA, 1.18 (0.83 - 1.68) for CR, 0.85 (0.60 - 1.20) for AT, and 1.18 (0.82 - 1.68) for PS. Multimedia Appendix 3 shows the full results of the analysis with coefficients for the included covariates. The results of the univariable Cox regression analysis (Multimedia Appendix 4) showed similar results to those of the multivariable analyses. Figure 2 presents the Kaplan-Meier survival curves for the occurrence of an MDE for each of the 5 iCBT components. We did not find a significant difference in the occurrence of an MDE in each iCBT component.

Table . Summary of the results of multiple cox regression analyses of each CBT[a] component on occurrence of an MDE[b] the results of the analysis for each component are shown in Multimedia Appendix 3.[c]

Component	Cases with MDE, n/N (completed cases)		HR[d] for incident MDE			
	Presence	Absence	HR	SE	95% CI	*P* value
SM[e]	75/808 (638)	58/818 (663)	1.26	0.18	0.88-1.79	.21
BA[f]	69/817 (648)	64/809 (653)	1.09	0.18	0.77-1.54	.63
CR[g]	72/811 (652)	61/815 (649)	1.18	0.18	0.83-1.68	.37
AT[h]	62/814 (647)	71/812 (654)	0.85	0.18	0.60-1.20	.35
PS[i]	71/811 (646)	62/815 (655)	1.18	0.18	0.82-1.68	.37

[a]CBT: cognitive behavioral therapy.

[b]MDE: major depressive episode

[c]Age, sex, baseline 9-item Patient Health Questionnaire point, and recruition site (4 universities) were adjusted in the Cox regression analysis.

[d]HR: hazard ratio.

[e]SM: self-monitoring.

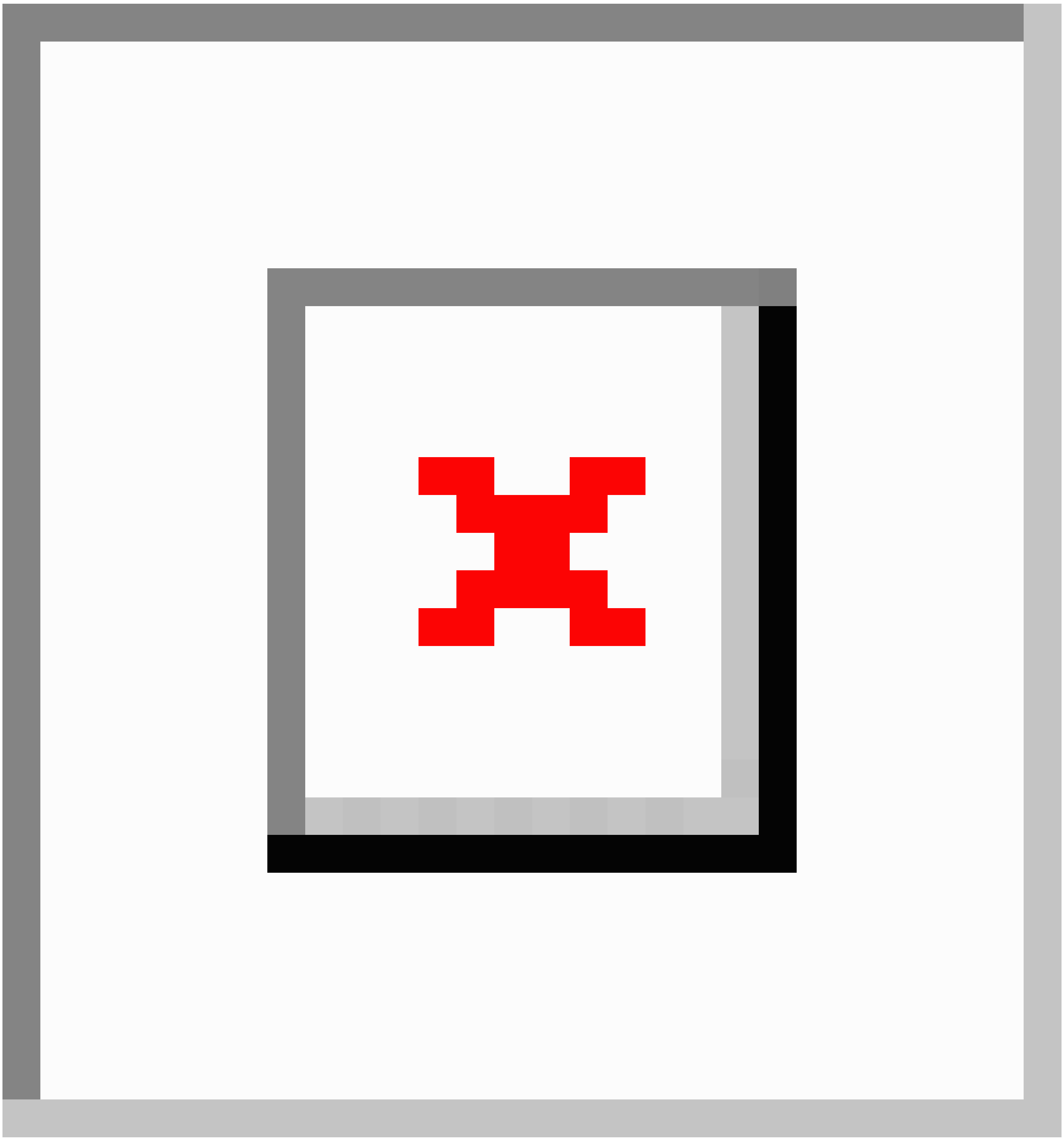
[f]BA: behavioral activation.

[g]CR: cognitive restructuring.

[h]AT: assertiveness training.

[i]PS: problem-solving.

Figure 2. Survival curves for MDEs for the 5 CBT components. AT: assertiveness training; BA: behavioral activation; CBT: cognitive behavioral therapy; CR: cognitive restructuring; MDE: major depressive episode; PS: problem-solving; S(t): survival function (major depressive episode–free survival); SM: self-monitoring.



Secondary Analysis

Table 3 shows the estimated least squares mean change scores of the PHQ-9 for participants allocated to the presence or absence of each component using the mixed models for repeated measures. Compared to the baseline, the estimated least squares mean change scores ranged between −1.77 and −1.97 at the 52-week follow-up for both the presence and absence groups across all iCBT components. Regardless of the allocated iCBT component, average depressive symptoms were reduced from baseline. However, no significant differences in change scores were found in the comparison between the presence group and the absence group across all iCBT components. Analysis of the GAD-7 showed similar results for anxiety symptoms, as shown in Table 4. Table 5 shows an analysis of the relevant CBT skills at week 52 for each CBT component. There was no difference in changes in CBT skills between the presence or absence of each iCBT component.

Table . Summary of the results of differences of change scores of PHQ-9[a] at 52 weeks using MMRM[b] CBT[c] components, week, CBT component by week interaction, university, age, and baseline scores as fixed effects, while participants were introduced as a random effect. Effect size was estimated by division of the estimated mean changes by the observed SD of baseline scores.

Component		Participants, n	Estimated least squares mean change scores of PHQ-9 within groups (95% CI)	Within-group effect size for baseline-week 52 change scores of PHQ-9 (95% CI)	Estimated differences of baseline-week 52 change scores between groups (95% CI)	Effect size for differences of baseline-week 52 change scores between groups (95% CI)	*P* values
SM [d]					−0.09 (−0.28 to 0.4)	−0.02 (−0.06 to 0.09)	.81
	Presence	808	−1.91 (−2.43 to −1.39)	−0.69 (−0.87 to −0.50)			
	Absence	818	−1.83 (−2.35 to −1.31)	−0.66 (−0.84 to −0.47)			
BA [e]					−0.14 (−0.43 to 0.25)	−0.03 (−0.09 to 0.05)	.71
	Presence	817	−1.94 (−2.46 to −1.41)	−0.70 (−0.88 to −0.51)			
	Absence	809	−1.8 (−2.32 to −1.28)	−0.65 (−0.83 to −0.46)			
CR [f]					0.21 (−0.29 to 0.39)	0.04 (−0.06 to 0.08)	.58
	Presence	811	−1.77 (−2.28 to −1.25)	−0.64 (−0.82 to −0.45)			
	Absence	815	−1.97 (−2.50 to −1.45)	−0.71 (−0.90 to −0.52)			
AT [g]					0.21 (−0.16 to 0.52)	0.04 (−0.04 to 0.11)	.57
	Presence	814	−1.77 (−2.29 to −1.24)	−0.64 (−0.82 to −0.45)			
	Absence	812	−1.97 (−2.49 to −1.45)	−0.71 (−0.90 to −0.52)			
PS [h]					−0.09 (−0.30 to 0.38)	−0.02 (−0.06 to 0.08)	.80
	Presence	811	−1.92 (−2.43 to −1.40)	−0.69 (−0.87 to −0.50)			
	Absence	815	−1.82 (−2.35 to −1.3)	−0.65 (−0.84 to −0.47)			

[a]PHQ-9: 9-item Patient Health Questionnaire.
[b]MMRM: mixed models for repeated measures.
[c]CBT: cognitive behavioral therapy.
[d]SM: self-monitoring.
[e]BA: behavioral activation.
[f]CR: cognitive restructuring.
[g]AT: assertiveness training.
[h]PS: problem-solving.

Table . Summary of the results of differences of change scores of GAD-7[a] at 52 weeks using MMRM[b] CBT[c] components, week, CBT component by week interaction, university, age, and baseline scores as fixed effects, while participants were introduced as a random effect. Effect size was estimated by division of the estimated mean changes by the observed SD of baseline scores.

Component		Participants, n	Estimated least squares mean change scores of GAD-7 within groups (95% CI)	Within-group effect size for baseline-week 52 change score of GAD-7 (95% CI)	Estimated differences of baseline-week 52 change scores between groups (95% CI)	Effect size for differences of baseline-week 52 change scores between groups (95% CI)	*P* values
SM [d]					0.27 (−0.46 to 1)	0.08 (−0.14 to 0.3)	.47
	Presence	808	−1.28 (−1.82 to −0.74)	−0.38 (−0.55 to −0.22)			
	Absence	818	−1.55 (−2.08 to −1.02)	−0.47 (−0.62 to −0.31)			
BA [e]					−0.32 (−1.04 to 0.41)	−0.09 (−0.31 to 0.12)	.39
	Presence	817	−1.57 (−2.11 to −1.04)	−0.47 (−0.63 to −0.31)			
	Absence	809	−1.26 (−1.79 to −0.72)	−0.38 (−0.54 to −0.22)			
CR [f]					0.30 (−0.43 to 1.02)	0.09 (−0.13 to 0.31)	.42
	Presence	811	−1.27 (−1.79 to −0.74)	−0.38 (−0.54 to −0.22)			
	Absence	815	−1.56 (−2.11 to −1.02)	−0.47 (−0.63 to −0.31)			
AT [g]					0.14 (−0.59 to 0.86)	0.04 (−0.18 to 0.26)	.71
	Presence	814	−1.35 (−1.89 to −0.81)	−0.41 (−0.57 to −0.24)			
	Absence	812	−1.48 (−2.01 to −0.95)	−0.44 (−0.60 to −0.29)			
PS [h]					−0.01 (−0.73 to 0.72)	0 (−0.22 to 0.22)	.99
	Presence	811	−1.42 (−1.95 to −0.88)	−0.43 (−0.59 to −0.26)			
	Absence	815	−1.41 (−1.95 to −0.87)	−0.42 (−0.59 to −0.26)			

[a]GAD-7: 7-item General Anxiety Disorder.
[b]MMRM: mixed models for repeated measures.
[c]CBT: cognitive behavioral therapy.
[d]SM: self-monitoring.
[e]BA: behavioral activation.
[f]CR: cognitive restructuring.
[g]AT: assertiveness training.
[h]PS: problem-solving.

Table . Summary of the estimated least squares mean differences of change in corresponding CBT[a] skills at 52 weeks using MMRM[b] in addition to CBT component, week, CBT component by week interaction, university, age, and baseline scores as fixed effects, while participants were introduced as a random effect.

Component	Estimated least squares mean differences of change in corresponding CBT skill (95% CI)
SM[c]	0.23 (−0.28 to 0.75)
BA[d]	−0.44 (−1.10 to 0.22)
CR[e]	−0.1 (−0.58 to 0.38)
AT[f]	0.14 (−0.39 to 0.66)
PS[g]	0.24 (−0.15 to 0.64)

[a]CBT: cognitive behavioral therapy.
[b]MMRM: mixed models for repeated measures.
[c]SM: self-monitoring.
[d]BA: behavioral activation.
[e]CR: cognitive restructuring.
[f]AT: assertiveness training.
[g]PS: problem-solving.

Discussion

Principal Findings

We conducted a fully factorial randomized trial of 5 iCBT skills to reduce the incidence of MDEs among university students. We did not observe significant preventive effects attributable to individual iCBT components in depression prevention. Secondary analyses showed no significant difference in change scores for PHQ-9, GAD-7, and CBT skills, regardless of the presence or absence of each iCBT component either. Participants exhibited improvements in depression and anxiety symptoms from the baseline to the 52 weeks, irrespective of the specific iCBT component allocation.

The significantly better PHQ-9 and GAD-7 scores compared to baseline at 52 weeks, irrespective of the iCBT component, replicate the results from our previous 8-week short-term analysis [12]. Multiple extra-intervention factors may have contributed to the general improvement effects. First, the frequent psychological assessments administered to all participants may have introduced a significant nonspecific therapeutic effect. This hypothesis is supported by findings from a prior meta-analysis of antidepressant efficacy trials, which demonstrated that follow-up assessments significantly influence placebo responses [25]. In this study, frequent follow-up visits significantly impacted depressive scale scores, where the extra follow-up visits accounted for approximately 34% - 44% of the overall placebo response observed over the course of the trials. These results emphasize that the therapeutic effect of these assessments is not only statistically significant but also cumulative and proportional to the number of interactions, thereby suggesting a potent nonspecific therapeutic effect stemming from the assessment frequency itself [25].

In our study, a rigorous schedule of assessments was implemented, occurring as many as 19 times throughout the duration of this study (specifically at weeks 1-8, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48, and 52). This high frequency of assessments might have played a significant role in mediating a therapeutic effect, potentially confounding the distinct impacts attributable to each individual iCBT component. It is important to note, however, that all participants in this study were subjected to the same number of assessments, making it challenging to isolate the effects of the assessments from those of the iCBT interventions themselves. Consequently, while it is tempting to speculate about the influence of frequent assessments, such speculation cannot be conclusively confirmed given the uniform assessment protocol applied across all participants.

Given these considerations, it is recommended that future research should explore the implications of varying the frequency of assessments by including experimental groups subjected to differing assessment schedules. This approach would allow for a more refined analysis of how frequent interactions might influence outcomes, thereby enabling a more accurate assessment of the true efficacy of the iCBT components. By distinguishing the effects of the intervention from those of the assessment frequency, future studies could yield more reliable and generalizable findings.

In addition to frequent assessments, application help notifications coupled with encouragement emails may have produced extra-intervention improvement effects. Regardless of allocated iCBT components, participants were instructed to keep the app installed until the end of the 52-week follow-up. This instruction also conveyed crisis management information, such as where to go for help in case of serious problems. Providing emergency contact information was reported to reduce rates of repeated self-harm [26]. Thus, it is possible that the crisis management information in our study could have contributed positively to mental health conditions. Future studies should consider incorporating groups with different frequencies of self-assessments to derive more reliable results.

Together with the frequent follow-up assessments and application help notifications, university grade level could have affected the results. A longitudinal study reported that depressive and anxiety symptoms were better in the fourth year of university than in the first 3 years [27]. This trend toward improvement may mask the effect of our iCBT intervention. Similarly, baseline depression level may have affected our results. As depression symptoms become milder, placebo effects become more prominent relative to the intervention. This assertion is supported by a meta-analysis examining the effects of antidepressants compared to placebo [28]. If we targeted university students with more severe depression symptoms, we may have a greater ability to detect intervention effects. On the other hand, another study assessing the effects of an iCBT program with healthy participants showed a significant improvement in subthreshold depressive symptoms [10]. Therefore, our failure to identify the effects of the intervention may not be attributed solely to the depressive severity of the participants.

The latter half of our study unfolded during the COVID-19 pandemic, accompanied by extensive social restrictions. This period has been scrutinized across various studies, notably in a systematic review of epidemiological research which reported a temporary surge in depression and anxiety within the first 3 months of the pandemic [29]. However, our own secondary analysis, using the same data set [30], did not find a statistically significant link between depressive symptoms, as measured by the PHQ-9, and the implementation of 4 distinct states of emergency in Japan aimed at mitigating the spread of the virus. Although our results suggest no direct correlation between these emergency measures and depressive symptoms, the overarching impact of the pandemic—its indirect influence on mental health through other unmeasured factors—remains a plausible explanation for variations in our findings. Consequently, we acknowledge the complexity of the pandemic's effects, which may interact with psychological outcomes in nuanced and varied ways that are not immediately apparent in our data.

The validated CBT Skills Scale [20] was used to measure the main 5 CBT skills. The results at 52 weeks showed no significant difference in change scores with or without each of the iCBT components. This finding was unexpected as our previously reported short-term analysis revealed significant improvement of CR and AT skills [12]. The 52-week results lack of significant differences may be due to the attenuation of skills learned during the first 8-week intervention period. However, some studies reported that the effect of CBT lasts longer than 46 months as an adjunct to pharmacotherapy for treatment-resistant depression [31]. These studies focused only on symptoms, not on CBT skills. Our report is the first to assess the persistence of CBT skills using our original CBT Skills Scale. Further research is needed to determine how long CBT skills are usually maintained. It is possible that maintenance CBT may be effective in maintaining CBT skills, despite the reported fact that maintenance CBT on depression prevention is superior to psychoeducation only among patients with a high risk of depression relapse [32].

Limitations

There are several limitations in our study. First, our participant pool was limited to 2 prefectures in Japan (ie, Kyoto and Aichi). Nonetheless, this well-powered multisite trial experienced minimal attrition and included 3 types of universities, national, public, and private, which attracted students from all regions of Japan. Additionally, we ensured adequate allocation concealment during randomization, resulting in a well-distributed sample. Nevertheless, caution is advised when generalizing these results. Second, our study did not account for potentially significant protective factors, such as good academic performance and financial stability. Although our research used a fully factorial design intended to evenly distribute unknown confounders across study groups, we cannot definitively exclude the influence of unmeasured variables. Third, our measures of MDEs, mental state, and CBT skills relied on self-administered questionnaires that, despite their established reliability, are susceptible to biases such as social-desirability. This susceptibility potentially obscures the true symptoms of depression, and the accuracy of the PHQ-9 may also be compromised by the repeated measurement effect. To mitigate these limitations and enhance assessment accuracy, face-to-face evaluations conducted by trained professionals would be more ideal. Finally, the validated CBT Skills Scale used in our study may not have been sensitive enough to capture subtle differences in skill acquisition and retention over the extensive follow-up period. While the scale is robust for assessing general CBT competency, it may not adequately reflect incremental changes or the long-term retention of specific skills. Future research should use such tools to ensure a more precise and sensitive evaluation of the long-term skill acquisition and retention of CBT skills.

Conclusions

This study did not show the specific impacts of individual components of iCBT on preventing MDEs. However, participants' scores on the PHQ-9 and GAD-7 generally improved over the year during their involvement in this study. In addition to the statistical phenomenon of regression to the mean and the natural course of the conditions, the psychoeducation given to all participants, the frequent psychological assessments, and the experience of being involved in this study itself may have contributed to some alleviation of depression and anxiety symptoms among university students. Future iCBT optimization trials should consider psychological assessment frequency, the impact of common intervention elements, the natural history of the mental state of the target group, and the baseline depression level.

Acknowledgments

This work was supported in part by JST SPRING (grant JPMJSP2110) and the World Health Organization Centre for Health Development (WHO Kobe Centre—WKC: K22001). Support was also given through grant-in-aid from Japan Society for the Promotion of Science to MS (21K03049), RT (22K21171), and TI (18K18643); Japan Agency for Medical Research and

Development (AMED; dk0307085) and WHO Mental Health Research (2021/HQ/WKC/0013) to TAF; and Suzuken Memorial Foundation, KDDI Foundation, and Pfizer Health Research Foundation to TU.

Conflicts of Interest

TU reports receiving lecture fees from Sumitomo Pharma and Shionogi Pharma. MS reports personal fees from SONY outside of this submitted work. NS reports receiving lecture fees from Mochida Pharma and Sumitomo Pharma. AT reports personal fees from Sumitomo Dainippon Pharma, Eisai, Janssen Pharmaceutical, Meiji-Seika Pharma, Mitsubishi Tanabe Pharma, Otsuka, Takeda Pharmaceutical, and Shionogi. HN received a research fund from GlaxoSmithKline and consulting fees from GlaxoSmithKline, Ono Pharmaceutical, Sony, and Kowa, outside of this submitted work. TAF reports personal fees from Boehringer-Ingelheim, Daiichi Sankyo, DT Axis, Kyoto University Original, Shionogi, SONY, and UpToDate, and a grant from DT Axis and Shionogi, outside of this submitted work. In addition, TAF has a patent 7448125 and a pending patent 2022-082495 and has licensed intellectual properties for Kokoro-app to Mitsubishi-Tanabe.

Multimedia Appendix 1
Combinations of cognitive and behavioral therapy skill components of cognitive behavioral therapy smartphone app.
[PDF File, 54 KB - mental_v11i1e56691_app1.pdf]

Multimedia Appendix 2
Baseline characteristics of total participants for 1-year analysis (N=1626) and by each component.
[PDF File, 70 KB - mental_v11i1e56691_app2.pdf]

Multimedia Appendix 3
Results of multiple Cox regression analyses.
[PDF File, 129 KB - mental_v11i1e56691_app3.pdf]

Multimedia Appendix 4
Results of simple Cox regression analyses.
[PDF File, 96 KB - mental_v11i1e56691_app4.pdf]

Checklist 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File, 1145 KB - mental_v11i1e56691_app5.pdf]

References

1. Auerbach RP, Mortier P, Bruffaerts R, et al. WHO world mental health surveys international college student project: prevalence and distribution of mental disorders. J Abnorm Psychol 2018 Oct;127(7):623-638. [doi: 10.1037/abn0000362] [Medline: 30211576]
2. Lépine JP, Briley M. The increasing burden of depression. Neuropsychiatr Dis Treat 2011;7(Suppl 1):3-7. [doi: 10.2147/NDT.S19617] [Medline: 21750622]
3. Suicide worldwide in 2019: global health estimates World Health Organization. 2021. URL: https://iris.who.int/bitstream/handle/10665/341728/9789240026643-eng.pdf?sequence=1 [accessed 2024-08-06]
4. Ministry of Health, Labour and Welfare 2022 white paper on suicide prevention. 2022. URL: https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/hukushi_kaigo/seikatsuhogo/jisatsu/jisatsuhakusyo2022.html [accessed 2024-08-06]
5. Caldwell DM, Davies SR, Hetrick SE, et al. School-based interventions to prevent anxiety and depression in children and young people: a systematic review and network meta-analysis. Lancet Psychiatry 2019 Dec;6(12):1011-1020. [doi: 10.1016/S2215-0366(19)30403-1] [Medline: 31734106]
6. Hetrick SE, Cox GR, Witt KG, Bir JJ, Merry SN. Cognitive behavioural therapy (CBT), third-wave CBT and interpersonal therapy (IPT) based interventions for preventing depression in children and adolescents. Cochrane Database Syst Rev 2016 Aug 9;2016(8):CD003380. [doi: 10.1002/14651858.CD003380.pub4] [Medline: 27501438]
7. Zhang Q, Wang J, Neitzel A. School-based mental health interventions targeting depression or anxiety: a meta-analysis of rigorous randomized controlled trials for school-aged children and adolescents. J Youth Adolesc 2023 Jan;52(1):195-217. [doi: 10.1007/s10964-022-01684-4] [Medline: 36229755]
8. Cuijpers P, Pineda BS, Quero S, et al. Psychological interventions to prevent the onset of depressive disorders: a meta-analysis of randomized controlled trials. Clin Psychol Rev 2021 Feb;83:101955. [doi: 10.1016/j.cpr.2020.101955] [Medline: 33333441]

9. Furukawa TA, Suganuma A, Ostinelli EG, et al. Dismantling, optimising, and personalising internet cognitive behavioural therapy for depression: a systematic review and component network meta-analysis using individual participant data. Lancet Psychiatry 2021 Jun;8(6):500-511. [doi: 10.1016/S2215-0366(21)00077-8] [Medline: 33957075]
10. Karyotaki E, Efthimiou O, Miguel C, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021 Apr 1;78(4):361-371. [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
11. Baumann M, Stargardt T, Frey S. Cost-utility of internet-based cognitive behavioral therapy in unipolar depression: a Markov model simulation. Appl Health Econ Health Policy 2020 Aug;18(4):567-578. [doi: 10.1007/s40258-019-00551-x] [Medline: 32060822]
12. Sakata M, Toyomoto R, Yoshida K, et al. Components of smartphone cognitive-behavioural therapy for subthreshold depression among 1093 university students: a factorial trial. Evid Based Ment Health 2022 Dec;25(e1):e18-e25. [doi: 10.1136/ebmental-2022-300455] [Medline: 35577537]
13. Uwatoko T, Luo Y, Sakata M, et al. Healthy campus trial: a multiphase optimization strategy (MOST) fully factorial trial to optimize the smartphone cognitive behavioral therapy (CBT) app for mental health promotion among university students: study protocol for a randomized controlled trial. Trials 2018 Jul 4;19(1):353. [doi: 10.1186/s13063-018-2719-z] [Medline: 29973252]
14. Haro JM, Arbabzadeh-Bouchez S, Brugha TS, et al. Concordance of the ComposIte International Diagnostic Interview Version 3.0 (CIDI 3.0) with standardized clinical assessments in the WHO world mental health surveys. Int J Methods Psychiatr Res 2006;15(4):167-180. [doi: 10.1002/mpr.196] [Medline: 17266013]
15. Shimoda H, Inoue A, Tsuno K, Kawakami N. One-year test-retest reliability of a Japanese web-based version of the WHO composite international diagnostic interview (CIDI) for major depression in a working population. Int J Methods Psychiatr Res 2015 Sep;24(3):204-212. [doi: 10.1002/mpr.1437] [Medline: 24677699]
16. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
17. Muramatsu K, Miyaoka H, Kamijima K, et al. The Patient Health Questionnaire, Japanese version: validity according to the Mini-International Neuropsychiatric Interview-Plus. Psychol Rep 2007 Dec;101(3 Pt 1):952-960. [doi: 10.2466/pr0.101.3.952-960] [Medline: 18232454]
18. Muramatsu K, Muramatsu Y, Miyaoka H. Validation and utility of a Japanese version of the GAD-7. In: Panminerva Medica 20th World Congress on Psychosomatic Medicine Abstracts Book: CiNii Research; 2009:79.
19. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
20. Sakata M, Toyomoto R, Yoshida K, et al. Development and validation of the Cognitive Behavioural Therapy Skills Scale among college students. Evid Based Ment Health 2021 May;24(2):70-76. [doi: 10.1136/ebmental-2020-300217] [Medline: 33402377]
21. Anderson-Bergman C. icenReg: regression models for interval censored data in R. J Stat Softw 2017 Nov;81(12):1-12. [doi: 10.18637/jss.v081.i12]
22. Anderson-Bergman C. An efficient implementation of the EMICM algorithm for the interval censored NPMLE. J Comput Graph Stat 2017 Apr 3;26(2):463-467. [doi: 10.1080/10618600.2016.1208616]
23. Mallinckrodt CH, Clark WS, David SR. Accounting for dropout bias using mixed-effects models. J Biopharm Stat 2001;11(1-2):9-21. [doi: 10.1081/BIP-100104194] [Medline: 11459446]
24. Luo Y, Funada S, Yoshida K, Noma H, Sahker E, Furukawa TA. Large variation existed in standardized mean difference estimates using different calculation methods in clinical trials. J Clin Epidemiol 2022 Sep;149:89-97. [doi: 10.1016/j.jclinepi.2022.05.023] [Medline: 35654267]
25. Posternak MA, Zimmerman M. Therapeutic effect of follow-up assessments on antidepressant and placebo response rates in antidepressant efficacy trials: meta-analysis. Br J Psychiatry 2007 Apr;190:287-292. [doi: 10.1192/bjp.bp.106.028555] [Medline: 17401033]
26. Scott A, Guo B. For which strategies of suicide prevention is there evidence of effectiveness? World Health Organization. 2012. URL: https://iris.who.int/bitstream/handle/10665/363870/9789289057233-eng.pdf?sequence=1&isAllowed=y [accessed 2024-08-07]
27. Gao W, Ping S, Liu X. Gender differences in depression, anxiety, and stress among college students: A longitudinal study from China. J Affect Disord 2020 Feb 15;263:292-300. [doi: 10.1016/j.jad.2019.11.121] [Medline: 31818792]
28. Stone MB, Yaseen ZS, Miller BJ, Richardville K, Kalaria SN, Kirsch I. Response to acute monotherapy for major depressive disorder in randomized, placebo controlled trials submitted to the US Food And Drug Administration: individual participant data analysis. BMJ 2022 Aug 2;378:e067606. [doi: 10.1136/bmj-2021-067606] [Medline: 35918097]
29. Salanti G, Peter N, Tonia T, et al. The impact of the COVID-19 pandemic and associated control measures on the mental health of the general population: a systematic review and dose-response meta-analysis. Ann Intern Med 2022 Nov;175(11):1560-1571. [doi: 10.7326/M22-1507] [Medline: 36252247]

30. Shiraishi N, Sakata M, Toyomoto R, et al. Dynamics of depressive states among university students in Japan during the COVID-19 pandemic: an interrupted time series analysis. Ann Gen Psychiatry 2023 Oct 10;22(1):38. [doi: 10.1186/s12991-023-00468-9] [Medline: 37814328]
31. Wiles NJ, Thomas L, Turner N, et al. Long-term effectiveness and cost-effectiveness of cognitive behavioural therapy as an adjunct to pharmacotherapy for treatment-resistant depression in primary care: follow-up of the CoBalT randomised controlled trial. Lancet Psychiatry 2016 Feb;3(2):137-144. [doi: 10.1016/S2215-0366(15)00495-2] [Medline: 26777773]
32. Stangier U, Hilling C, Heidenreich T, et al. Maintenance cognitive-behavioral therapy and manualized psychoeducation in the treatment of recurrent depression: a multicenter prospective randomized controlled trial. Am J Psychiatry 2013 Jun;170(6):624-632. [doi: 10.1176/appi.ajp.2013.12060734] [Medline: 23732968]

Abbreviations

AT: assertiveness training
BA: behavioral activation
CBT: cognitive behavioral therapy
CR: cognitive restructuring
GAD-7: 7-item General Anxiety Disorder
iCBT: internet-based cognitive behavioral therapy
ITT: intention-to-treat
MDD: major depressive disorder
MDE: major depressive episode
PHQ-9: 9-item Patient Health Questionnaire
PS: problem-solving
SM: self-monitoring
WHO: World Health Organization

Edited by J Torous; submitted 24.01.24; peer-reviewed by J Wang, P Frazier; revised version received 11.06.24; accepted 12.06.24; published 24.09.24.

Please cite as:
Nakagami Y, Uwatoko T, Shimamoto T, Sakata M, Toyomoto R, Yoshida K, Luo Y, Shiraishi N, Tajika A, Sahker E, Horikoshi M, Noma H, Iwami T, Furukawa TA
Long-Term Effects of Internet-Based Cognitive Behavioral Therapy on Depression Prevention Among University Students: Randomized Controlled Factorial Trial
JMIR Ment Health 2024;11:e56691
URL: https://mental.jmir.org/2024/1/e56691
doi:10.2196/56691

Digital Phenotyping of Mental and Physical Conditions: Remote Monitoring of Patients Through RADAR-Base Platform

Zulqarnain Rashid[1], PhD; Amos A Folarin[1,2,3,4,5], PhD; Yuezhou Zhang[1], PhD; Yatharth Ranjan[1], MSc; Pauline Conde[1], BSc; Heet Sankesara[1], BSc; Shaoxiong Sun[1,6], PhD; Callum Stewart[1], PhD; Petroula Laiou[1], PhD; Richard J B Dobson[1,2,3,4,5], PhD

1
2
3
4
5
6

Corresponding Author:
Amos A Folarin, PhD

Abstract

Background: The use of digital biomarkers through remote patient monitoring offers valuable and timely insights into a patient's condition, including aspects such as disease progression and treatment response. This serves as a complementary resource to traditional health care settings leveraging mobile technology to improve scale and lower latency, cost, and burden.

Objective: Smartphones with embedded and connected sensors have immense potential for improving health care through various apps and mobile health (mHealth) platforms. This capability could enable the development of reliable digital biomarkers from long-term longitudinal data collected remotely from patients.

Methods: We built an open-source platform, RADAR-base, to support large-scale data collection in remote monitoring studies. RADAR-base is a modern remote data collection platform built around Confluent's Apache Kafka to support scalability, extensibility, security, privacy, and quality of data. It provides support for study design and setup and active (eg, patient-reported outcome measures) and passive (eg, phone sensors, wearable devices, and Internet of Things) remote data collection capabilities with feature generation (eg, behavioral, environmental, and physiological markers). The back end enables secure data transmission and scalable solutions for data storage, management, and data access.

Results: The platform has been used to successfully collect longitudinal data for various cohorts in a number of disease areas including multiple sclerosis, depression, epilepsy, attention-deficit/hyperactivity disorder, Alzheimer disease, autism, and lung diseases. Digital biomarkers developed through collected data are providing useful insights into different diseases.

Conclusions: RADAR-base offers a contemporary, open-source solution driven by the community for remotely monitoring, collecting data, and digitally characterizing both physical and mental health conditions. Clinicians have the ability to enhance their insight through the use of digital biomarkers, enabling improved prevention, personalization, and early intervention in the context of disease management.

(JMIR Ment Health 2024;11:e51259) doi:10.2196/51259

KEYWORDS

digital biomarkers; mHealth; mobile apps; Internet of Things; remote data collection; wearables; real-time monitoring; platform; biomarkers; wearable; smartphone; data collection; open-source platform; RADAR-base; phenotyping; mobile phone; IoT

Introduction

Background

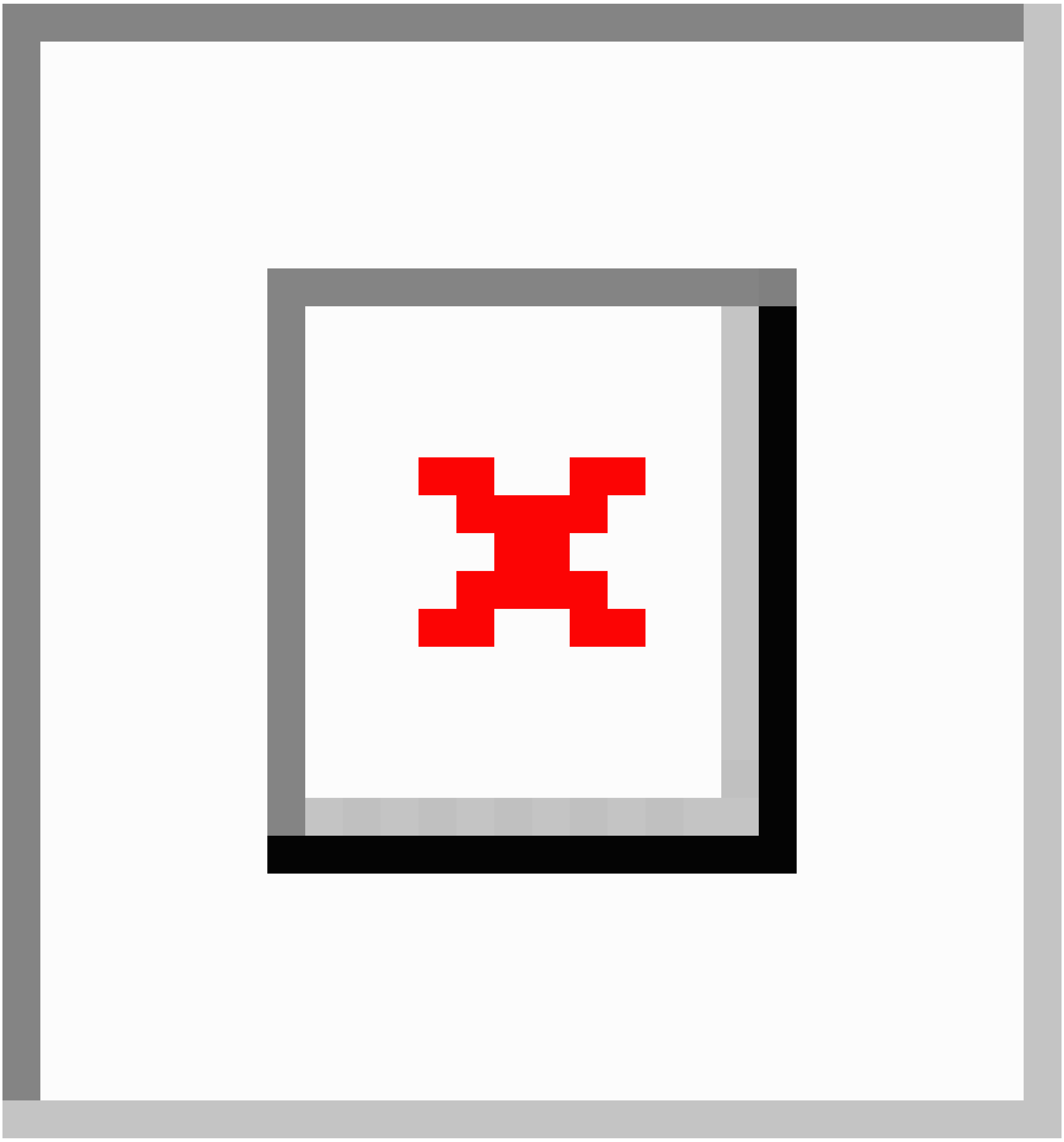
Digital biomarkers offer a host of advantages for measuring our health over traditional assessment approaches that are typically confined to clinical settings, including decentralization, scalability, sampling frequency and real-time measurement, and affordability. However, significant challenges remain with implementing digital biomarkers.

Digital biomarkers developed from sensor data can help with prevention and early intervention to better diagnose and manage disease. Collected data should be reliable and of high quality reflecting the true condition of the patients, and many studies have attempted to measure the effectiveness of digital biomarkers for various clinical use cases [1].

RADAR-Base Platform

High-quality longitudinal data collected for long periods and at scale are a key requirement for digital biomarker development. The widespread availability of smartphones, more capacious mobile networks, and the development of new wearable sensors have enabled measurement of a growing set of physiological and phenomenological parameters relevant to physical and mental diseases. To facilitate wearable and smartphone data remote collection at scale and digital biomarker development, the RADAR-base platform was developed and released under the open-source Apache 2 License in January 2018 [2,3]. RADAR-base comprises an Apache Kafka-based back end deployed onto Kubernetes infrastructure and 2 mobile apps. The cross-platform (Android, iOS) Cordova Active Remote Monitoring (aRMT) app for active monitoring of participants requires conscious action (eg, questionnaires, audio questions, timed tests), and the Passive Remote Monitoring (pRMT) app, a native Android app, does passive monitoring (without direct action from participants) via phone, wearable devices, and Internet of Things. A high-level overview of the platform is shown in Figure 1.

Figure 1. Overview of the RADAR-base platform. Current data sources: Empatica E4, Pebble 2, Fitbit, Biovotion, Faros, Garmin, Active Remote Monitoring questionnaire app and Passive Remote Monitoring app. API: application programming interface; IoT: Internet of Things.



The aRMT app renders questionnaires using JavaScript Object Notation (JSON) definition files, which store data in key-value pairs. Within a single questionnaire file exists a collection of questions, each containing attributes such as field name, label or text, input type (checkbox, free text, etc), choices, and more. Subsequently, the aRMT app uses these files to display the questionnaire within the user interface. This facilitates the dynamic deployment of questionnaires for wide-ranging project requirements. Example questionnaires used in the aRMT app in existing projects include measures of self-esteem (Rosenberg Self-Esteem Scale), depression (8-item Patient Health Questionnaire), and ecological momentary assessment. These questionnaires include traditional question sets, while others take a different approach. For instance, the speech questionnaire requires users to record themselves reading a specific text or answering a question, rather than responding to written prompts.

The passive application runs in the background, requiring minimal or no input from participants. Data are collected from smartphone “sensors” and from integrated wearable devices. The catalogue of devices currently integrated into the pRMT app includes onboard Android smartphone sensors, Empatica E4, Pebble 2 smartwatch, Biovotion Everion, Faros 180 and 360, Fitbit, Garmin Vivosmart, and Oura Ring. Pluggable capability is provided to integrate new wearable devices offering a native software development kit (SDK) (eg, Empatica E4) or through third-party vendor’s Representational State Transfer Application Programming Interface (REST API) (eg, Fitbit via the back end REST Collector + REST-Authorizer for OAuth-2 Flows).

A common task is the exploration of collected raw data. RADAR-base includes capability for data aggregation, management of studies, and real-time visualization in Grafana dashboards [4]. In addition to the near real-time visualization provided by the dashboard, the RADAR-base platform includes a Python package designed for data processing, feature generation, and visualization. This package offers a range of standard tools for exploratory visualizations of collected data. It also simplifies the implementation of feature generation pipelines, allowing users to take data exported from a RADAR-base project and generate processed data (high-level features), along with any associated labels, in a format suitable for use with commonly used machine learning libraries.

Objectives of This Study

The platform is used in a wide range of research and clinical studies and is available under a range of service models, depending on the requirement of the project. This paper extends the foundation laid by the initial iteration of the RADAR-base platform [3], describing its architecture and technical components in thorough detail. This work aims to encapsulate and outline a pertinent assortment of research and clinical studies that leverage the RADAR-base platform for data collection and digital phenotyping. The focus of this paper will be on distilling prevalent usage patterns and addressing the challenges encountered in using the platform for diverse research endeavors.

Related Work and Comparison With Other Platforms

Numerous studies are validating digital biomarkers for disorders and their effectiveness, and studies [5] and [6] were conducted to assess the usefulness of digital biomarkers for mood and depression. Both studies had smaller cohorts of only 60 and 59 participants, respectively, which limits the studies’ findings. Digital biomarkers are being studied with a view to replacing or augmenting traditional markers for disorders and a number of barriers have been identified; these barriers include standardization and regulation, and studies are undergoing to address challenges and streamline digital biomarkers in health care [7]. Open source software platform for end-to-end digital biomarker development digital biomarker discovery pipeline was developed to standardize digital biomarker development. Digital biomarker discovery pipeline modules calculate and use resting heart rate, glycemic variability, insulin sensitivity status, exercise response, inflammation, heart rate variability, activity, sleep, and circadian patterns to predict health outcomes [8]. “mCerebrum: A Mobile Sensing Software Platform for Development and Validation of Digital Biomarkers and Interventions” supports high-rate data collections from multiple sensors with real-time assessment of data quality and development of digital biomarkers [9]. Guidelines for developing digital biomarkers are proposed in the study [10].

Intel Context Sensing SDK is a library for Android and Windows with specific context states; it, however, provides only front-end components [11]. The EmotionSense app is developed by the University of Cambridge to sense emotions with implications for psychological therapy and improving well-being; however, it is focused only on depression [12]. Medopad, now known as Huma, provides solutions for different health care issues with symptom tracking; this is a commercial solution and mainly focuses on phone sensors and active-monitoring methods [13]. Personal health INTERVENTION tool kit allows users to build health apps based on existing infrastructure [14]. ResearchKit is an open-source framework for building apps specifically for iOS. ResearchKit makes it easier to enroll participants and conduct studies; however, new wearable device integration requires strong programing skills and it does not include a data management solution [15].

ResearchStack is an SDK and user experience framework for building research study apps on Android, with a similar application domain as ResearchKit [16]. Both ResearchKit and ResearchStack provide software libraries, frameworks, and development tools that require extensive programing skills to create apps. A framework to create observational medical studies for mobile devices without extensive programing skills was presented [17]. LAMP (Learn Assess Manage and Prevent) platform [18] provides an app and back-end infrastructure for clinical relevant studies; app can adapt to different studies with input from patients.

One of the distinctive features of the RADAR-base platform is its use of Confluent platform technologies [19], which are built around Apache Kafka. This choice forms the foundation for a highly scalable end-to-end solution for event-driven messaging, capable of addressing diverse use cases such as high throughput, low-latency messaging, real-time data processing, and fault tolerance. Deployable as microservices with Kubernetes cluster [20], the platform offers seamless integration for new sensors and data sources with minimal effort. Number of new sensors are added to the platform since its inception for clinical studies.

Minimal effort to integrate new sensors and devices offers opportunity to capture signals of various diseases and use cases.

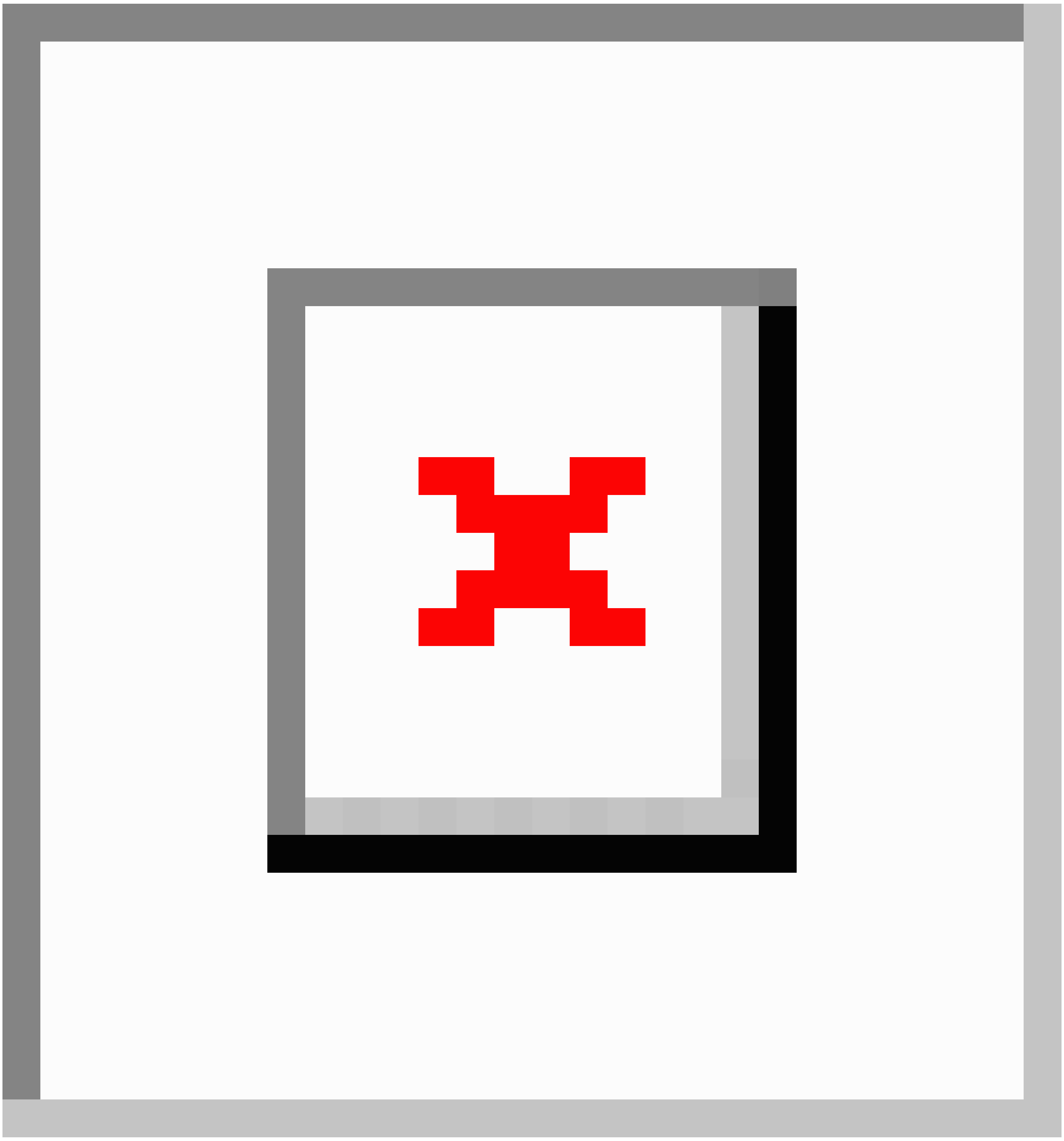
There is a need for a systematic approach to assess the quality and use of digital biomarkers to ensure an appropriate balance between their safety and effectiveness. Coravos et al [21] outline key considerations for the development and evaluation of digital biomarkers, examining their role in clinical research and routine patient care. RADAR-base provides a secure and effective digital biomarker ecosystem, ensuring transparency of the algorithms, interoperable components with open interfaces to accelerate the development of new multicomponent systems, and high-integrity measurement systems.

Methods

Digital Phenotyping of Disease

A key feature of RADAR-base is its extensibility allowing new wearables and sensors to be readily integrated into the platform to collect new modes of data depending on the study requirements. Collected raw data from phone and wearable sensors can then be aggregated and converted into low-level features and subsequently high-level features, representing digital biomarkers. Figure 2 exemplifies the process for major depressive disorder (MDD) for wearable and phone sensor–collected data, for example, phone microphone–collected audio data can provide different speech features and respiratory acoustics that could help identify respiration and physiological stress; similarly, low-level acceleration provided–actigraphy features can be used to identify psychomotor retardation. Developed digital biomarkers may provide insight into the disorder and could be used in clinical trials to ascertain their usefulness in management of the disease.

Figure 2. Collected raw data transition into high-level features which gives insight into major depressive disorder. ePRO: electronic patient-reported outcomes; HR: heart rate; HRV: heart rate variation.



The next two sections discuss the development of high-level features for MDD and epilepsy and how these features are being used to explore and manage disease. Textbox 1 shows selected features extracted from sensor data and Table 1 shows digital biomarkers developed for different disorder areas [22-24]. An open-source feature generation pipeline has been created to enhance and standardize the analysis of data generated by RADAR-base. This pipeline facilitates the extraction of features and biomarkers, enabling cross-disease symptom analysis. With its capabilities to ingest, analyze, visualize, and export RADAR-base data, the pipeline simplifies and establishes a convention for data scientists. This streamlines the process of feature-based analysis, ensuring consistency and enabling researchers to gain valuable insights from the data. Pipelines are readily extended and published on the RADAR-base pipeline catalogue [25].

Textbox 1. Features extracted from the sensors integrated into the RADAR-base platform.

Apps

- Active Remote Monitoring
- Passive Remote Monitoring

Devices

- Empatica E4
- Biovotion Everion
- Bittium Faros 180
- Faros 360
- Fitbit
- Garmin
- Smartphone
- Oura Ring

Sensors/raw data

- Acceleration
- Blood volume pulse
- Electrodermal activity
- Interbeat interval
- Temperature
- Blood pulse wave
- Galvanic skin response
- Heart rate
- Oxygen saturation
- Led current
- Photoplethysmogram raw
- Electrocardiogram
- Gyroscope
- Light
- Magnetic field
- Location
- Microphone
- Step count
- Usage event
- User interaction
- Activity levels
- Activity log record
- Intraday calories
- Intraday steps
- Resting heart rate
- Sleep classic
- Sleep stage
- SMS text message unread
- Bluetooth devices

- Phone battery level
- Phone contact list
- Garmin stress tracking
- Garmin relaxation
- Breathing timer
- Garmin Vo2max
- Garmin body
- Battery energy
- Monitor
- SpO2

Selected features

- Sleep duration
- Sleep architechture
- Sleep stability
- Sleep quality
- Sleep efficiency
- Sleep Fragmentation Index
- Sleep onset latency
- Sleep onset latency variance
- Sleep midpoint
- Sleep midpoint variance
- Insomnia
- Hypersomnia
- Unlock times/duration
- Unlock duration min/max
- Median interval between two unlocks
- Step epoch
- Daily step sum
- Moderate walking duration
- Maximum nonstop duration
- Maximum nonstop step count
- Activity level
- Actigraphy
- Respiratory acoustics
- Heart rate
- Heart rate variation
- Oxygen saturation
- Skin conductance
- Electronic patient-reported outcome
- Ambient light
- Activity
- Phone use
- Bluetooth

- Max/min/mean/SD of nearby Bluetooth devices (NBDC)
- NBDC entropy
- NBDC frequency features
- Location
- Location variance
- Moving time
- Moving distance
- Number of location clusters
- Location entropy
- Homestay
- Location frequency features
- Gait
- Median gait cycles
- Frequency of gait
- Median force
- Change in total sleep
- Social jet lag

Table . Digital biomarkers generated from extracted features and associated studies.

Digital biomarker	Device	Sensor/raw data	RADAR-CNS[a]	RADAR-AD[b]	AIMS-2-TRIALS[c]	ART[d]	BigData@Heart	RALPMH[e]	COVID-Collab
Total sleep	Fitbit/Garmin	Sleep stage	✓	✓	✓	✓	✓	✓	✓
Social interactions	Smartphone	Bluetooth	✓		✓	✓			
Gait patterns	Smartphone	Acceleration	✓	✓		✓			
Respiration	Garmin	Garmin respiration rate						✓	✓
Psychological stress	Garmin	Garmin stress			✓		✓	✓	✓
Phone use	Smartphone	User interaction	✓		✓	✓	✓		
Ambulatory mobility	Smartphone	Location	✓			✓			✓
Fatigue	Garmin	Garmin body battery energy						✓	✓
Seizures	Empatica E4/Biovotion Everion	Acceleration, EDA[f], PPG[g]	✓						
Step	Fitbit/Garmin	Step count	✓	✓		✓			✓
Activity	Fitbit/Garmin	Activity logs	✓	✓	✓	✓		✓	✓
Speech	Smartphone	Microphone	✓		✓			✓	
Resting HR[h]	Fitbit/Garmin	PPG	✓	✓		✓		✓	✓
HR variability	Fitbit/Garmin	PPG	✓	✓		✓		✓	✓
Sleep variance	Fitbit/Garmin	Sleep stage	✓	✓		✓		✓	✓
Mobility variance	Smartphone	Location	✓	✓		✓		✓	✓
Respiratory acoustics	Smartphone	Microphone	✓	✓				✓	

[a]RADAR-CNS: Remote Assessment of Disease and Relapse—Central Nervous System.

[b]RADAR-AD: Remote Assessment of Disease and Relapse–Alzheimer Disease.

[c]AIMS-2-TRIALS: Autism Innovative Medicine Studies–2–Trials.

[d]ART: ADHD Remote Technology.

[e]RALPMH: Remote Assessment of Lung Disease and Impact on Physical and Mental Health.

[f]EDA: electrodermal activity.

[g]PPG: photoplethysmography.

[h]HR: heart rate.

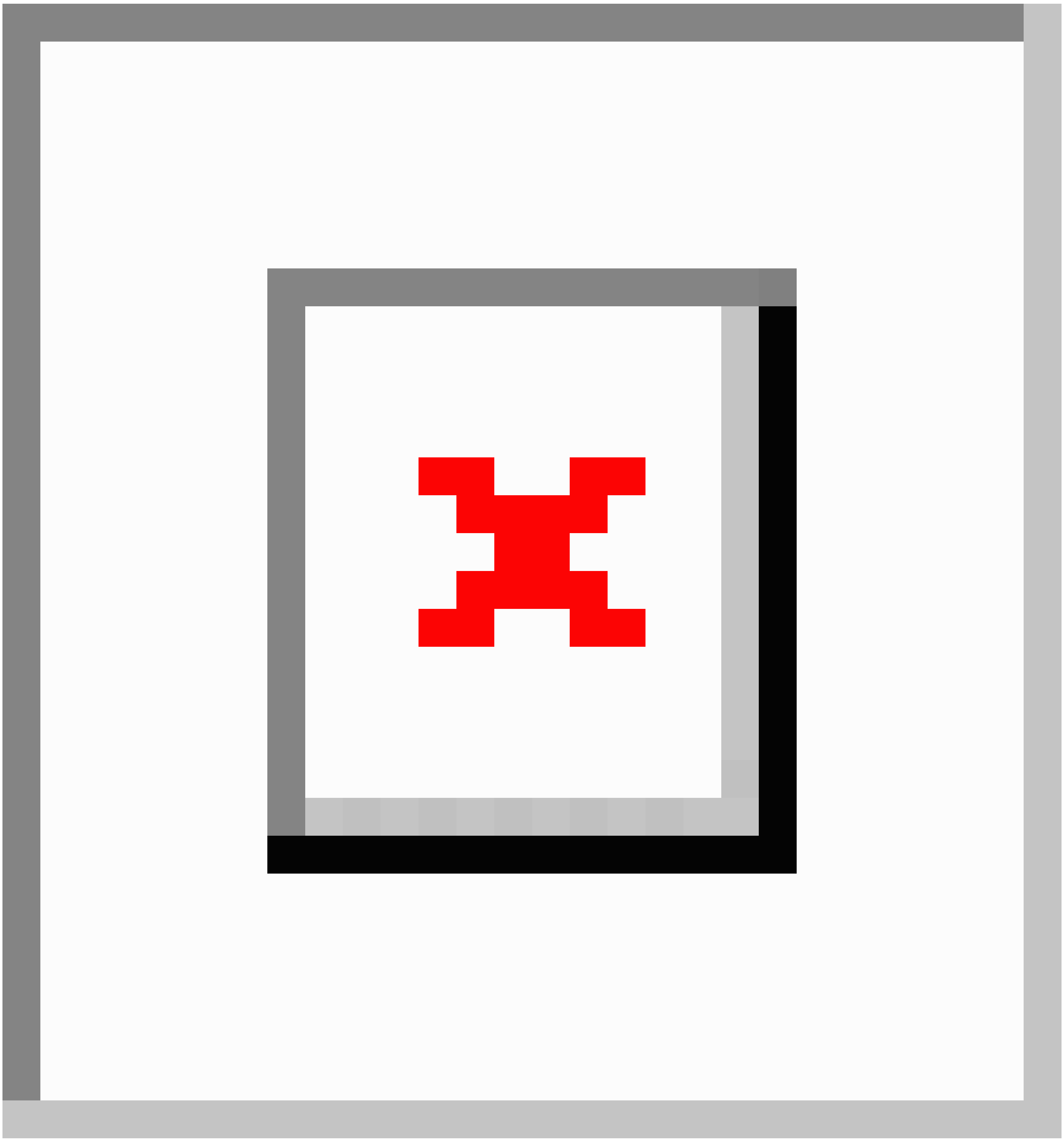
Major Depressive Disorder

MDD is associated with a wide range of negative outcomes including premature mortality, reduced quality of life, and loss of occupational function, and it is often experienced alongside physical comorbidity and approximately 55% will go on to develop chronic depression, characterized by periods of recovery and relapse [26].

The pRMT app provides a comprehensive solution for activity monitoring by using wearable device sensors and smartphones to collect data without requiring any input from the wearer. It leverages a range of sensors, including GPS, accelerometer, gyroscope, communication logs, ambient noise and light levels, and screen interactions. Through this approach, the app can effectively and passively gather diverse data streams, enabling a seamless and unintrusive data collection process for various applications. These sensors along with the Fitbit watch have the potential to identify changes in sleep, communication, and activity patterns associated with depressive episodes.

The smartphone-embedded Bluetooth sensor can be used to record individuals' local proximity information, such as the nearby Bluetooth device count (NBDC) that includes the Bluetooth signal of other phone users. The NBDC data have the potential to reflect changes in people's behaviors and statuses during the depressive state [27]. An illustration is given in Figure 3.

Figure 3. (**A**) A schematic diagram showing an individual's nearby Bluetooth devices count (NBDC) in different scenarios in daily activities and life. (**B**) Pearson correlation heat map for top 20 features of walking data (median rankings from all models). (**C**) Exemplar geolocation data correspond to a biweekly segment of a study participant, The red dots denote an individual's home location cluster, whereas longitude and latitude along the axes are expressed in decimal degrees. (**D**) The PHQ-8 scores and a select 4 sleep features of 1 participant with an obvious increasing trend in PHQ-8 score at 13th PHQ-8 record. PHQ-8: 8-item Patient Health Questionnaire.

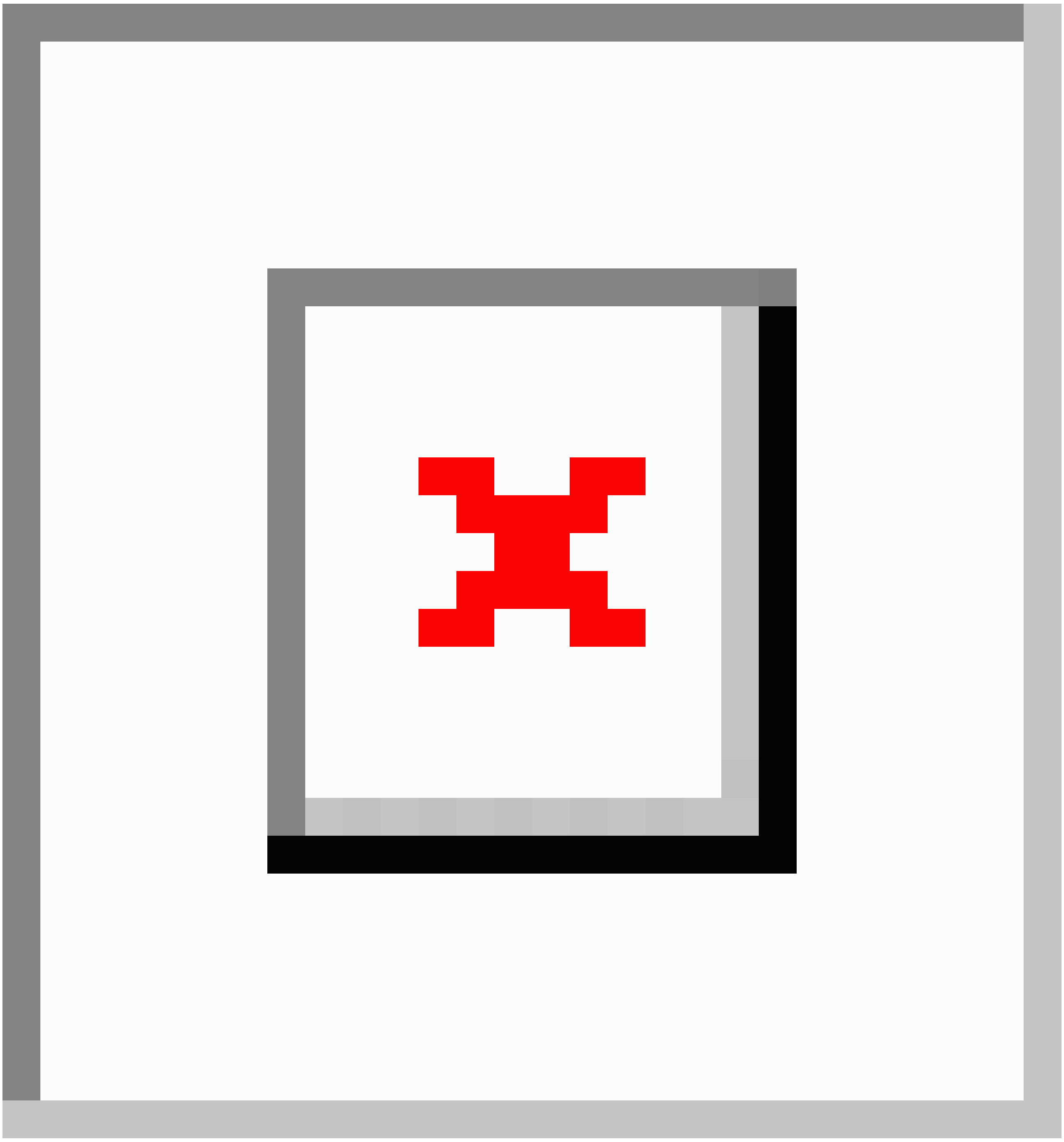


Speech characteristics, such as speaking rate, pitch, pause duration, and energy, collected via smartphone microphones, have been used to detect depression with a prediction accuracy of 81.3% [28]. The aRMT app delivers validated questionnaires, cognitive games, speech tasks, or electronic diaries using the experienced sampling methodology to provide a fine-grained understanding of mood changes and stressors in the context of daily life. The aRMT app has been used to measure effect, cognition, and mood and behavior in real time, with evidence highlighting the increased validity of this methodology in comparison with traditional retrospective reports. The aRMT assessments of positive and negative affect have also been found to be reliably indicative of mood state and have been associated with MDD symptoms [29].

Epilepsy

Numerous epilepsy research studies based on epilepsy monitoring units have shown the possibility of capturing characteristic movement associated with myoclonic seizure manifestations using wearable sensors [30]. Using pRMT app–integrated wearable sensors, it is possible to record several signals associated with seizure including motor components, using inertial sensors such as accelerometry and surface electromyography, various features of heart rate variations captured by wearable electrocardiogram and photoplethysmography, and alteration of the autonomic nervous system with electrocardiogram, photoplethysmography, and electrodermal activity sensors, with different levels of signal and seizure detection accuracy. Figure 4 shows seizures detected with the collected data using the Empatica E4 wearable.

Figure 4. Empatica E4 sensor data. The area between the vertical dashed lines indicates a focal seizure with a motor component. (**A**) accelerometer, (**B**) photoplethysmography (PPG) blood volume pulse, (**C**) PPG interbeat interval, and (**D**) electrodermal activity. EEG: electroencephalogram.



Participant Recruitment Process

Participants are recruited using different methods, including through clinical services, hospitals, and remotely including through a citizen science approach, depending on the study requirements. A number of recruitment strategies are supported by the platform including the following:

1. All participants are recruited at once, and the study starts simultaneously.
2. Participants enter the study in a "batch" mode.
3. Participants are recruited continuously until the desired sample size or date is reached ("stream mode").

Simultaneous recruitment from multiple sites is possible supporting recruitment of diverse population groups for the same study.

Projects Using RADAR-Base Platform

Table 2 presents a summary of selected projects using the platform with disorders they are focused on, along with the cohort size and sensors being used. It also lists the main objectives of the project. A brief summary of the projects listed in Table 2 are provided in the next sections. In some projects, wearables data collected through the platform augment existing collected data, for example, historical clinical records or baseline assessments.

Table . Projects summary with disease area and study size including devices used and main objectives.

Project and disease area		Size n	Enrolled, n	Devices/data types	Main objectives
RADAR-CNS [a]					
	Depression	600	623	Fitbit, phone sensors, questionnaires	Depressive relapse
	Epilepsy	200	145	Biovotion Everion, Empatica E4, questionnaires	Epilepsy seizure and preictal seizure detection
	Multiple sclerosis	500	430	Fitbit, Bittium Faros, phone sensors, questionnaires	Trajectory of disease, characterization, relapsing/remitting of disease symptoms
ART-CARMA[b]					
	Cardiometabolic risk factors	300	200 (ongoing)	Empatica EmbracePlus, phone sensors, questionnaires	Pretreatment initiation through to treatment initiation, titration, and the subsequent period
ART[c]					
	Attention-deficit/hyperactivity disorder	40	40	Fitbit, phone sensors, questionnaires	To establish a remote assessment and monitoring system for adults and adolescents with ADHD[d]
RADAR-AD[e]					
	Alzheimer disease	200	229	Fitbit, phone sensors, questionnaires	Feasibility of remote-monitoring technologies for AD[f]
AIMS-2-TRIALS[g]					
	Autism	500	300 (ongoing)	Empatica E4, Fitbit, phone sensors, questionnaires	Biology of autism to tailor treatments and develop new therapies and medicines
BigData@Heart					
	Atrial fibrillation	160	160	Phone sensors, questionnaires	Comparison of 2 strategies of rate control, based either on initial treatment with digoxin or β-blockers
DynaMORE[h]					
	Mental health	—[i]	—	Phone sensors, questionnaires	Developing an in-silico model of stress resilience
CONVALESCENCE					
	Long-term effects of COVID-19	800	363 (ongoing)	Garmin Vivoactive, phone sensors, questionnaires	Characterization, determinants, mechanisms, and consequences of the long-term effects of COVID-19
COVID-Collab					

Project and disease area		Size n	Enrolled, n	Devices/data types	Main objectives
	COVID-19	15,000	17,667	Fitbit, questionnaires	Behavior and physical and mental health occur in response to COVID-19 infection, persistent symptoms, and the pandemic in general
RALPMH[j]					
	Lung disease	60	60	Garmin Vivoactive, phone sensors, Questionnaires	Feasibility of remote-monitoring technologies for high-burden pulmonary disorders
EDIFY					
	Eating disorder	500	10 (ongoing)	Oura Ring, phone sensors, questionnaires	Delineating illness and recovery trajectories to inform personalized prevention and early intervention in young people
UNFOLD					
	Psychosis	50	—	Questionnaires	To characterize the processes involved in developing an identity as a person in recovery
Jovens na Pandemia & MAAY Study					
	Depression	280	—	Phone sensors, questionnaires	Remotely monitor behavioral and symptom changes associated with behavioral interventions in children and adolescents

[a]RADAR-CNS: Remote Assessment of Disease and Relapse—Central Nervous System.

[b]Attention-Deficit/Hyperactivity Disorder Remote Technology Study of Cardiometabolic Risk Factors and Medication Adherence.

[c]ART: ADHD Remote Technology.

[d]ADHD: attention-deficit/hyperactivity disorder

[e]RADAR-AD: Remote Assessment of Disease and Relapse–Alzheimer Disease.

[f]AD: Alzheimer disease.

[g]AIMS-2-TRIALS: Autism Innovative Medicine Studies–2–Trials.

[h]DynaMORE: Dynamic Modelling of Resilience.

[i]Not available.

[j]RALPMH: Remote Assessment of Lung Disease and Impact on Physical and Mental Health.

Remote Assessment of Disease and Relapse–Central Nervous System

Remote Assessment of Disease and Relapse–Central Nervous System (RADAR-CNS) was a cohort study that developed new ways of monitoring MDD, epilepsy, and multiple sclerosis using wearable devices and smartphone technology. Patients' data were collected continuously for 24 months [31]. More than 1200 participants took part in the study in different disease areas, and participant recruitment was done via clinics and hospitals. Different study protocols with different wearable devices were used for each disease. Participants were recruited from 6 different sites from different countries. Digital biomarkers developed through the remotely collected data give a better understanding of the diseases and will help clinicians manage them timely [27,32,33].

Attention-Deficit/Hyperactivity Disorder Remote Technology Study of Cardiometabolic Risk Factors and Medication Adherence

The Attention-Deficit/Hyperactivity Disorder Remote Technology Study of Cardiometabolic Risk Factors and Medication Adherence (ART-CARMA) aims to obtain real-world data from the patient's daily life to explore the extent to which attention-deficit/hyperactivity disorder (ADHD) medication and physical activity, individually and jointly, may influence cardiometabolic risks in adults with ADHD. The second objective is to obtain valuable real-world data on adherence to pharmacological treatment and its predictors and correlates. The long-term goal is to use collected data to improve the management of cardiometabolic disease in adults with ADHD and to improve ADHD medication treatment adherence and the personalization of treatment [34]. For this cohort, 2 study sites in London and Barcelona are concurrently recruiting the participants using the platform.

ADHD Remote Technology

The ADHD Remote Technology (ART) was a pilot project focused on developing a novel remote assessment system for ADHD. ART assessed the feasibility and validity of remote researcher-led administration and self-administration of modified versions of 2 cognitive tasks sensitive to ADHD, a 4-choice reaction time task (Fast task) and a combined Continuous Performance Test/Go No-Go task (CPT/GNG) [35]. A cohort of 40 participants was recruited, 20 controls and 20 patients with ADHD.

Remote Assessment of Disease and Relapse–Alzheimer's Disease

Remote Assessment of Disease and Relapse–Alzheimer's Disease (RADAR-AD) aimed to transform Alzheimer disease patient care through remote assessment using mobile technologies such as smartphones or fitness trackers [36]. The project developed the technology to identify which clinical or physiological features, digital biomarkers, can be measured remotely to predict deterioration in function. RADAR-AD created a pipeline for developing, testing, and implementing remote measurement technologies with patients involved at each stage. Complete details of the study protocol and pipeline development are explained in the study by Muurling et al [37]. It was an augmentation study in which 300 participants took part. Three different categories of participants were recruited: controls, mild cognitively impaired or prodromal Alzheimer disease, and Alzheimer dementia.

Autism Innovative Medicine Studies–2–Trials

The Autism Innovative Medicine Studies–2–Trials (AIMS-2-TRIALS) program includes a range of studies to explore how autism develops, from before birth to adulthood, and how this varies in different people. AIMS-2-TRIALS is looking for biological markers which indicate whether a person has or may develop particular characteristics [38]. AIMS-2-TRIALS collects both active and passive data in clinical assessment settings and in home-based and ambulatory settings. Fitbit is used for remote data collection, and Empatica E4 is used for local data collection at hospitals. Digital markers could help identify those who may ultimately benefit from particular treatments. Medicines will also be tested to help with social difficulties, repetitive behaviors, and sensory processing. Remote-monitoring data are augmenting the clinical data.

Rate Control Therapy Evaluation in Permanent Atrial Fibrillation (BigData@Heart)

The Rate Control Therapy Evaluation in Permanent Atrial Fibrillation (RATE-AF) study was designed to compare 2 strategies of rate control, based on either initial treatment with digoxin or β-blockers in 160 patients with atrial fibrillation (AF) in need for rate control therapy. Monitoring with wearable devices, phone sensors, and questionnaires was conducted over a continuous 6-month period. Objectives of the project included discovering new phenotypes, developing reliable subphenotyping, and informing new taxonomies of heart failure based on a better understanding of underlying disease processes. Sleep, heath rate, heart rate variation, and activity data were collected to develop new phenotypes. This work is additionally supported by the EU IMI2 BigData@Heart major program [39].

Dynamic Modelling of Resilience

Dynamic Modelling of Resilience (DynaMORE) generated the first personalized in-silico model of mental health in the face of adversity or stress resilience. The model is based on and validated against unique multiscale longitudinal real-world empirical data sets, collected through neuroimaging, experimental assessments, questionnaires, and remote monitoring using the pRMT app and a wearable device. The model will substantially deepen scientific understanding of the mechanisms of resilience, supporting the creation of mechanistically targeted interventions for the primary prevention of stress-related disorders. On this basis, DynaMORE developed an entirely new mobile health (mHealth) product incorporating the RADAR-base platform that will include model-based prognostic tools for real-time and real-life monitoring of at-risk subjects and for automated decision-making about timed, personalized interventions.

CONVALESCENCE

CONVALESCENCE is focused on the characterization, determinants, mechanisms, and consequences of the long-term effects of COVID-19, providing the evidence base for health care services [40]. It is an existing large longitudinal cohort being further characterized and augmented by incorporating wearables data. Deep phenotyping and remote assessment using mobile devices and smartphones through the RADAR-base platform is being used to identify subclinical damage or dysfunction in individuals with long-term COVID-19.

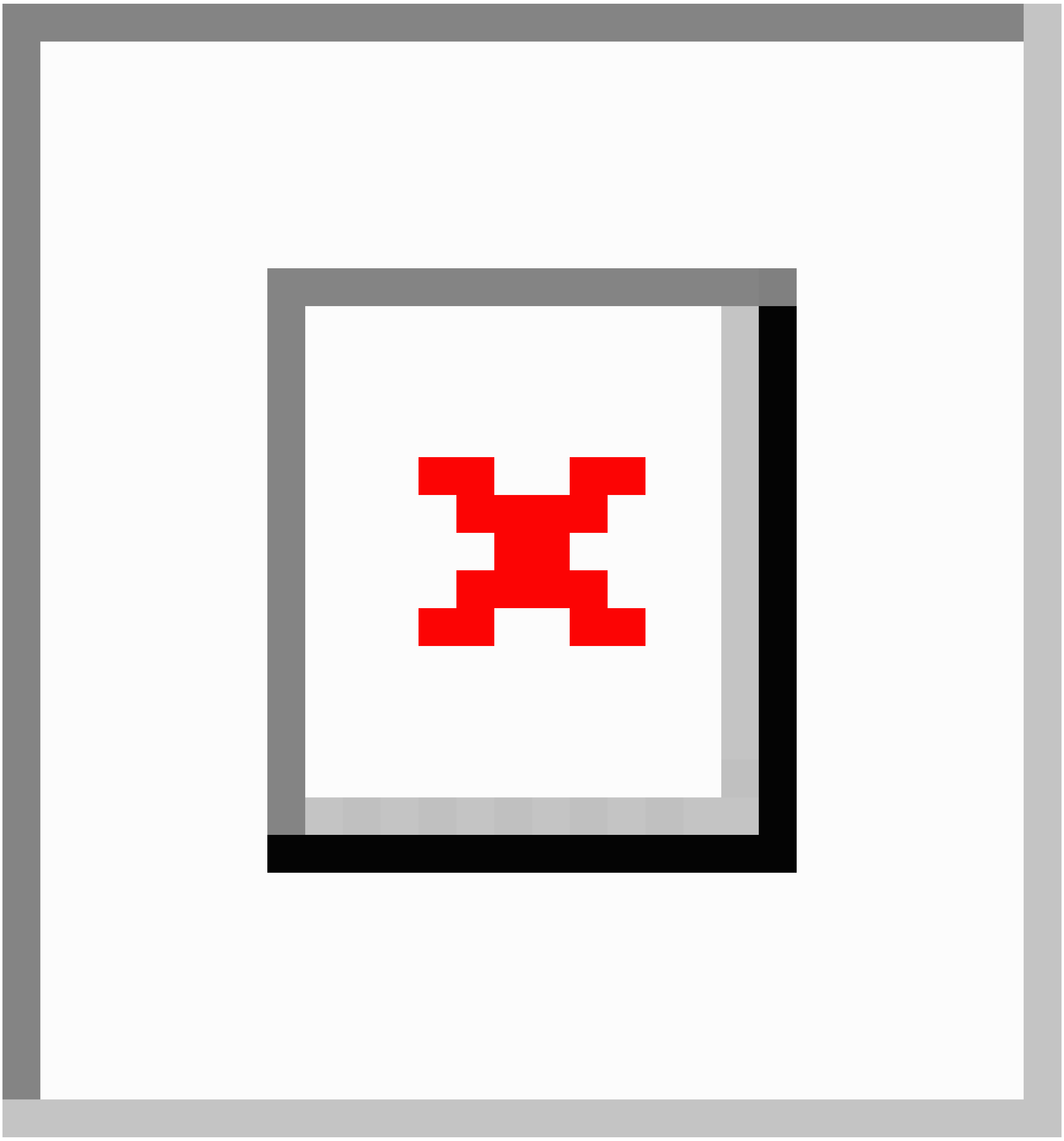
COVID-Collab

COVID-Collab is a citizen science project with members of the public volunteering to donate their wearable data and complete diagnosis and symptom questionnaires. The main aim was to investigate the ongoing COVID-19 outbreak: (1) establish whether wearable data can be used to diagnose COVID-19 infection and (2) characterize the disease symptoms and evolution. A key feature of the study is the use of wearable data to investigate changes in mental health and physiological measurements such as heart rate during infection with coronavirus [41].

Remote Assessment of Lung Disease and Impact on Physical and Mental Health

Chronic lung disorders such as chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis are characterized by exacerbations and decline over time. Twenty participants were recruited in each of 3 cohorts (chronic obstructive pulmonary disease, idiopathic pulmonary fibrosis, and posthospitalization COVID). Data collection is being done remotely using the RADAR-base platform for different devices, including Garmin wearable devices and smart spirometers, mobile app questionnaires, surveys, and finger pulse oximeters. The Remote Assessment of Lung Disease and Impact on Physical and Mental Health (RALPMH) project focuses on the feasibility of remote monitoring in chronic lung disorders and provides a reference infrastructure for future studies [42]. Figure 5 shows an overview of the RALPMH study.

Figure 5. Various data sources will be used to collect both active and passive data to gain a unique perspective into patient health in the Remote Assessment of Lung Disease and Impact on Physical and Mental Health lung disorder study. aRMT: Active Remote Monitoring; FEV-1: forced expiratory volume in the first second of expiration; FVC: forced vital capacity; HR: heart rate; PPG: photoplethysmography.



Eating Disorder (EDIFY)

The objective of the EDIFY study is to undertake a longitudinal comparison of the biopsychosocial symptom profiles of those with early- and late-stage eating disorders (EDs) and recovery trajectories of those with early-stage EDs. This will provide evidence that will help inform decision-making of targeted intervention and preventative treatments across EDs for those with early and more progressed forms of illness. ED patients do not like to wear any wearable device or have any information displayed on their phone regarding their calories and daily workout. The Oura ring is being experimented for the first time with ED patients at scale as it provides no feedback to the patients during the study.

Ethical Considerations

Prior to initiating each project, we obtained approval for each study from the relevant ethics committee and conducted a data protection impact assessment (DPIA), which is equivalent to an institutional review board in some countries. Each study adhered to its unique data collection protocol, and both the General Data Protection Regulation or DPIA and the protocol were approved by the research ethics committee and the data protection officer, respectively. In the study by Ranjan et al [3], comprehensive information regarding data protection, privacy, and the process of pseudonymization is expounded.

Using RADAR-Base Platform and Hosting Models

RADAR-base is of interest to a wide range of mHealth communities, from academic research to industry and wearable vendors interested in collecting data remotely or integrating new data sources into the platform. The platform is freely available as an open-source (Apache 2 License) GitHub repository [43]. More details of the platform can be found on the official RADAR-base website [44]. A detailed quickstart, deployment details, and developer documentation are made available on the platform Confluence Wiki [45]. Docker images for all the components are available at Docker Hub [46] and a Kubernetes stack is also available for the deployment [47]. The aRMT app questionnaires and protocol implementation is explained in “RADAR-base/RADAR-aRMT-protocols” [48]. An exemplar DPIA that explains the data collection procedure is shared in Multimedia Appendix 1. In principle, 3 hosting models are available for using the platform.

Self-Hosting

The platform, along with its accompanying apps and related elements, is open source, allowing users to host it on private or local servers and tailor it to suit their study requirements. Under the self-hosting model, users possess full autonomy over deployment, infrastructure, and data collection.

Supported Hosting

Under the supported hosting model, a third-party provider of RADAR-base can deploy the platform on local or public cloud such as Amazon Web Services. The costing and support arrangements will be determined by factors such as the study’s duration, participant count, and data throughput, primarily based on sensor types selected. In this setup, users will share control over the infrastructure and data with the RADAR-base team. Integration with electronic case report forms such as REDCap (Research Electronic Data Capture) also allows more sensitive data to be segregated from the managed service and be retained entirely under the user’s control.

Fully Managed Hosting

This option is available to completely outsource the deployment, and hosting to a third party provides the platform deployed on their infrastructure and hosts projects as a service to, for example, researchers.

Results

Overview

A large number of mental and physical health research studies have used the RADAR-base platform for remote data collection with funding from many major funding agencies. This includes more than 50 use cases exploring more than 30 disorder areas with more than 150,000 participants enrolled to date. Major disease areas that are using the platform are MDD, eating disorder, multiple sclerosis, ADHD, autism, epilepsy, atrial fibrillation, Alzheimer Disease, and COVID-19. Projects using the platform are collecting various health parameters depending on the disease area requirement. Data collected relate to cognition, mood, voice, digital usage, geolocation, and heart rate, to name a few. Figures 6 and 7 show examples of the status of collected data, their compliance, and quality for different studies. Numerous challenges addressed by the platform include completeness of data, quality and accuracy of data, participant engagement, and remote data collection.

Figure 6. Active Remote Monitoring app questionnaire completion rate from a single patient from the Remote Assessment of Disease and Relapse—Central Nervous System major depression study. ESM: experienced sampling methodology; PHQ8: 8-item Patient Health Questionnaire; RSES: Rosenberg Self-Esteem Scale.

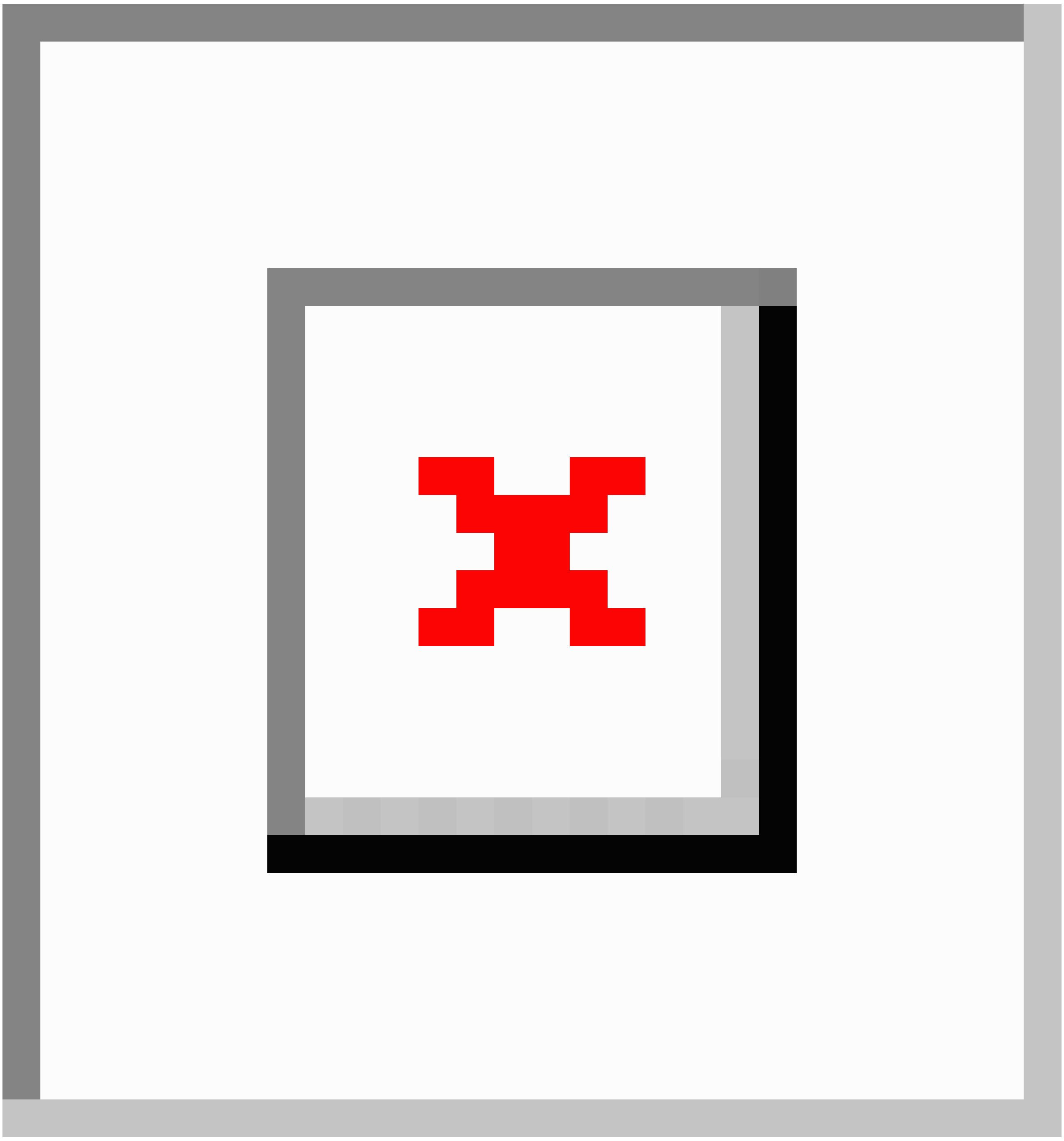
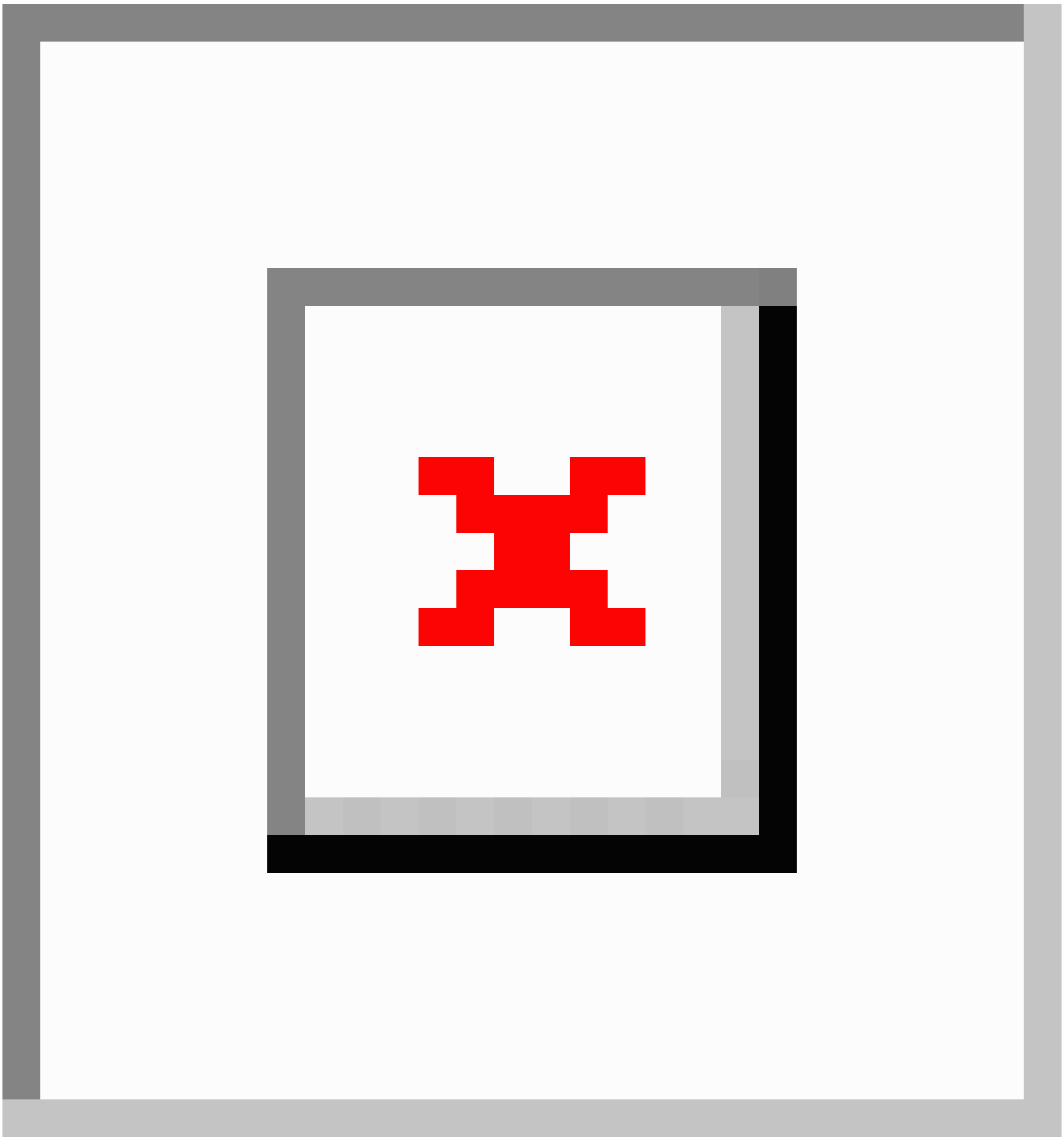


Figure 7. Contiguity of phone sensor data over the first year of enrollment collected through RADAR-base for a patient in the Remote Assessment of Disease and Relapse—Central Nervous System major depressive disorder study. The intensity of the color represents how many hours in a day a particular modality is present.



Digital Biomarker Development

The RADAR-base platform has effectively transformed low-level sensor data into digital biomarkers through feature generation pipelines. This process involves extracting relevant characteristics and patterns from the raw data, enabling the creation of meaningful and actionable insights. These digital biomarkers hold immense potential in various disease areas, aiding clinicians in making informed decisions, facilitating early intervention, and contributing to the prevention of relapse.

Participant Retention, Engagement, and Data Quality

Studies using the RADAR-base platform for data collection and feature extraction have reported insights into data quality, participant engagement, and retention [49,50]. In the Major Depressive Disorder Study, 623 participants were enrolled, with 79.8% (n=497) of them participating for the maximum study duration (11-24 months). In addition, further analysis revealed that wearable data stream had the highest data available rate (n=407 participants with more than 90% data completeness and n=99 participants with more than 50% completeness) across all data streams and found several indicators impacting participant engagement and retention, such as age and mental health status

[50]. These findings illustrate the feasibility of remote data collection for clinical applications and provide the insights and experiences for future mobile health studies.

Discussion

Results suggest that the RADAR-base platform can effectively gather critical health data outside of traditional clinical settings and generate digital biomarkers, improving continuous monitoring and management of physical and mental conditions. It provides a platform that focuses on safety and effectiveness by ensuring transparency in the algorithms used to generate biomarkers. In addition, the interoperable components with open interfaces in RADAR-base facilitate the development of new multicomponent systems. This interoperability ensures that various digital biomarker sources can be integrated into a comprehensive and cohesive platform for health care purposes. Moreover, RADAR-base emphasizes high-integrity measurement systems, which means that the data collected and the biomarkers generated are reliable and accurate. This emphasis on data quality is essential for building trust in the digital biomarker ecosystem and encouraging its adoption in clinical research and routine patient care.

Overall, the systematic approach and emphasis on safety, effectiveness, transparency, and interoperability offered by RADAR-base can contribute to the advancement of digital biomarkers and their integration into health care systems, ultimately benefiting patient outcomes and medical research.

Acknowledgments

This study has received support from the European Union (EU)/European Federation of Pharmaceutical Industries and Associations (EFPIA) Innovative Medicines Initiative (IMI) Joint Undertaking 2 (Remote Assessment of Disease and Relapse—Central Nervous System [RADAR-CNS] grant 115902). This communication reflects the views of the RADAR-CNS consortium, and neither IMI nor the EU and EFPIA are liable for any use that may be made of the information contained herein. The authors receive funding support from the National Institute for Health and Care Research (NIHR) Biomedical Research Centre at South London and Maudsley National Health Service (NHS) Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health. The authors also acknowledge the support of NIHR University College London Hospitals Biomedical Research Centre. This work is also supported by the Medical Research Council/Arts and Humanities Research Council/Economic and Social Research Council Adolescence, Mental Health and the Developing Mind initiative as part of the EDIFY program (grant MR/W002418/1).

Data Availability

The collected data that are used to develop biomarkers in the mentioned studies are typically available upon reasonable request to the principle investigators of each study; requests may be directed to authors ZR, AAF, and RJBD. The data are not publicly accessible due to the ethical constraints of the studies. Authors will formally request approval from the project's principal investigator before sharing the data.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Exemplar RADAR-base platform data protection impact assessment.
[DOCX File, 32 KB - mental_v11i1e51259_app1.docx]

References

1. Motahari-Nezhad H, Fgaier M, Mahdi Abid M, Péntek M, Gulácsi L, Zrubka Z. Digital biomarker-based studies: scoping review of systematic reviews. JMIR Mhealth Uhealth 2022 Oct 24;10(10):e35722. [doi: 10.2196/35722] [Medline: 36279171]
2. Ranjan Y, Kerz M, Rashid Z, et al. RADAR-base: a novel open source m-health platform. Presented at: 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers; Oct 8-12, 2018; Singapore p. 223-226. [doi: 10.1145/3267305.3267579]
3. Ranjan Y, Rashid Z, Stewart C, et al. RADAR-base: open source mobile health platform for collecting, monitoring, and analyzing data using sensors, wearables, and mobile devices. JMIR Mhealth Uhealth 2019 Aug 1;7(8):e11734. [doi: 10.2196/11734] [Medline: 31373275]
4. Grafana. URL: https://grafana.com [accessed 2023-07-26]
5. Opoku Asare K, Moshe I, Terhorst Y, et al. Mood ratings and digital biomarkers from smartphone and wearable data differentiates and predicts depression status: a longitudinal data analysis. Perv Mob Comput 2022 Jul;83:101621. [doi: 10.1016/j.pmcj.2022.101621]
6. Jacobson NC, Summers B, Wilhelm S. Digital biomarkers of social anxiety severity: digital phenotyping using passive smartphone sensors. J Med Internet Res 2020 May 29;22(5):e16875. [doi: 10.2196/16875] [Medline: 32348284]

7. Babrak LM, Menetski J, Rebhan M, et al. Traditional and digital biomarkers: two worlds apart? Digit Biomark 2019 Aug;3(2):92-102. [doi: 10.1159/000502000] [Medline: 32095769]
8. Bent B, Wang K, Grzesiak E, et al. The digital biomarker discovery pipeline: an open-source software platform for the development of digital biomarkers using mHealth and wearables data. J Clin Transl Sci 2020 Jul 14;5(1):e19. [doi: 10.1017/cts.2020.511] [Medline: 33948242]
9. Hossain SM, Hnat T, Saleheen N, et al. mCerebrum: a mobile sensing software platform for development and validation of digital biomarkers and interventions. Proc Int Conf Embed Netw Sens Syst 2017 Nov:7. [doi: 10.1145/3131672.3131694] [Medline: 30288504]
10. Torous J, Rodriguez J, Powell A. The new digital divide for digital biomarkers. Digit Biomark 2017 Sep;1(1):87-91. [doi: 10.1159/000477382] [Medline: 29104959]
11. Intel® Context Sensing SDK. Intel. URL: https://software.intel.com/en-us/context-sensing-sdk [accessed 2018-11-15]
12. Mood-tracking app paves way for pocket therapy. University of Cambridge. 2013 May 8. URL: https://www.cam.ac.uk/research/news/mood-tracking-app-paves-way-for-pocket-therapy [accessed 2022-11-15]
13. Huma. URL: https://www.huma.com/ [accessed 2024-07-17]
14. Eckhoff RP, Kizakevich PN, Bakalov V, Zhang Y, Bryant SP, Hobbs MA. A platform to build mobile health apps: the personal health intervention toolkit (PHIT). JMIR Mhealth Uhealth 2015 Jun 1;3(2):e46. [doi: 10.2196/mhealth.4202] [Medline: 26033047]
15. Jardine J, Fisher J, Carrick B. Apple's Researchkit: smart data collection for the smartphone era? J R Soc Med 2015 Aug;108(8):294-296. [doi: 10.1177/0141076815600673] [Medline: 26268915]
16. ResearchStack. URL: http://researchstack.org [accessed 2018-08-28]
17. Zens M, Grotejohann B, Tassoni A, Duttenhoefer F, Südkamp NP, Niemeyer P. Development of a modular research platform to create medical observational studies for mobile devices. JMIR Res Protoc 2017 May 23;6(5):e99. [doi: 10.2196/resprot.7705] [Medline: 28536095]
18. Torous J, Wisniewski H, Bird B, et al. Creating a digital health smartphone app and digital phenotyping platform for mental health and diverse healthcare needs: an interdisciplinary and collaborative approach. J Technol Behav Sci 2019 Jun;4(2):73-85. [doi: 10.1007/s41347-019-00095-w]
19. Confluent. URL: https://www.confluent.io [accessed 2018-06-19]
20. Kubernetes. URL: https://kubernetes.io [accessed 2024-04-10]
21. Coravos A, Khozin S, Mandl KD. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. NPJ Digit Med 2019;2(1). [doi: 10.1038/s41746-019-0090-4]
22. Sun S, Folarin AA, Zhang Y, et al. The utility of wearable devices in assessing ambulatory impairments of people with multiple sclerosis in free-living conditions. Comput Methods Programs Biomed 2022 Dec;227:107204. [doi: 10.1016/j.cmpb.2022.107204] [Medline: 36371974]
23. Sun S, Folarin AA, Zhang Y, et al. Challenges in the analysis of mobile health data from smartphones and wearable devices to predict depression symptom severity. arXiv. Preprint posted online on Dec 20, 2022. [doi: 10.48550/arxiv.2212.10540]
24. Zhang Y, Folarin AA, Sun S, et al. Longitudinal relationships between depressive symptom severity and phone-measured mobility: dynamic structural equation modeling study. JMIR Ment Health 2022 Mar 11;9(3):e34898. [doi: 10.2196/34898] [Medline: 35275087]
25. RADAR-base/radarpipeline wiki. GitHub. URL: https://github.com/RADAR-base/radarpipeline/wiki [accessed 2023-03-08]
26. Matcham F, Barattieri di San Pietro C, Bulgari V, et al. Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): a multi-centre prospective cohort study protocol. BMC Psychiatry 2019 Feb 18;19(1):72. [doi: 10.1186/s12888-019-2049-z] [Medline: 30777041]
27. Zhang Y, Folarin AA, Sun S, et al. Predicting depressive symptom severity through individuals' nearby Bluetooth device count data collected by mobile phones: preliminary longitudinal study. JMIR Mhealth Uhealth 2021 Jul 30;9(7):e29840. [doi: 10.2196/29840] [Medline: 34328441]
28. Sanchez MH, Vergyri D, Ferrer L, et al. Using prosodic and spectral features in detecting depression in elderly males. Presented at: Interspeech 2011; Aug 27-31, 2011; Forence, Italy. [doi: 10.21437/Interspeech.2011-751]
29. Cummins N, Dineley J, Conde P, et al. Multilingual markers of depression in remotely collected speech samples. Res Sq. Preprint posted online on Nov 9, 2022. [doi: 10.21203/rs.3.rs-2183980/v1]
30. Bruno E, Biondi A, Böttcher S, et al. Remote assessment of disease and relapse in epilepsy: protocol for a multicenter prospective cohort study. JMIR Res Protoc 2020 Dec 16;9(12):e21840. [doi: 10.2196/21840] [Medline: 33325373]
31. RADAR-CNS. URL: http://radar-cns.org [accessed 2023-05-25]
32. Sun S, Folarin AA, Ranjan Y, et al. Using smartphones and wearable devices to monitor behavioral changes during COVID-19. J Med Internet Res 2020 Sep 25;22(9):e19992. [doi: 10.2196/19992] [Medline: 32877352]
33. Zhang Y, Folarin AA, Sun S, et al. Relationship between major depression symptom severity and sleep collected using a wristband wearable device: multicenter longitudinal observational study. JMIR Mhealth Uhealth 2021 Apr 12;9(4):e24604. [doi: 10.2196/24604] [Medline: 33843591]
34. ART-CARMA - the ADHD remote technology study of cardiometabolic risk factors and medication adherence. King's College London. URL: https://www.kcl.ac.uk/research/

art-carma-the-adhd-remote-technology-study-of-cardiometabolic-risk-factors-and-medication-adherence [accessed 2022-09-21]
35. ADHD Remote Technology (ART). King's College London. URL: https://www.kcl.ac.uk/research/adhd-remote-technology-art [accessed 2022-09-21]
36. RADAR-AD. URL: https://www.radar-ad.org [accessed 2022-09-21]
37. Muurling M, de Boer C, Kozak R, et al. Remote monitoring technologies in Alzheimer's disease: design of the RADAR-AD study. Alzheimers Res Ther 2021 Apr 23;13(1):89. [doi: 10.1186/s13195-021-00825-4] [Medline: 33892789]
38. AIMS-2-TRIALS. URL: https://www.aims-2-trials.eu [accessed 2022-09-21]
39. BigData@Heart. URL: https://www.bigdata-heart.eu [accessed 2022-09-21]
40. Convalescence Long COVID Study. COVID-19 Longitudinal Health and Wellbeing National Core Study. URL: https://www.ucl.ac.uk/covid-19-longitudinal-health-wellbeing/convalescence-long-covid-study [accessed 2022-09-21]
41. Covid Collab. URL: https://covid-collab.org [accessed 2022-09-21]
42. Ranjan Y, Althobiani M, Jacob J, et al. Remote assessment of lung disease and impact on physical and mental health (RALPMH): protocol for a prospective observational study. JMIR Res Protoc 2021 Oct 7;10(10):e28873. [doi: 10.2196/28873] [Medline: 34319235]
43. RADAR-base. GitHub. URL: https://github.com/RADAR-base [accessed 2018-06-04]
44. RADAR-base. URL: https://radar-base.org [accessed 2018-06-04]
45. Catalogue of devices and data sources. Confluence. URL: https://radar-base.atlassian.net/wiki/spaces/RAD/pages/87359489/Data+Source+Catalog [accessed 2018-06-04]
46. RADAR-Docker 2.2.0. GitHub. URL: https://github.com/RADAR-base/RADAR-Docker [accessed 2018-06-04]
47. RADAR-Kubernetes. GitHub. URL: https://github.com/RADAR-base/RADAR-Kubernetes [accessed 2022-09-26]
48. Protocols. GitHub. URL: https://github.com/RADAR-base/RADAR-aRMT-protocols [accessed 2023-03-08]
49. Muurling M, de Boer C, Hinds C, et al. Feasibility and usability of remote monitoring in Alzheimer's disease. Dig Health 2024 Apr;10:20552076241238133. [doi: 10.1177/20552076241238133] [Medline: 38601188]
50. Matcham F, Leightley D, Siddi S, et al. Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): recruitment, retention, and data availability in a longitudinal remote measurement study. BMC Psychiatry 2022 Feb 21;22(1):136. [doi: 10.1186/s12888-022-03753-1] [Medline: 35189842]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
AF: atrial fibrillation
aRMT: Active Remote Monitoring
DPIA: data protection impact assessment
ED: eating disorder
JSON: JavaScript Object Notation
LAMP: Learn Assess Manage and Prevent
MDD: major depressive disorder
mHealth: mobile health
NBDC: nearby Bluetooth device count
pRMT: Passive Remote Monitoring
REDCap: Research Electronic Data Capture
REST: Representational State Transfer
SDK: software development kit

Edited by J Torous; submitted 27.07.23; peer-reviewed by A Tacchino, A Baglione, V Gupta; revised version received 07.05.24; accepted 08.05.24; published 23.10.24.

Please cite as:
Rashid Z, Folarin AA, Zhang Y, Ranjan Y, Conde P, Sankesara H, Sun S, Stewart C, Laiou P, Dobson RJB
Digital Phenotyping of Mental and Physical Conditions: Remote Monitoring of Patients Through RADAR-Base Platform
JMIR Ment Health 2024;11:e51259
URL: https://mental.jmir.org/2024/1/e51259
doi:10.2196/51259

Combining mHealth Technology and Pharmacotherapy to Improve Mental Health Outcomes and Reduce Human Rights Abuses in West Africa: Intervention Field Trial

Dror Ben-Zeev[1], PhD; Anna Larsen[1,2], MPH, PhD; Dzifa A Attah[3], MPhil, PhD; Kwadwo Obeng[4], MBChB; Alexa Beaulieu[1], MPH; Seth M Asafo[3], MPhil, PhD; Jonathan Kuma Gavi[3], PhD; Arya Kadakia[1], BA; Emmanuel Quame Sottie[3], BA; Sammy Ohene[3], MBChB; Lola Kola[5,6], PhD; Kevin Hallgren[1], PhD; Jaime Snyder[1,7], PhD; Pamela Y Collins[8], MPH, MD; Angela Ofori-Atta[3], MA, PhD; M-Healer Research Team[3]

1

2

3

4

5

6

7

8

Corresponding Author:
Dror Ben-Zeev, PhD

Abstract

Background: In West Africa, healers greatly outnumber trained mental health professionals. People with serious mental illness (SMI) are often seen by healers in “prayer camps” where they may also experience human rights abuses. We developed “M&M,” an 8-week-long dual-pronged intervention involving (1) a smartphone-delivered toolkit designed to expose healers to brief psychosocial interventions and encourage them to preserve human rights (M-Healer app), and (2) a visiting nurse who provides medications to their patients (Mobile Nurse).

Objective: We examined the feasibility, acceptability, safety, and preliminary effectiveness of the M&M intervention in real-world prayer camp settings.

Methods: We conducted a single-arm field trial of M&M with people with SMI and healers at a prayer camp in Ghana. Healers were provided smartphones with M-Healer installed and were trained by practice facilitators to use the digital toolkit. In parallel, a study nurse visited their prayer camp to administer medications to their patients. Clinical assessors administered study measures to participants with SMI at pretreatment (baseline), midtreatment (4 weeks) and post treatment (8 weeks).

Results: Seventeen participants were enrolled and most (n=15, 88.3%) were retained. Participants had an average age of 44.3 (SD 13.9) years and 59% (n=10) of them were male. Fourteen (82%) participants had a diagnosis of schizophrenia and 2 (18%) were diagnosed with bipolar disorder. Four healers were trained to use M-Healer. On average, they self-initiated app use 31.9 (SD 28.9) times per week. Healers watched an average of 19.1 (SD 21.2) videos, responded to 1.5 (SD 2.4) prompts, and used the app for 5.3 (SD 2.7) days weekly. Pre-post analyses revealed a significant and clinically meaningful reduction in psychiatric symptom severity (Brief Psychiatric Rating Scale score range 52.3 to 30.9; Brief Symptom Inventory score range 76.4 to 27.9), psychological distress (Talbieh Brief Distress Inventory score range 37.7 to 16.9), shame (Other as Shamer Scale score range 41.9 to 28.5), and stigma (Brief Internalized Stigma of Mental Illness Scale score range 11.8 to 10.3). We recorded a significant reduction in days chained (1.6 to 0.5) and a promising trend for reduction in the days of forced fasting (2.6 to 0.0, *P*=.06). We did not identify significant pre-post changes in patient-reported working alliance with healers (Working Alliance Inventory), depressive symptom severity (Patient Health Questionnaire-9), quality of life (Lehman Quality of Life Interview for the Mentally Ill), beliefs about medication (Beliefs about Medications Questionnaire–General Harm subscale), or other human rights abuses. No major side effects, health and safety violations, or serious adverse events occurred over the course of the trial.

Conclusions: The M&M intervention proved to be feasible, acceptable, safe, and clinically promising. Preliminary findings suggest that the M-Healer toolkit may have shifted healers’ behaviors at the prayer camp so that they commit fewer human rights abuses.

(JMIR Ment Health 2024;11:e53096) doi:10.2196/53096

KEYWORDS

mHealth; mobile health; app; apps; applications; human rights; Africa; healers; healer; alternative; complementary; CAM; schizophrenia; bipolar disorder; depression; bipolar; depressive; mental; schizophrenic; schizophrenics; psychosocial; training; abuse; behavior change; behaviour change; medication; medications; pharmacology; pharmacological; pharmacotherapy; feasibility, acceptability, safety

Introduction

Mental health systems in West Africa are constrained in their resources, infrastructure, and access to trained mental health personnel, limiting their capacity to deliver care [1,2]. In part due to these limitations, and in part due to local belief systems that frame mental health problems as spiritual rather than medical or psychological in nature [3,4], people with serious mental illness (SMI) in the region are often treated by traditional or faith healers, who are in abundance [5,6].

Healers often provide services at spiritual centers or "prayer camps"—rustic facilities where family members bring their relatives with mental illness, developmental disabilities, or substance use problems to be "healed" [7]. Pragmatically, prayer camps serve as West Africa's de facto inpatient units and psychiatric residential homes. Once there, individuals seeking care may be retained at prayer camps for weeks, months, or even years at a time. Prayer camps are unregulated by any authorities, for the most part. Healers rarely have training in the etiology, assessment, or treatment of psychopathology, and their clients will seldom receive psychotropic medications or evidence-based psychosocial interventions. Healers provide spiritual consultation, prescribe prayer, and administer various ceremonial or herbal remedies [8,9]. In the absence of both training and resources to manage mental illness, healers may also engage in practices that have dangerous effects and constitute human rights abuses. These may include forced fasting, flogging, confinement in overcrowded or unsanitary conditions, and chaining patients to trees or concrete slabs so that they do not escape the camp's grounds [9,10]. Chaining or similar forced mechanical restraining of people with SMIs, particularly those exhibiting signs of psychosis, are common practices in many low- and middle-income countries and are psychologically and physically damaging [11-13]. Despite the well-documented harsh practices used by some healers, and the negative attention their practices receive from Western media and human rights groups, these paraprofessionals continue to receive referrals and fill a societal need.

Our multinational team works closely with traditional and faith healers and people with SMI in West Africa. Through consultation, bilateral information sharing, articulation of mutual respect, and transparency in our activities, we have fostered partnerships with healers and their prayer camp communities [14-16]. These ties have served as the building blocks for cross-sector collaborations that are designed to improve the safety and quality of care that people with mental illness receive at prayer camps.

Findings from these collaborations have led us to develop M&M, a dual-pronged intervention involving (1) a smartphone-delivered mobile health treatment support toolkit (M-Healer app) designed to expose healers to brief psychosocial interventions, encourage them to preserve human rights in their practice, and prompt them to monitor the status of the people they serve, and (2) a visiting community nurse who provides pharmacological care directly to their patients at the prayer camp (Mobile Nurse). The combination of healer-facing support technology, which can be installed on mobile devices that are widely accessible to healers throughout the region, and patient-facing pharmacotherapy delivered by medically trained personnel, which are more widely available than specialty mental health providers, may be a scalable strategy for addressing unmet mental health and quality-of-care needs in the region. Here we report on the first field trial of the integrated M&M intervention in a prayer camp in Ghana.

Methods

Study Design

The study involved a field trial of the integrated M&M intervention at a large prayer camp in Ghana. The objectives of the study were to (1) examine the safety and acceptability of the intervention among people with mental illness who are treated at the camp (hereinafter referred to as "patients") and the prayer camp staff who provide them with services (hereinafter referred to as "healers") and (2) evaluate the preliminary effectiveness of the intervention package.

Ethical Considerations

The study was approved by the institutional review boards of the University of Washington (00015549) and the University of Ghana (00001276). All study participants provided informed consent. Both healer and patient participants were compensated 75 Ghanaian Cedi (US $5.83) for their engagement in each assessment interview with study personnel.

Study Procedures

Initially, members of the investigative team met with the leader of the prayer camp (hereinafter referred to as "the Prophet") to describe the project (ie, objectives, procedures, and timeline), introduce members of the team, and demonstrate the M-Healer technology on a smartphone. During this meeting, the team was introduced to the 4 healers who would participate in the study. In a second meeting, the team met with the healers separately to discuss their routine practices at the camp, describe the "enhanced" services they would be trained to deliver in the context of the study, solicit feedback, cocreate the study's operational timeline, and address any questions or concerns.

Candidate patients for intervention were suggested by prayer camp staff and screened on-site by study team personnel. The initial screening was conducted by study doctors (psychiatry residents). Participants were evaluated for general medical and psychiatric fitness for participation. Medical assessment included measurement of temperature, pulse, BMI, blood pressure, random blood sugar, blood hemoglobin level, and malaria. The study inclusion criteria were as follows: (1) being aged 18 years or older; (2) speaking English or Twi; (3) being a current inpatient staying at a study prayer camp; (4) having a diagnosis of schizophrenia spectrum disorder (ie, schizophrenia, schizoaffective disorder, delusional disorder, or schizophreniform disorder), bipolar disorder, or major depressive disorder, as determined by study staff administering the relevant sections of the Structured Clinical Interview for the *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition) during screening. Candidates were excluded if they had a serious physical illness or needed urgent medical attention (eg, they had a high fever, serious infection, visible injury, or hemorrhage). Individuals who met the criteria and who were interested in participating were enrolled in the study. Participants were provided with a snack and meal by the study staff during assessment visits.

Intervention Description

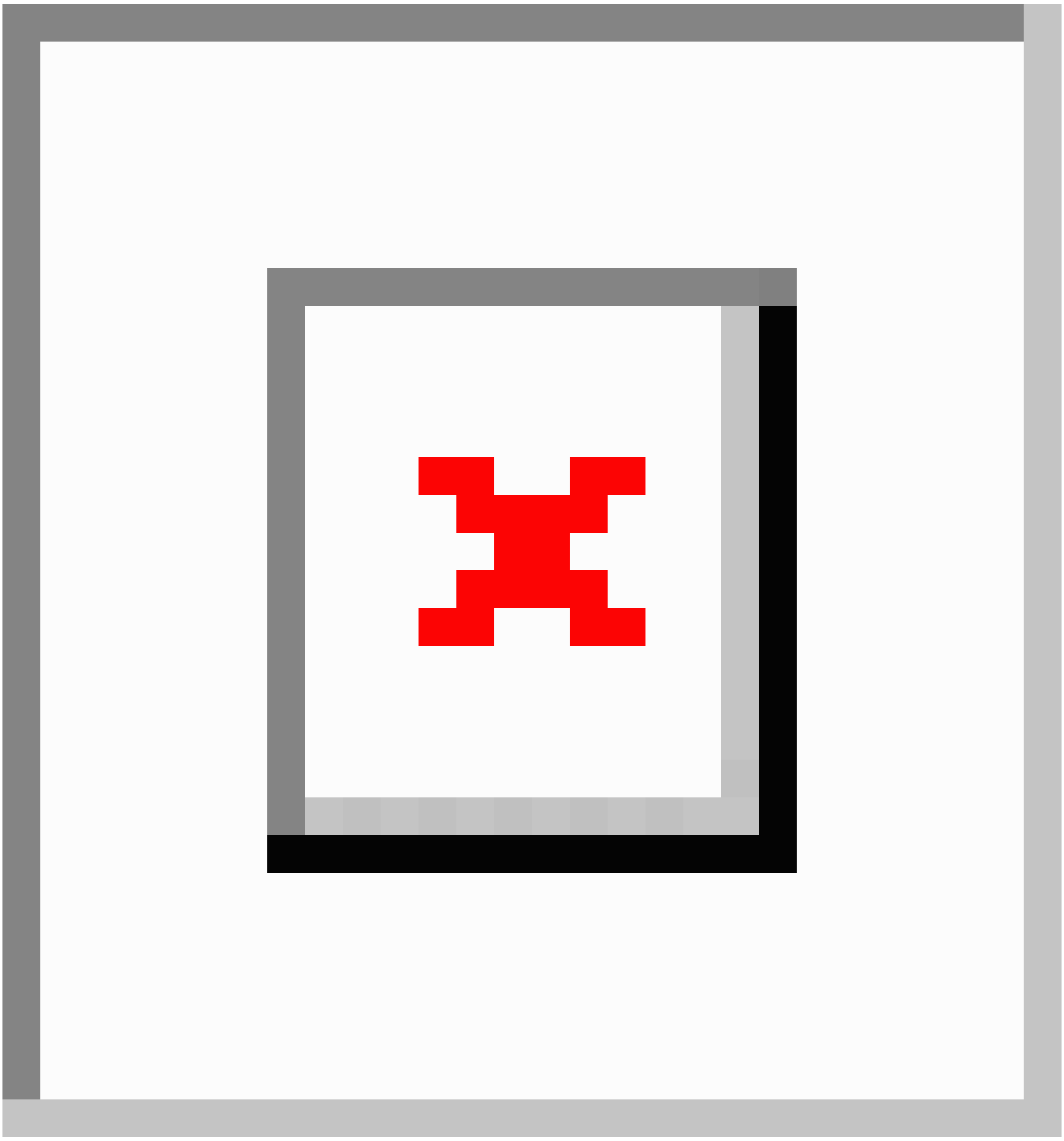
M&M is an 8-week-long dual-pronged intervention package comprising psychoeducation, skills training, and treatment support scaffolding tools delivered to healers via the M-Healer toolkit app, and pharmacotherapy that is administered directly to patients at prayer camps by a visiting mobile nurse.

Intervention Component 1: M-Healer Toolkit

M-Healer is an Android Smartphone app that was specifically designed to serve as a digital toolkit for healers providing care to people with mental illness in West Africa [16]. The goal of the toolkit is to provide exposure to psychoeducational materials and to support basic monitoring of patient progress. Once installed on an Android device, the app does not require an active data plan or internet connectivity to operate, mitigating possible disruptions in functionality in resource-constrained environments. M-healer is available in English and Twi, the 2 most common languages spoken in Ghana. The app has 4 core functionalities. First, M-Healer provides psychoeducation on the administration of psychosocial interventions such as guided relaxation techniques, rapport building, verbal de-escalation, challenging dysfunctional beliefs about psychiatric symptoms, and preservation of human rights and dignity in practice. All content is accessible "on demand" from the home screen and delivered as brief digital animations or audio recordings (Figure 1). Second, M-Healer allows healer users to create a list of active patients currently at the camp to support basic tracking and monitoring of individual progress. New patients are assigned a number, and healers have the option to also record an audio identifier for each individual. Every day, M-Healer prompts the healer to check in with each patient and provide a rating on whether they are doing better, worse, or the same as the day before. Ratings are assigned in the form of emojis with happy, sad, or neutral facial expressions. Once entered, a daily rating is added to a series of small icons showing patient progress over several days, which enables the healer to track patient progress over time (Figure 1). Third, M-Healer has push notifications, prompting users to watch the "video of the day" drawn from the bank of psychosocial digital animation training videos. Fourth, alongside these user-facing functionalities, M-Healer also tracks the frequency and duration of use in a given time frame, allowing the study staff and the nurse to monitor healers' interactions with the M-Healer toolkit.

Healer participants were provided Android smartphones with M-Healer installed and activated, which they were allowed to retain at the conclusion of the study. Healers were trained by 2 practice facilitators (study coinvestigators) and the study nurse on how to use M-Healer. Practice facilitation was conducted over 2 sessions in a designated space within the prayer camp, which was removed from the areas where patients resided. In the first session, the practice facilitators established rapport and trust, introduced the M-Healer toolkit, and taught some technical and clinical skills to the healer. This session was interactive and involved demonstrations of how to use the system and its features. Each healer had an opportunity to practice these features with the assistance of the facilitators and the nurse. To help the healers understand how they could use M-Healer skills in caring for their patients, facilitators led role-plays, which involved the use of preselected videos and providing real-time instruction as the healers navigated through their devices. Role-playing skills in the videos involved the participation of both cofacilitators or a cofacilitator and a volunteer who modeled as a healer and patient. Role-playing sessions on how to add patients and track their progress involved all healers. As the lead facilitator led the training, the other study team members walked around to assist healers who were experiencing navigational or technical difficulties. To conclude the session, healers were tasked to watch a video, listen to an audio module, and check in and track the progress of their patients on a daily basis.

Figure 1. M-Healer home screen, patient check-in page, and on-demand human rights module.



The second facilitation session was designed to strengthen rapport and trust, refine technical skills, and strengthen clinical skill acquisition. The facilitators and study nurse reviewed the healers' experiences of using M-Healer and their use data. Healers' efforts in using the app were also acknowledged and praised. Healers were then asked to demonstrate how to watch a video or listen to an audio module, add a new patient for daily monitoring, and record their patients' progress. The facilitators addressed any questions and concerns. At the end of the session, the healers were reminded to watch a video, listen to an audio module, and check in with their patients daily moving forward. After these 2 training sessions, the healers continued to receive ongoing M-Healer support from the study nurse during weekly site visits. During these visits, the nurse had opportunities to interact with the healers, inquire about and address their difficulties, check and record their user data, and encourage their use of the app.

Intervention Component 2: Mobile Nurse

A registered mental health nurse trained to use the World Health Organization's Mental Health Gap Action Programme–Intervention Guide (mhGAP-IG; version 2.0) served as the intervention's mobile nurse. Participants were prescribed medications by the screening study doctors. When possible, the prescribers opted to recommend long-acting injectable medications for participants with symptoms of psychosis to mitigate the risk of poor adherence to oral medication regimens. The use of long-acting injectable medications was facilitated by the fact that most of the participants had received treatment with antipsychotics prior to their entry into the camp. Participants ultimately decided whether they would take medication and specified their preferred medication modality. The nurse monitored participants' BMI, blood pressure, pulse, and temperature weekly and monitored their random blood sugar monthly. The nurse conducted weekly assessments of response to treatment and medication side effects using mhGAP-IG guidance. The nurse was supervised clinically by 2 study psychiatrists and instructed to call them if they encountered any urgent treatment-related problems that were not clearly described by the mhGAP-IG or the study protocol. The clinical team maintained active, real-time telecommunication to facilitate timely consultation and a weekly clinical team meeting to review nonurgent inquiries.

Data Collection

Study assessors (masters'-level clinical psychologists) administered a battery of measures to patient participants at baseline, midintervention (week 4), and post intervention (week 8). Assessments were administered in English or Twi using a tablet-based data collection software REDCap (Research Electronic Data Capture) to facilitate secure, password-protected, on-site digital data collection. We collected information on demographics, psychological symptoms, psychological distress, quality of life, and prayer camp experiences.

Patient participants' psychiatric symptoms were measured at baseline, midtreatment, and post treatment using three measures: (1) the Brief Psychiatric Rating Scale (BPRS) [17], an 18-item clinical assessor–rated measure of current indicators of psychiatric illness including somatic concern, anxiety, emotional withdrawal, conceptual disorganization, feelings of guilt, tension, mannerism and posturing, grandiosity, depressive mood, hostility, suspiciousness, hallucinatory behavior, motor retardation, uncooperativeness, unusual thought content, blunted affect, excitement, and disorientation; the study assessors rated each indicator from 1 ("not present") to 7 ("extremely severe"), with higher scores indicating more severe symptoms (range 18-126); (2) the Brief Symptom Inventory (BSI) [18], a 53-item self-report assessment measuring 9 symptom dimensions; patients rated how much each symptom has bothered them in the past 8 days from 0 ("not at all") to 4 ("extremely"), with higher scores indicating more severe distress over the past 7 days (range 0-212); and (3) the Patient Health Questionnaire-9 [19], a 9-item depressive symptom scale in which respondents self-report how often (1="not at all" to 4="nearly every day") they experience symptoms including feeling depressed and hopeless over the prior 2 weeks.

The following measures were also administered at baseline, midtreatment, and post treatment. Psychological distress was measured using the Talbieh Brief Distress Inventory [20], a 24-item self-report measure using a Likert scale for prior-month frequency rating (0="not at all" to 4="extremely"). Higher scores represent more severe distress. Quality of life was measured using the global quality-of-life items from the Lehman Quality of Life Interview for the Mentally Ill [21], whereby participants rate how they feel about various aspects of their life from 1 ("terrible") to 7 ("delighted"), and higher scores indicate higher quality of life. Shame was measured using the Other as Shamer Scale [22], an 18-item self-report (range 8-32; higher scores correspond to higher experiences of guilt and shame). Patient beliefs about medication were assessed using the Beliefs about Medications Questionnaire–General Harm subscale [23], internalized stigma was measured with the Brief Internalized Stigma of Mental Illness scale [24], and patient working alliance with the healers at the camp was assessed with an adapted version of the Working Alliance Inventory [25].

Human rights abuses were measured at baseline, midtreatment, and post treatment. At each assessment, study assessors interviewed patient participants and asked them to self-report on their experiences during the preceding week, up until and including the day of the assessment. Specially, they were asked to report the number of days they were chained or shackled (ie, "days chained"), the number of days they were forced to take herbal remedies, the number of hours they were held in isolation, and the total number of times they were touched on their genitals in an uncomfortable manner in the previous week.

To characterize the sample before treatment, we collected data on individuals' history of sexual abuse (Sexual Abuse Severity Scale) [26] substance use (Tobacco, Alcohol, Prescription Medications, and Other Substances scale) [27], and social support (Oslo Social Support Scale) [28] at baseline.

To measure the feasibility of the M-Healer intervention, we collected data post treatment on healers' use of the M-Healer technology, as well as knowledge and skills related to engagement with M-Healer. At each visit, study staff viewed the password-protected M-Healer data analytics screen on the healers' phones and recorded data including the number of days on which the app was used, the number of videos watched, the number of user-initiated app interactions, and the number of prompts to which the healers responded. To assess fidelity, healers participated in a 12-item verbally administered knowledge checklist whereby healers agreed or disagreed with statements regarding M-Healer knowledge domains. Healers performed 3 behavioral tests to demonstrate skills learned in M-Healer modules by interacting with the study staff acting as a standardized patients in role-played scenarios common to prayer camp settings. Study staff rated healer-demonstrated skills in a checklist.

Statistical Analysis

Descriptive statistics were used to characterize patient participants and their psychiatric and psychosocial factors at all study visits. We used paired samples *t* tests to compare baseline to midtreatment and baseline to posttreatment mean scores. For participants missing <50% of scale items for a psychosocial scale, item-level scores were imputed as the median score across the participant's existing scale items (person-median imputation) [29]. Psychosocial scores were not analyzed for participants missing ≥50% of scale items. Quantitative analyses were conducted using Stata (version 17; StataCorp).

Results

In total, 29 patients were screened for study eligibility among whom 19 (65.5%) were eligible to participate and 17 (89.5%) were enrolled (Figure 2). The majority of participants (n=15, 88.3%) were retained for midtreatment (4 weeks) and posttreatment (8 weeks) follow-up visits. Participants had an average age of 44.3 (SD 13.9) years; 59% (n=10) of them were male, and 82% (n=14) of them identified as Christian (Table 1). Fourteen (82%) participants had a diagnosis of schizophrenia, 1 had a comorbid diagnosis of major depressive disorder, and 2 were diagnosed with bipolar disorder. The average number of weeks spent residing in the camp was 72.6 (SD 111.9), and this was the first experience in a prayer camp for 41% (n=7) of participants.

We found a significant and clinically meaningful reduction in psychiatric symptom severity from pretreatment to posttreatment as measured by the BPRS and the BSI (Table 2). Subjective ratings of psychological distress and shame were significantly lower post treatment than before treatment. Participants had significantly reduced internalized stigma about their mental health conditions from pretreatment to posttreatment. We recorded a significant reduction in days chained from pretreatment to posttreatment.

Figure 2. Flowchart for prayer camp participant enrollment. *Reasons for ineligibility were not mutually exclusive. **Malaria Rapid Diagnostic Test positive: n=1; hemoglobin level < 8% and symptoms: n=1; pulse > 120 bpm and symptoms: n=2. ***Participants returned for the 8-week visit.

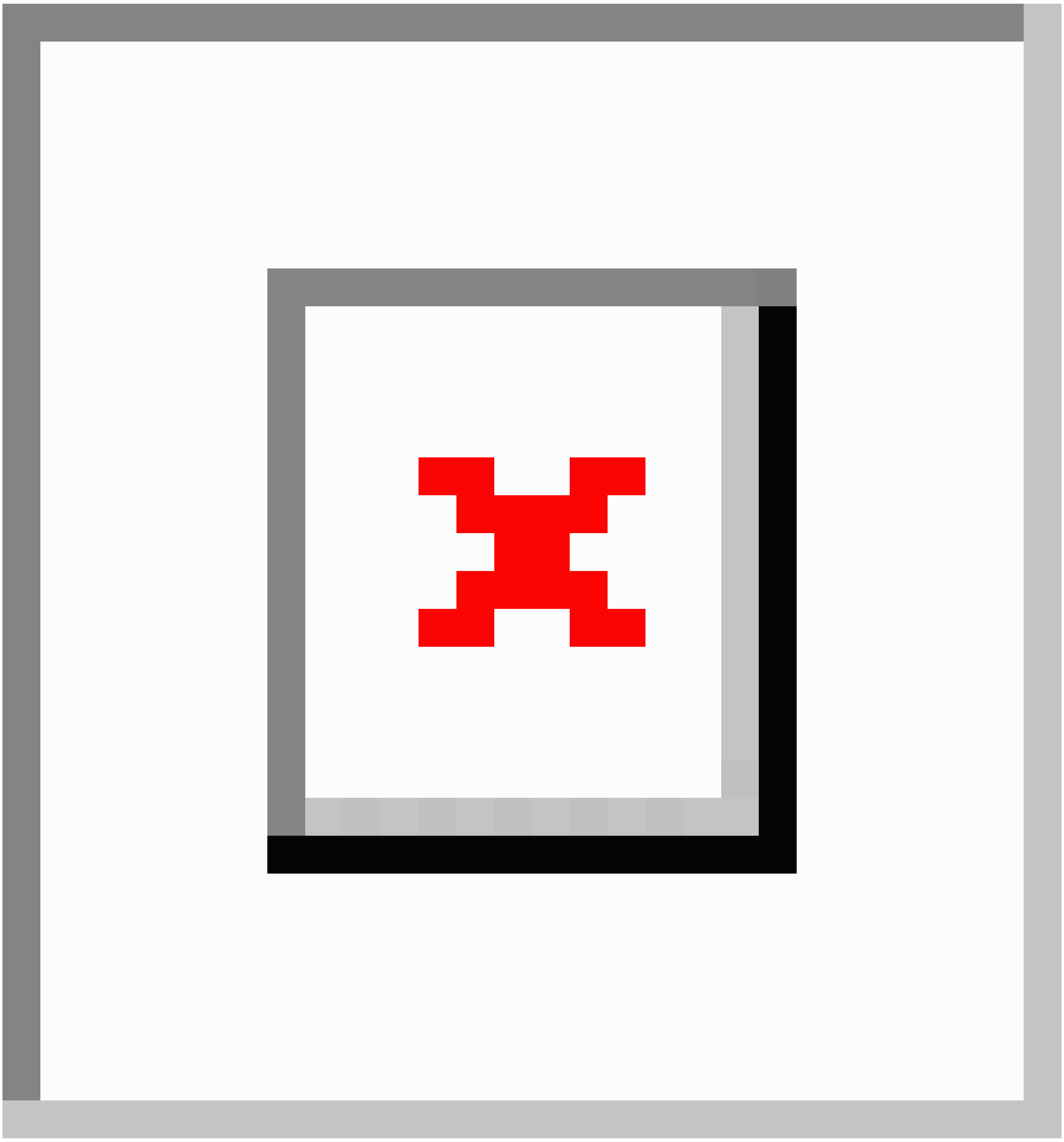


Table . Study sample characteristics (N=17).

Demographic characteristics		Value
Age (years; n=16), mean (SD)		44.3 (13.9)
Sex, n (%)		
	Female	7 (41.0)
	Male	10 (59.0)
Educational attainment, n (%)		
	Primary	2 (11.8)
	Junior high school	6 (35.3)
	Senior high school	5 (29.4)
	Tertiary or higher	4 (23.5)
Marital status, n (%)		
	Married	6 (35.0)
	Divorced or separated	2 (12.0)
	Single	9 (53.0)
Religion n (%)		
	Christian	14 (82.0)
	No religion	1 (6.0)
	Other	1 (6.0)
	Missing	1 (6.0)
Weeks residing in this prayer camp (n=14), mean (SD)		72.6 (111.9)
First experience residing in a prayer camp		
	Yes	7 (41)
	No	8 (47)
	Missing data	2 (12)
Social support[a] (n=12), mean (SD)		7.8 (2.6)
Ever use tobacco[b], n (%)		2 (12)
Ever use alcohol[b], n (%)		2 (12)
Ever use drugs[b], n (%)		2 (12)
Ever use prescription drugs[b], n (%)		0 (0)
Psychiatric diagnosis n (%)		
	Schizophrenia	14 (82.0)
	Bipolar disorder	2 (12.0)
	Major depressive disorder (comorbid with a primary diagnosis of schizophrenia)	1 (6.0)

[a]3-item Oslo Social Support Scale; absolute range 3-12; a higher score indicates a higher level of social support.
[b]Tobacco, alcohol, prescription medications, and other substances scale.

Table . Summary of M&M treatment outcomes.

	Pretreatment baseline assessment (n=17)		Midtreatment 4-week assessment (n=15)			Posttreatment 8-week assessment (n=15)		
	N	Mean (SD)	N	Mean (SD)	*P* value	N	Mean (SD)	*P* value
BPRS[a]	16	52.3 (24.0)	15	41.4 (10.6)	*.01*[b]	15	30.9 (8.7)	*<.001*
BSI[c]	16	76.4 (56.6)	14	53.9 (42.7)	*.004*	13	27.9 (23.7)	*<.001*
TBDI[d]	15	37.7 (23.5)	13	27.8 (20.9)	*.002*	14	16.9 (12.9)	*<.001*
OAS[e]	15	41.9 (21.3)	12	35.1 (20.6)	.64	13	28.5 (21.5)	*.01*
BMQ-Harm[f]	14	11.8 (3.6)	12	10.0 (4.1)	*.007*	12	10.3 (4.9)	.40
ISMI-10[g]	14	30.4 (3.9)	—[h]	—	—	13	24.9 (5.9)	*<.001*
PHQ-9[i]	15	10.4 (6.5)	14	9.2 (7.9)	.35	15	8.5 (8.5)	.21
Lehman Quality of Life Scale for the Mentally Ill[j]	14	20.1 (7.8)	11	23.5 (10.0)	.40	15	22.5 (9.2)	.26
Working Alliance Inventory–Bond Subscale[k]	13	14.8 (5.6)	11	13.8 (5.1)	.80	12	15.4 (3.9)	.59
Days chained[l]	16	1.6 (2.8)	14	0.5 (1.9)	.09	15	0.5 (1.8)	*.047*
Days forced to fast[l]	16	2.6 (5.5)	14	0.3 (1.1)	.12	15	0.0 (0.0)	.06
Days forced to take herbal remedies[l]	16	0.06 (0.3)	12	0.6 (2.0)	.40	15	0.0 (0.0)	.32
Hours kept isolated[l]	16	3.0 (8.2)	14	0.0 (0.0)	.32	15	0.0 (0.0)	.14
Times touched in an uncomfortable manner[l]	16	0.1 (0.3)	14	0.0 (0.0)	.14	15	0.0 (0.0)	.14

[a]BPRS: Brief Psychiatric Rating Scale—an 18-item scale with an absolute range of 18-126; higher scores indicate higher symptom severity.
[b]Italicized values are statistically significant at *P*<.05.
[c]BSI: Brief symptom inventory—a 53-item scale with an absolute range of 0-212; higher scores indicate higher symptom severity.
[d]TBDI: Talbieh Brief Distress Inventory—a 24-item scale with an absolute range of 0-96; higher scores indicate greater distress.
[e]OAS: Other as Shamer Shame Scale—an 18-item scale with an absolute range of 0-90; higher scores indicate greater shame.
[f]BMQ-Harm: Beliefs about Medications Questionnaire–Harm Subscale—a 4-item scale with an absolute range of 4-20; higher scores indicate unfavorable thoughts about medications.
[g]ISMI-10: Internalized Stigma of Mental Illness Inventory—a 10-item scale with an absolute range of 10-40; higher scores indicate greater internalized stigma.
[h]Not available.
[i]PHQ-9: Patient Health Questionnaire-9—a 9-item scale with an absolute range of 0-27; higher scores indicate higher symptom severity.
[j]A 6-item scale with an absolute range of 6-42; higher scores indicate better quality of life.
[k]A 4-item scale with an absolute range of 4-20; higher scores indicate better working alliance.
[l]Frequency over the past week.

Psychiatric symptoms measured with the BPRS and BSI were reduced midtreatment compared to those at baseline. Unfavorable beliefs about medications had decreased by the midtreatment visit compared to those at baseline. We observed a trend toward statistical significance for reduction in forced fasting by the posttreatment follow-up visit (*P*=.06). Over the intervention period, we did not identify changes in patient-reported working alliance with healers, depressive symptom severity, quality of life, days forced to take herbal remedies, hours retained in isolation, or number of times a participant reported being touched in an uncomfortable manner.

Results with imputed variables were not meaningfully different from those obtained from raw data. Notably, a significant decrease in unfavorable beliefs about medications from pretreatment to posttreatment was identified in the nonimputed data set; yet, this change was not detected in the imputed data set. Participants' scales with imputed values were predominantly only missing 1-2 items, which were imputed using person-median values per scale to retain the available data.

In the initial visit to the prayer camp, 15 participants consented to receiving pharmacotherapy. For schizophrenia treatment, fluphenazine decanoate (25 mg/mL) was the most commonly prescribed pharmacotherapy (n=12, 80.0%), followed by trihexyphenidyl (5 mg tablet; n=6, 40.0%) and risperidone (2 mg; n=3, 20.0%); some participants were prescribed multiple medications. Among participants with bipolar disorder, fluphenazine decanoate (25 mg), risperidone (25 mg), and trihexyphenidyl (5 mg) were each prescribed to 6.7% (n=1) of participants. Overall, 110 medication follow-up visits were conducted by the mobile nurse.

Four healers were trained to use M-Healer and their usage of the app was monitored weekly. M-Healer feasibility measures indicated that healers had an average of 31.9 (SD 28.9) user-initiated app uses per weekly visit, watched 19.1 (SD 21.2) videos, responded to 1.5 (SD 2.4) prompts, and used the app for 5.3 (SD 2.7) days a week. Healers had posttreatment average knowledge assessment checklist scores of 11.75 (absolute range 9-14, total possible score=16). Two healers successfully demonstrated all 3 behavioral skills, 1 healer demonstrated 2 skills, and 1 healer demonstrated 1 skill.

Discussion

Principal Results

The combination treatment package deployed in this study was designed to inform and shape West African healers' practices and to provide symptomatic relief to individuals with psychiatric illnesses receiving services at their prayer camps. The intervention proved to be feasible and acceptable to healers and their patients. Healers were able to navigate the M-Healer system successfully on the study smartphone devices that were provided to them. Healers expressed a clear understanding of M-Healer functionalities, watched psychoeducational videos, and listened to the audio lessons over the study period. Healers in the study self-initiated M-Healer use throughout the intervention period above and beyond system-prompted interactions and chose to view M-Healer videos and listen to audio lessons at a rate that far exceeded our expectations. Healers' use of the app was very frequent—almost daily. When evaluated, healers were able to describe the content accurately and to demonstrate skills that they had learned from the app.

Healers agreed to grant the mobile nurse access to the individuals they treated at the prayer camp, the majority of whom were placed in locked dormitories, with some shackles. The nurse was able to assess, provide pharmacotherapy to, and monitor these patients weekly without difficulty or obstruction, despite these austere conditions. All patients approached by our study team expressed openness to meeting with the nurse, and all but 1 expressed interest in receiving pharmacotherapy to manage their psychiatric symptoms and improve their health.

The combined intervention proved to be clinically promising. Patient participants experienced significant reductions in the severity of their psychiatric symptoms, psychological distress, subjective feelings of shame, and internalized stigma over the course of the intervention. Moreover, patient participants reported a significant reduction in the days during which they were chained or shackled by the healers and a nonsignificant trend of reduction in the days during which they were forced to fast. These findings suggest that exposure to the training content delivered by the M-Healer toolkit may have shifted healer behaviors after the intervention was implemented at the prayer camp's study site. One participant was discharged from the prayer camp and allowed to return to their family over the course of the intervention by prayer camp staff.

The intervention proved to be safe. The study helped promote better care in the prayer camp's study site. Over the course of the intervention, 3 participants were identified by our study staff as requiring immediate medical attention and were referred to the district hospital where they received care. The nurse monitored medication side effects weekly. No major side effects, health and safety violations, serious adverse events, or other major complications occurred over the course of the field trial.

Poststudy debriefing interviews with healers indicated that they felt some ambivalence about M-Healer content. On one hand, they enjoyed the opportunity to learn about novel psychosocial illness management strategies and indicated that the digital toolkit informed the way they engaged with their patients. On the other, some expressed frustration that the toolkit's human rights module seemed too critical of their practices (eg, M-Healer's guidance to refrain from chaining patients). Given that current prayer camp practices do not prioritize preservation of human rights, this tension may have been inevitable. A more complete report on the findings from these interviews will be provided in a separate qualitative study.

Limitations

The study has several limitations. First, the study involved patient participants and healers at a single prayer camp, limiting study generalizability. Prayer camps vary dramatically in terms of their size, practices, and openness to research and outsider involvement in their day-to-day activities. Future multisite research in several prayer camps can help determine the generalizability of our findings to other settings. Second, we relied on study participants' self-reports to determine whether they experienced human rights abuses at the camp. Participants may have underreported these events due to shame, fear of retaliation, or to protect camp staff. Third, the study sample was small, allowing only for preliminary examination of treatment effects. Despite the exploratory nature of our evaluation of outcomes, promising evidence of clinical effectiveness was recorded. Future research involving larger samples will facilitate more robust evaluation of the effects of the intervention on participants' health and well-being. Finally, prayer camp staff had previous experience being involved in research with our team. It will be important to examine the intervention in research-naïve prayer camps to evaluate whether such relationships are necessary for successful intervention uptake and outcomes.

Conclusions

In West Africa, traditional and faith healers greatly outnumber trained mental health professionals. Healers hold political capital and are respected by the local population. These key stakeholders can serve as facilitators, conduits, obstructionists, or blockades to the delivery of evidence-based mental health services in their communities. Engaging healers and training them to provide compassionate psychosocial interventions themselves and opening the doors of their prayer camps to medical professionals who can also deliver pharmacological interventions have the potential to enhance the regional capacity to address unmet mental health needs.

Previous research has demonstrated that pharmacotherapy can be conducted successfully at prayer camps. While psychotropics may help effectively manage the psychiatric symptoms of patients staying at prayer camps, those improvements did not translate to reductions in the days during which patients were shackled [14]. This study demonstrated that adding a key component—a healer-facing digital health training toolkit—can impact how healers interact with their patients, including a reduction in the use of mechanical restraints. These findings are encouraging and suggest that the M&M intervention holds promise as a dual-pronged model for both improving mental health outcomes and reducing human rights abuses in West African prayer camps. If future research finds similarly robust treatment effects on a larger scale, this will provide important information for Ghanian policy makers. Such findings would demonstrate that a combination of accessible technologies coupled with individuals trained to deliver interventions in the field may help reconfigure how prayer camps serve people with SMI. Ghanaian government mandates, monitoring, and enforcement over the last few years are beginning to produce positive effects on the reduction of human rights abuses in prayer camps. In addition to barring harmful practices, the government can also play a key role in supporting the adoption of new models of care. Through targeted trainings and appropriate oversight, individuals who are already working in the field (eg, government-employed district nurses or members of the Psych Corps—psychology graduates who are posted to health facilities across the country as part of their national service) could be leveraged to train healers and other paraprofessionals in the use of evidence-based digital mental health technologies. Government investment in such force-multiplying activities and technologies can prove to be a pragmatic and scalable approach for addressing significant professional workforce shortages in the region.

Acknowledgments

The study was supported by a grant award from the US National Institute of Mental Health (R01MH127531-01A1; DBZ and AOA). The authors thank the camp leader for permitting the study to be conducted at his faith healing camp and for the collaboration he fostered between camp staff and the study team. The authors acknowledge and thank the head of the district health management team for sending a district nurse to the study site to engage with the study nurse to facilitate continuation of pharmacological care, the district health management team that facilitated the acquisition of national health insurance cards for the study, and the study participants.

Conflicts of Interest

DBZ has financial interests in Merlin LLC and FOCUS technology. He has had an intervention content licensing agreement with Pear Therapeutics and has provided consultation services to Trusst Health, K Health, Boehringer Ingelheim, eQuility, Deep Valley Labs, and Otsuka Pharmaceuticals. KO and SO have provided consultation services to Janssen pharmaceuticals.

References

1. Esan O, Abdumalik J, Eaton J, Kola L, Fadahunsi W, Gureje O. Global mental health reforms: mental health care in anglophone West Africa. Psychiatr Serv 2014 Sep;65(9):1084-1087. [doi: 10.1176/appi.ps.201300300] [Medline: 25179185]
2. Ofori-Atta A, Read U, Lund C. A situation analysis of mental health services and legislation in Ghana: challenges for transformation. Afr J Psych 2010;13(2):99-108. [doi: 10.4314/ajpsy.v13i2.54353]
3. Gureje O, Nortje G, Makanjuola V, Oladeji BD, Seedat S, Jenkins R. The role of global traditional and complementary systems of medicine in the treatment of mental health disorders. Lancet Psychiatry 2015 Feb;2(2):168-177. [doi: 10.1016/S2215-0366(15)00013-9] [Medline: 26359753]
4. Ae-Ngibise K, Cooper S, Adiibokah E, et al. 'Whether you like it or not people with mental problems are going to go to them': a qualitative exploration into the widespread use of traditional and faith healers in the provision of mental health care in Ghana. Int Rev Psychiatry 2010;22(6):558-567. [doi: 10.3109/09540261.2010.536149] [Medline: 21226644]
5. Roberts H. A way forward for mental health care in Ghana? Lancet 2001 Jun 9;357(9271):1859. [doi: 10.1016/S0140-6736(00)05020-0] [Medline: 11410206]
6. Roberts M, Mogan C, Asare JB. An overview of Ghana's mental health system: results from an assessment using the World Health Organization's Assessment Instrument for Mental Health Systems (WHO-AIMS). Int J Ment Health Syst 2014;8:16. [doi: 10.1186/1752-4458-8-16] [Medline: 24817908]

7. Read U. “I want the one that will heal me completely so it won't come back again”: the limits of antipsychotic medication in rural Ghana. Transcult Psychiatry 2012 Jul;49(3-4):438-460. [doi: 10.1177/1363461512447070] [Medline: 22722982]
8. Patel V. Traditional healers for mental health care in Africa. Glob Health Action 2011;4. [doi: 10.3402/gha.v4i0.7956] [Medline: 21845145]
9. Esan O, Appiah-Poku J, Othieno C, et al. A survey of traditional and faith healers providing mental health care in three sub-Saharan African countries. Soc Psychiatry Psychiatr Epidemiol 2019 Mar;54(3):395-403. [doi: 10.1007/s00127-018-1630-y] [Medline: 30456425]
10. “Like a Death Sentence”: Abuses against Persons with Mental Disabilities in Ghana. Human Rights Watch. 2012. URL: https://www.hrw.org/report/2012/10/02/death-sentence/abuses-against-persons-mental-disabilities-ghana#:~:text=People%20with%20mental%20disabilities%20suffer,their%20right%20under%20international%20law [accessed 2024-03-14]
11. Living in Chains: Shackling of People with Psychosocial Disabilities Worldwide. Human Rights Watch. 2020. URL: https://www.hrw.org/report/2020/10/06/living-chains/shackling-people-psychosocial-disabilities-worldwide [accessed 2024-03-14]
12. Drew N, Funk M, Tang S, et al. Human rights violations of people with mental and psychosocial disabilities: an unresolved global crisis. Lancet 2011 Nov 5;378(9803):1664-1675. [doi: 10.1016/S0140-6736(11)61458-X] [Medline: 22008426]
13. Mental health and human rights. Report of the United Nations High Commissioner for Human Rights. United Nations General Assembly. 2017. URL: https://www.un.org/disabilities/documents/reports/ohchr/a_hrc_34_32_mental_health_and_human_rights_2017.pdf [accessed 2023-09-28]
14. Ofori-Atta A, Attafuah J, Jack H, Baning F, Rosenheck R, Joining Forces Research Consortium. Joining psychiatric care and faith healing in a prayer camp in Ghana: randomised trial. Br J Psychiatry 2018 Jan;212(1):34-41. [doi: 10.1192/bjp.2017.12] [Medline: 29433613]
15. Ben-Zeev D. Mobile health for mental health in West Africa: the case for Ghana. Psychiatr Serv 2018 Jul 1;69(7):741-743. [doi: 10.1176/appi.ps.201700555] [Medline: 29540120]
16. Ben-Zeev D, Meller S, Snyder J, et al. A digital toolkit (M-Healer) to improve care and reduce human rights abuses against people with mental illness in West Africa: user-centered design, development, and usability study. JMIR Ment Health 2021 Jul 2;8(7):e28526. [doi: 10.2196/28526] [Medline: 34255712]
17. Overall JE, Gorham DR. The Brief Psychiatric Rating Scale (BPRS): recent developments in ascertainment and scaling. Psychopharmacol Bull 1988;24(1):97-99.
18. Derogatis LR. Brief symptom inventory. Eur J Psychol Assess 1975.
19. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
20. Ritsner M, Rabinowitz J, Slyuzberg M. The Talbieh brief distress inventory: a brief instrument to measure psychological distress among immigrants. Compr Psychiatry 1995;36(6):448-453. [doi: 10.1016/s0010-440x(95)90253-8] [Medline: 8565450]
21. Lehman AF. A quality of life interview for the chronically mentally ill. Eval Program Plan 1988 Jan;11(1):51-62. [doi: 10.1016/0149-7189(88)90033-X]
22. Allan S, Gilbert P, Goss K. An exploration of shame measures: II: psychopathology. Pers Individual Differ 1994 Nov;17(5):719-722. [doi: 10.1016/0191-8869(94)90150-3]
23. Andersson Sundell K, Jönsson AK. Beliefs about medicines are strongly associated with medicine-use patterns among the general population. Int J Clin Pract 2016 Mar;70(3):277-285. [doi: 10.1111/ijcp.12781] [Medline: 26916721]
24. Boyd JE, Otilingam PG, Deforge BR. Brief version of the Internalized Stigma of Mental Illness (ISMI) scale: psychometric properties and relationship to depression, self esteem, recovery orientation, empowerment, and perceived devaluation and discrimination. Psychiatr Rehabil J 2014 Mar;37(1):17-23. [doi: 10.1037/prj0000035] [Medline: 24660946]
25. Horvath AO, Greenberg LS. Development and validation of the working alliance inventory. J Couns Psychol 1989;36(2):223-233. [doi: 10.1037//0022-0167.36.2.223]
26. Swinson KC. The Sexual Assault Severity Scale: a comprehensive measure of assault severity.: Aquila; 2013.
27. McNeely J, Wu LT, Subramaniam G, et al. Performance of the Tobacco, Alcohol, Prescription medication, and other Substance use (TAPS) Tool for substance use screening in primary care patients. Ann Intern Med 2016 Nov 15;165(10):690-699. [doi: 10.7326/M16-0317] [Medline: 27595276]
28. Kocalevent RD, Berg L, Beutel ME, et al. Social support in the general population: standardization of the Oslo Social Support Scale (OSSS-3). BMC Psychol 2018 Jul 17;6(1):31. [doi: 10.1186/s40359-018-0249-9] [Medline: 30016997]
29. Bono C, Ried LD, Kimberlin C, Vogel B. Missing data on the center for epidemiologic studies depression scale: a comparison of 4 imputation techniques. Res Social Adm Pharm 2007 Mar;3(1):1-27. [doi: 10.1016/j.sapharm.2006.04.001] [Medline: 17350555]

Abbreviations

BPRS: Brief Psychiatric Rating Scale
BSI: Brief Symptom Inventory

mhGAP-IG: Mental Health Gap Action Programme–Intervention Guide
REDCap: Research Electronic Data Capture
SMI: serious mental illness

Edited by J Torous; submitted 25.09.23; peer-reviewed by E Sackey, S Joshi; revised version received 21.02.24; accepted 22.02.24; published 27.03.24.

Please cite as:
Ben-Zeev D, Larsen A, Attah DA, Obeng K, Beaulieu A, Asafo SM, Gavi JK, Kadakia A, Sottie EQ, Ohene S, Kola L, Hallgren K, Snyder J, Collins PY, Ofori-Atta A, M-Healer Research Team
Combining mHealth Technology and Pharmacotherapy to Improve Mental Health Outcomes and Reduce Human Rights Abuses in West Africa: Intervention Field Trial
JMIR Ment Health 2024;11:e53096
URL: https://mental.jmir.org/2024/1/e53096
doi:10.2196/53096

Efficacy of BrighterSide, a Self-Guided App for Suicidal Ideation: Randomized Controlled Trial

Natasha Josifovski[1,2], PhD; Michelle Torok[1,2], PhD; Philip Batterham[3], PhD; Quincy Wong[4], PhD; Joanne R Beames[1,2], PhD; Adam Theobald[1], BPsyc (Hons), MPsyc (Clin); Sarah Holland[1], BComm; Kit Huckvale[5], PhD; Jo Riley[6], BSc; Nicole Cockayne[1,2], MPH, BSc (Hons); Helen Christensen[1,2], PhD; Mark Larsen[1,2], DPhil

1

2

3

4

5

6

Corresponding Author:
Natasha Josifovski, PhD

Abstract

Background: Self-guided digital interventions can reduce the severity of suicidal ideation, although there remain relatively few rigorously evaluated smartphone apps targeting suicidality.

Objective: This trial evaluated whether the BrighterSide smartphone app intervention was superior to a waitlist control group at reducing the severity of suicidal ideation.

Methods: A total of 550 adults aged 18 to 65 years with recent suicidal ideation were recruited from the Australian community. In this randomized controlled trial, participants were randomly assigned to receive either the BrighterSide app or to a waitlist control group that received treatment as usual. The app was self-guided, and participants could use the app at their own pace for the duration of the study period. Self-report measures were collected at baseline, 6 weeks, and 12 weeks. The primary outcome was severity and frequency of suicidal ideation, and secondary outcomes included psychological distress and functioning and recovery. Additional data were collected on app engagement and participant feedback.

Results: Suicidal ideation reduced over time for all participants, but there was no significant interaction between group and time. Similar improvements were observed for self-harm, functioning and recovery, days out of role, and coping. Psychological distress was significantly lower in the intervention group at the 6-week follow-up, but this was not maintained at 12 weeks.

Conclusions: The BrighterSide app did not lead to a significant improvement in suicidal ideation relative to a waitlist control group. Possible reasons for this null finding are discussed.

Trial Registration: Australian New Zealand Clinical Trials Registry ACTRN12621000712808; https://trialsearch.who.int/Trial2.aspx?TrialID=ACTRN12621000712808

(JMIR Ment Health 2024;11:e55528) doi:10.2196/55528

KEYWORDS

suicidal ideation; suicide prevention; digital health; clinical trial

Introduction

One in 6 adults experience serious thoughts of suicide during their lives, with an estimated prevalence of 3.4% of adults experiencing suicidal ideation in a 12-month period [1]. Data from the World Health Organization suggest that those with suicidal ideation are 10 times more likely to make a suicide attempt across their lifetime than those without suicidal ideation [2]. In recent years, there has been a shift in how people seek mental health support for a suicidal crisis. For example, during the 2021-2022 financial year, Lifeline—an Australian crisis service—answered 1,142,234 calls, a 56% increase since 2019 [3]. At the same time, hospitals in Australia experienced a 14.3% reduction in mental health–related presentations [4]. One potential way to offset the high demand for crisis support services is to make self-guided digital health interventions publicly available for those in distress. These interventions have the potential to increase access to mental health care and enhance the capacity of mental health systems to respond to persons in crisis by offering high-fidelity, evidence-based therapeutic support, anonymously and at low to no cost, which can be readily accessed anywhere [5,6].

Meta-analytic evidence suggests that digital health interventions can effectively reduce the severity of suicidal thoughts [7].

Those interventions that specifically target suicidality are more effective than generalized mental health apps for reducing suicide-related outcomes [7]. However, despite the huge potential for self-guided smartphone interventions to address service access gaps for those experiencing suicidal distress, there are currently few digital interventions targeting suicidal ideation that have been rigorously tested in randomized controlled trials (RCTs). Even fewer of these have been in general adult populations.

One exception to the above is the web-based self-help program Living with Deadly Thoughts, which was adapted from the Dutch program Living Under Control [8]. Drawing from principles of cognitive behavioral therapy and dialectical behavior therapy, this RCT involved access to 6 online modules for community-recruited adults experiencing suicidal ideation. While the Dutch program found a small but significant effect in reducing suicidal thoughts, the English-adapted Living with Deadly Thoughts program found no difference between intervention and control groups in an Australian study [8]. One possibility for this discrepancy may be that the English study was underpowered to detect an effect size that would be comparable with the Dutch version. However, there may be additional nuances given the difference in the recruited population—the Dutch study included only participants who experienced mild to moderate suicidal thoughts (defined as a score between 1 and 26 on the Beck Scale for Suicidal Ideation) and who were not severely depressed (defined as a scores greater than 39 on the Beck Depression Inventory). Conversely, Living with Deadly Thoughts did not include any cutoffs for either depression or suicidal thoughts as exclusion criteria. An effect on suicidal ideation could therefore be dependent on the severity of suicidal thoughts, in that those with thoughts that are more severe may not benefit from the modules involved. Alternatively, or as well as this, the program may involve other factors that are not generalizable, and an effective program to reduce suicidal ideation may require greater insight from those with lived experience.

In the context of these gaps, we developed BrighterSide. BrighterSide is a self-guided app based on a mix of cognitive behavioral therapy and dialectical behavior therapy, alongside elements of acceptance and commitment therapy and positive psychology. These therapeutic approaches have previously been demonstrated to be efficacious in reducing suicidal ideation when delivered through web-based programs [9,10]. The app involves 5 modules that each contain activities to encourage the user to develop and practice strategies to manage their suicidal thoughts. None of the activities last more than 5 minutes and users have complete control over which modules and activities they wish to complete. Furthermore, the app was co-designed with lived experience advisors in order to consider how best to maximize engagement and how best to support those with suicidal ideation or behaviors (see Torous et al [11] for a discussion on how co-design with consumers may enhance app engagement).

This study aimed to determine if those using the BrighterSide app would have a greater reduction in suicidal thoughts relative to the waitlist control group in a community trial in the adult population. We hypothesized that those in the intervention group would demonstrate significantly lower suicidal ideation, the primary outcome, at 6 and 12 weeks after baseline compared to a waitlist control group. We also hypothesized that the intervention group would report fewer incidents of self-harm and suicide attempts at 6 and 12 weeks after baseline and that the intervention group would report greater improvements in their ability to cope at 6 and 12 weeks after baseline.

Methods

Ethical Considerations

This study is reported as per the Consolidated Standards of Reporting Trials (CONSORT; Checklist 1) guideline (complete supplementary information is provided in Multimedia Appendix 1). The trial protocol was approved by the University of New South Wales Human Research Ethics committee (HC210196) and prospectively registered on the Australian New Zealand Clinical Trials Registry (ACTRN12621000712808).

Trial Design

The trial was a single-blind, 2-arm parallel RCT. Participants were randomized 1:1 between intervention and waitlist-control groups. Researchers were blind to group allocation. Those allocated to the intervention group received immediate access to the BrighterSide app upon completing the baseline survey. Those allocated to the control group received access to the app at the end of the trial.

Participants

Individuals were eligible for the study if they were (1) aged 18 to 65 years, (2) had experienced suicidal ideation within the previous 2 weeks, (3) owned an iPhone (with iOS 13.0 or higher) or Android (with Android 6.0 or higher) smartphone, (4) were fluent in English, and (5) currently lived in Australia. There were no specific exclusion criteria as this enabled a more heterogeneous sample.

Recruitment took place in 2 waves: the first was from June 30, 2021, to July 23, 2021, and the second was from September 20, 2021, to September 24, 2021 (see the Sample Size section). All data collection was completed by December 24, 2021. Participants were recruited via multiple channels. First, the trial was advertised on the Black Dog Institute (BDI) website. Second, recent visitors to the Black Dog Institute Online Clinic (a free mental health assessment tool) who had indicated recent suicidal ideation and consented to be contacted for future research opportunities were sent an invitation email by the clinic team. Third, the study was advertised on Facebook and Instagram via the BDI and Lifeline social media channels and paid advertising.

All individuals who responded to a study advertisement first completed an online screening questionnaire to determine eligibility. Those who screened as ineligible were redirected to a web page with information on crisis services. Eligible participants were presented with the participant information statement and digital consent form. Consenting participants then completed baseline questionnaires online via BDI's bespoke trial software, upon completion of which participants were randomized. Randomization was performed by the trial management software using a block size of 4. Participants were notified of group allocation, but investigators remained blinded.

Participants in the intervention group were sent an email on baseline completion with a link to download the BrighterSide app from the Apple App Store or Google Play Store, along with a unique link that provided access to the app content. This prevented use of the app by users who were not registered to the trial. There was no specific timeframe to download the app, nor were any reminders sent if participants did not download it. Participants in the control group were sent an email with the link to access the app on completion of the final 12-week questionnaire (or, if not completed, at 13 weeks). Participants did not receive reimbursement or incentives to participate in the study. Participants were permitted to engage in other treatment, preexisting or new, while participating in the study.

Intervention

BrighterSide is a self-guided smartphone app for adults experiencing suicidal ideation. The app was designed to help users develop and practice strategies to manage suicidal thoughts and was derived from content from the Living with Deadly Thoughts online program [8]. A multidisciplinary team of clinicians, researchers, lived experience advisors, designers, and developers engaged in a human-centered design process to update the original content into a more engaging form, ensuring the language, design, and user journey within the app were clear, simple, and supportive for adults experiencing suicidal thinking. This multidisciplinary collaboration was engaged across all functions of the app, including the safety planning and check-in features.

The app contained 5 modules: Understand your Thoughts, Prevent a Crisis, Navigate your Emotions, Navigate your Thoughts, and Plan for the Future (see Table S1 in Multimedia Appendix 1 for brief details of each module, and Multimedia Appendix 2 for selected screenshots of each module). Each module contained interactive activities (eg, guided breathing). The content was based on cognitive behavioral therapy and dialectical behavior therapy, with elements of acceptance and commitment therapy and positive psychology. Users could access content in any order they wished or could choose a guided option that progressed through the modules in a specific order. The app included a safety planning function where users could list warning signs, helpful techniques, and supportive contact details and share them with others via email. The app also included a daily "check-in" feature that asked whether the user was safe and linked them to their safety plan and crisis contact numbers. Finally, in addition to the modules, the app also included "distraction activities," such as Bubble Pop, a built-in game simulating the motion of popping bubble wrap, and "calming activities," such as guided mindfulness recordings. Participants were free to engage with the app in whichever way they chose, including after the final data collection at 12 weeks, and the trial did not mandate frequency or patterns of use. A guided option was available if participants preferred, which allowed them to prioritise modules in a specific order depending on their main concern. The intervention did not include reminders to use the app. Participants in the waitlist control group received an email with details of crisis support services, and access to the app was granted after the 12-week study period.

Outcomes

Standard demographic data were collected at baseline, with outcome measures collected at baseline, 6 weeks, and 12 weeks. Self-report questionnaires were administered online via the BDI's Research Engine platform, and participants were sent a link via email at 6 and 12 weeks to complete the follow-up time points.

The primary outcome was the frequency and severity of suicidal ideation, as measured by the Suicidal Ideation Attributes Scale (SIDAS) [12]. This comprises 5 items measuring frequency of ideation, controllability, closeness to attempt, level of distress, and the impact on daily functioning. The total score is in the range 0 to 50, with higher scores indicating higher levels of suicidal ideation. Scores of 21 or higher are indicative of high risk for suicidal behavior [12]. The Cronbach α for this study was α=.803.

Secondary outcomes included self-harm behavior, coping strategies (abbreviated Coping Orientation to Problems Experienced Inventory; Brief-COPE [13]) functioning and recovery (Functioning and Recovery Scale; FRS [14]), psychological distress (Distress Questionnaire-5; DQ5 [15]), help-seeking (modified Actual Health-Seeking Questionnaire; ASHQ [16]), and days out of role (WHO Disability Assessment Schedule; WHODAS–1 item [17]). The distress, functioning and recovery, and days out of role measures were added following lived experience consultation—these measures were added following initial registration of the trial, but prior to recruitment opening.

At 6 weeks, participants in the intervention group completed additional measures in relation to the app: appropriateness of intervention (Implementation Appropriateness Measure; IAM [18]), the Digital Working Alliance Inventory (DWAI [19]) and a bespoke questionnaire seeking feedback on the BrighterSide app. Participants in the intervention group were also invited on completion of the 12-week measures to participate in a semistructured interview to provide detailed feedback on their experience using BrighterSide.

Safety Monitoring

Data on adverse events and serious adverse events related to suicidal ideation and suicide attempts, respectively, were routinely collected at all time points though the self-report outcome measures already described. Specifically, severe suicidal ideation was indicated by a total score of 21 or higher on the SIDAS, and a recent suicide attempt was marked by either a self-report of 1 or more suicide attempts in the past 6 weeks (in the self-harm behavior questionnaire), or as indicated by a score of 10 ("I have made an attempt") on question 3 of the SIDAS.

If a participant indicated they had recently attempted suicide (as described above), an email was automatically sent to them with support contact details and to arrange for a follow-up phone call by the research team. If a response was not received within 1 business day, a second email was sent to the participant.

Follow-up phone calls ensured that the participant was safe and offered a referral to Lifeline, a telephone crisis support service who would be able to call the participant and offer specialized support. In the case of the intervention group, participants were also routinely asked during questionnaires if their suicidal ideation and/or suicide attempt was related to use of the BrighterSide app. If they indicated yes, the follow-up call would be identical to that described above, with an addition to seek clarity on if and how use of the app contributed to their suicidal thoughts or behaviours.

Patient and Public Involvement

The BDI's lived experience advisory team collaborated on the content and design of the app. Furthermore, the lived experience advisors were consulted on the trial design and recommended including outcome measures related to functioning and recovery (measured with the FRS), psychological distress (measured with the DQ5), and days out of role (measured with WHODAS-1).

Sample Size

The initial recruitment target for the trial was 394 participants, with 197 participants per arm. This allowed for detection of a small to medium effect size (d=0.3) in the primary outcome (severity of suicidal ideation) between the intervention and control arms with 80% power (α=.05), allowing for 40% attrition at postintervention follow-up. This was informed by 3 previous trials that incorporated the underlying intervention content in a web-based program, reporting a pooled effect size of 0.31 and average attrition rate of 35% [8-10]. After observing a higher than estimated attrition rate at the 6-week follow-up (n=223; 56.6%), the recruitment target was raised to 546 (n=273 per arm) to maintain statistical power.

Statistical Methods

Demographic and clinical characteristics were compared between study arms using 2-sided independent sample t tests, χ^2 tests, or the Fisher exact test. Mixed model repeated measures (MMRM) analyses with maximum likelihood estimation and an unstructured covariance matrix were used to evaluate the efficacy of the BrighterSide app relative to the control condition. The primary outcome was severity of suicidal ideation as assessed by the SIDAS over time (baseline to 6 weeks; baseline to 12 weeks). The mixed model approach incorporates all available data, including participants with missing follow-up data points, under the missing-at-random assumption that is robust to data that are missing contingent on observed variables. Analyses were performed under the intention-to-treat principle by a statistician who was blinded to group allocation. Analyses of secondary outcomes used the same methods for continuous outcomes. The frequency of adverse events, including severe suicidal ideation, recent self-harm, and recent suicide attempts, were compared between groups using a χ^2 test of independence.

Descriptive statistics were used to evaluate use (indicated by the number of modules completed, recorded using app analytics) of the BrighterSide app by participants in the intervention condition. Interview data collected from a subset of participants in the intervention group were analyzed thematically by one of the authors [20]. An inductive approach, independent of a theoretical confirmative method, was used to identify and group themes. Themes were refined to determine the final coding framework.

Post Hoc Analysis

A post hoc analysis using logistic regression was performed to assess whether follow-up attrition rates at 6 and 12 weeks after baseline could be predicted by any factors measured at baseline.

Results

Recruitment and Baseline Characteristics

A total of 795 participants were assessed for eligibility, of whom 550 were randomized (see Figure S1 in Multimedia Appendix 1). A total of 275 participants were randomized to each group. All participants were analyzed under the intention-to-treat principle, except 1 (in the intervention group) who withdrew and requested that their data not be retained. Baseline characteristics for participants are presented in Table 1; the groups did not differ across any measure.

Table . Baseline characteristics and clinical outcomes for each group. Significance values refer to comparisons of the 2 groups using a 2-sided independent-sample *t* test, except where footnoted.

		Total (n=549)	BrighterSide (n=274)	Control (n=275)	*P* value
Characteristics					
	Female, n (%)	399 (72.7)	203 (74.1)	196 (71.3)	.84[a]
	Age (years), mean (SD)	39.1 (13.5)	39.1 (13.5)	39.1 (13.5)	.98
Actual Help-Seeking Questionnaire					
	Mental health help sought in past 3 months, n (%)	453 (82.5)	226 (82.5)	227 (82.6)	.54[a]
Suicidal Ideation Attributes Scale					
	Mean score (SD)	25.8 (10.1)	25.7 (10.3)	25.8 (10)	.89
	Score ≥21, n (%)	366 (66.7)	178 (64.9)	188 (68.4)	.26[a]
Self-harm (%)					
	Have you ever harmed yourself on purpose? (score >0)	416 (75.8)	208 (75.9)	208 (75.6)	.51[a]
	In the last six weeks have you harmed yourself on purpose? (score >0)	202 (36.8)	101 (36.9)	101 (36.7)	.54[a]
	Have you ever attempted to take your own life? (score >0)	291 (53)	152 (55.5)	139 (50.6)	.14[a]
	In the last six weeks have you attempted to take your own life? (score >0)	38 (6.9)	19 (6.9)	19 (6.9)	.45[a]
Functioning and Recovery Scale score, mean (SD)		13.3 (2.4)	13.3 (2.4)	13.4 (2.3)	.55
Distress Questionnaire-5 score, mean (SD)		19.2 (2.8)	19.3 (2.7)	19.2 (2.9)	.67
WHO Disability Assessment Schedule (1 item for days out of role) score, mean (SD)		10.1 (8.9)	10.4 (9.2)	9.9 (8.7)	.45
Abbreviated Coping Orientation to Problems Experienced Inventory score, mean (SD)					
	Problem-focused score	18.4 (4.9)	18.3 (5.1)	18.4 (4.8)	.79
	Emotion-focused score	27.6 (4.9)	27.7 (5.2)	27.5 (4.7)	.53
	Avoidant-focused score	18.2 (3.6)	18.3 (3.6)	18.1 (3.7)	.69

[a]*P* value refers to the X^2 or Fisher exact test.

Primary and Secondary Outcomes

The mixed effects models for the primary and secondary outcomes are shown in Table 2. The main effect of time was significant for suicidal ideation, functioning and recovery, days out of role, psychological distress, problem-focused coping, and avoidant coping. There were no main effects for condition for any of the measures, and the time by group interaction was only significant for psychological distress. Residuals for SIDAS scores were nonnormal; however, results were identical under a negative binomial MMRM. Therefore, while the primary outcome, suicidal ideation, demonstrated a significant reduction over time, this did not significantly differ between groups (Figure 1). The Cohen *d* effect size for suicidal ideation between the intervention and control groups from baseline to 6 weeks was *d*=–0.03, and *d*=–0.15 from baseline to 12 weeks. Negative effect sizes favor the control group. See Table S2 in Multimedia Appendix 1 for summary of Cohen *d* effect sizes.

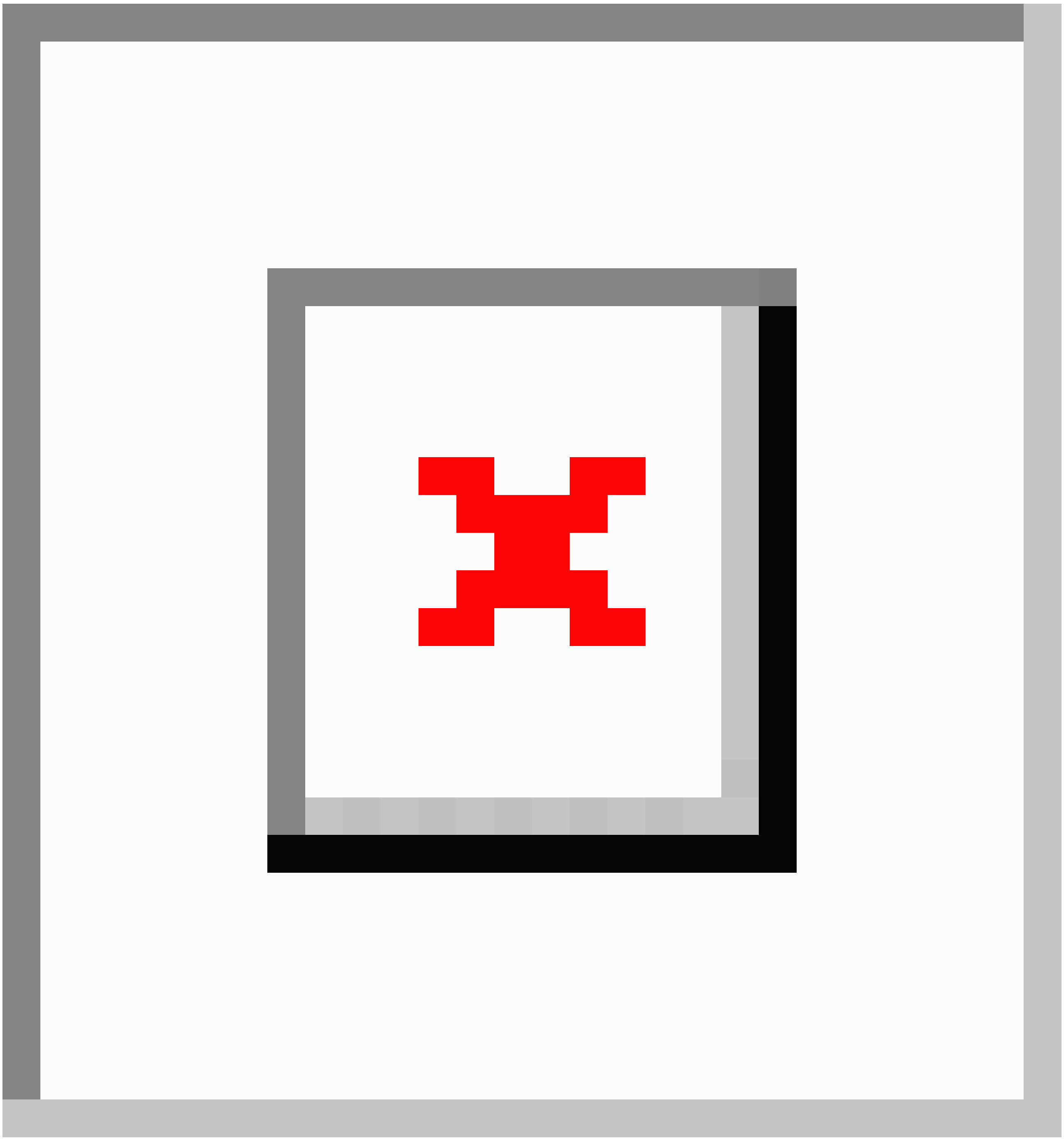
Table . Omnibus mixed model repeated measures ANOVA time (baseline; 6 weeks; 12 weeks) × group (BrighterSide; control).

		F test (*df*)	*P* value
Suicidal Ideation Attributes Scale			
	Time	124.11 (2,223.1)	<.001[a]
	Group	0.01 (1,404.3)	.92
	Time × group	0.00 (2,223.1)	.99
Functioning and Recovery Scale			
	Time	25.89 (2,242.7)	<.001[a]
	Group	0.02 (1,384.1)	.88
	Time × group	1.18 (2,247.7)	.31
WHO Disability Assessment Schedule (1 item for days out of role)			
	Time	13.69 (2,198.6)	<.001[a]
	Group	0.04 (1,375.9)	.83
	Time × group	1.05 (2,198.6)	.35
Distress Questionnaire-5			
	Time	47.04 (2,234.0)	<.001[a]
	Group	1.74 (1,396.3)	.18
	Time × group	3.62 (2,234.0)	.03[a]
Self-harm			
	Time	37.29 (2,214.8)	<.001[a]
	Group	0.00 (1,411.6)	.97
	Time × group	0.35 (2,214.8)	.70
Brief-COPE [b] (problem)			
	Time	3.38 (2,221.4)	.01[a]
	Group	0.71 (1,380.3)	.40
	Time × group	1.03 (2,221.4)	.35
Brief-COPE (emotional)			
	Time	0.04 (2,225.7)	.96
	Group	2.11 (1,370.2)	.15
	Time × group	1.01 (2,225.7)	.36
Brief-COPE (avoidant)			
	Time	18.35 (2,225.6)	<.001[a]
	Group	1.24 (1,381.4)	.26
	Time × group	2.45 (2,225.6)	.08

[a]Significant at α=.05.

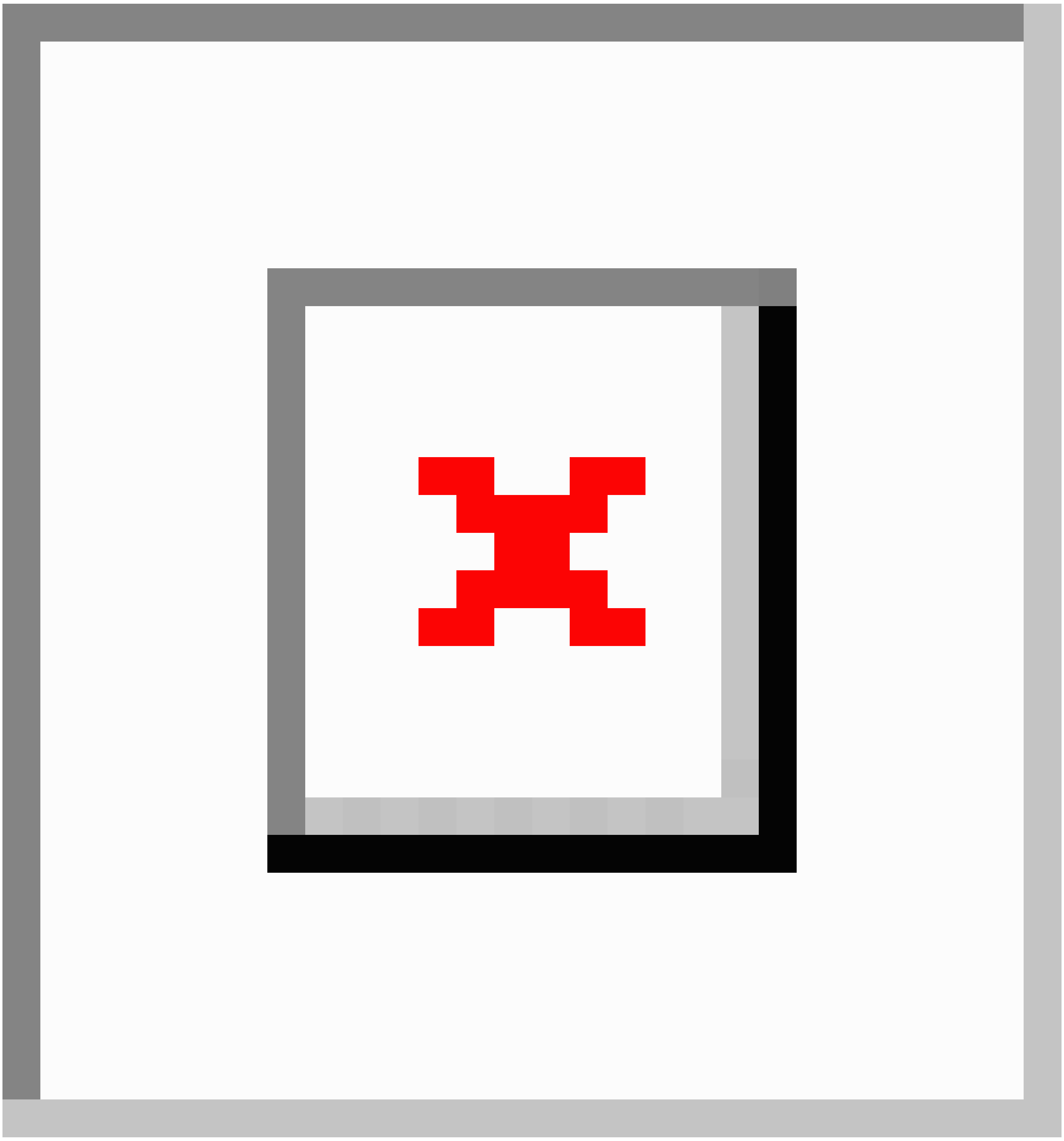
[b]Brief-COPE: abbreviated Coping Orientation to Problems Experienced Inventory.

Figure 1. Suicidal ideation (Suicidal Ideation Attributes Scale; SIDAS) scores by group at baseline, 6 weeks, and 12 weeks. Error bars represent 1 SE.



Follow-up analysis of the interaction between time and group in the DQ5 demonstrated a significant difference between the intervention and control groups at 6 weeks ($t_{244.60}$=2.68, 95% CI 0.25-1.67; *P*=.01), where the BrighterSide group reported lower scores (mean 17.36, SE 0.27) than the control group (mean 18.22, SE 0.25; Cohen *d*=0.26). However, this difference was no longer significant at 12 weeks ($t_{189.61}$=1.05, 95% CI –0.41 to 1.35; *P*=.29; Cohen *d*=−0.01) (Figure 2).

Figure 2. Distress (Distress Questionnaire-5; DQ5) scores by group at baseline, 6 weeks, and 12 weeks. Error bars represent 1 SE.



Adverse Events

Comparison of the frequency of severe suicidal ideation, self-harm, and suicide attempts between the control and intervention groups yielded no significant differences at either 6 weeks or 12 weeks (Table S3 in Multimedia Appendix 1). Given the target population were people experiencing suicidal ideation at baseline, the frequency of adverse events is not remarkable at either follow-up time point.

Attrition

We used a post hoc logistic regression analysis modeling predictors of attrition at 6 weeks and 12 weeks. SIDAS and FRS were chosen as variables to investigate for any potential moderating effects, rather than all surveys, as doing so may have violated the assumption of no multicollinearity (full results available in Table S4 in Multimedia Appendix 1). The only significant predictors of attrition were group at 12 weeks and the SIDAS × group interaction at 12 weeks. The interaction reflects a greater proportion of participants with lower levels of suicidal ideation (SIDAS scores of 0-20) dropping out in the intervention group (77/96, 80%) compared to the control group (48/87, 59%).

User Engagement Outcomes

Analytic information regarding the frequency of use of the app was obtained. Overall, 188 intervention participants (n=275, 68.4% of the baseline intervention group) enrolled into the BrighterSide app, which is the initial onboarding task after downloading and opening the app. Table 3 presents descriptive statistics on module engagement across the course of the trial.

Table . Engagement with activities by module in the BrighterSide app (n=275).

	Module									
	Understand Your Thoughts		Prevent a Crisis		Navigate Your Emotions		Navigate Your Thoughts		Plan for the Future	
	No.[a]	%[b]	No.	%	No.	%	No.	%	No.	%
Engaged with ≥1 activity in module	100	36.4	58	21.1	57	20.7	40	14.5	47	17.1
Engaged with ≥50%of activities in module	95	34.5	57	20.7	33	12	13	4.7	27	9.8
Completed all activities in module	19	6.9	22	8	8	2.9	6	2.2	4	1.5

[a]No.: number of users.

[b]%: percentage of baseline sample randomised to the intervention group.

In addition to module engagement, we also reviewed engagement with distraction and calming activities. We found that 69 users (25.1% of the intervention group) engaged with at least 1 calming activity and 43 users (15.6% of the intervention group) engaged with at least 1 distraction activity. Given the low percentage of engagement with these activities, they were not considered further.

Use and Acceptability Outcomes

Tables S5 and S6 in Multimedia Appendix 1 present the results of the appropriateness (IAM) and therapeutic alliance (DWAI) measures collected at the 6-week time point. The average of the appropriateness scales was 3.63 (SD 0.89) on a range of 1 to 5, indicating that the participants' response to the app was positively skewed. This result is consistent with another mental health–oriented app using this same measure, which returned an average of 3.6 [21]; these authors regarded this as an indication of appropriateness. Each question on the DWAI averaged a response between 2.11 and 3.28 across all participants, on a scale of 1 (not at all) to 4 (completely), suggesting that typically, participants indicated they somewhat to mostly agreed with each statement.

Participants were also given a study-specific acceptability survey, where only 44 of 86 responders (51%) agreed the app met their needs, but 73 of those responders (85%) agreed it was easy to use (agreement was determined by a score of 5 or higher on a 7-point scale; see Table S7 in Multimedia Appendix 1 for full results). This survey also included questions regarding the use of the safety plan feature. This showed that 31 of the 86 (36%) responders to this question had filled out the safety plan, and 100% of these did so alone; 81% (25/31) did not share the safety plan with someone else, and the remaining 19% (6/31) did so with one other person. Finally, 58% (18) of those who completed their safety plan looked back on it. For those who did not fill out the safety plan, the most frequent responses when asked why were that they did not feel they had the time; they found it too confronting or overwhelming; they had completed a safety plan elsewhere; or they simply did not know how to navigate to it in the app.

Semistructured Interviews

Of the 69 intervention participants who completed the 12-week follow-up, 6 agreed to be interviewed about their experiences of using the BrighterSide app. Themes and subthemes that emerged throughout the interviews are described in Table S8 in Multimedia Appendix 1. The most notable patterns were that participants found the distraction and mindfulness activities particularly useful, and that the check-in function alleviated the burdensomeness often associated with reaching out to support networks. At the same time, some participants said that while not all of the information in the modules was helpful to them, it may be useful for people who have little or no experience with professional mental health care. Participants felt that the safety plan was useful as a reference tool for when they may be entering a crisis, and that the ability to share this plan with others was helpful since they often are unable to recognize their own warning signs.

Discussion

Principal Findings

The primary objective of this study was to determine if use of the BrigherSide app significantly decreased suicidal ideation at 6 weeks and 12 weeks after baseline compared to a waitlist control group. While there was a significant overall decrease in suicidal ideation from baseline to 12 weeks, there was no difference between the 2 groups. However, the intervention group did report significantly lower psychological distress at 6 weeks compared to the control group, although this difference was no longer significant at 12 weeks. There was no difference in rates of adverse events (severe suicidal ideation, recent self-harm, or recent suicide attempts), nor were there any significant moderators of demographic variables (age and gender) on attrition.

There may be several possible reasons for the null effect on suicidal ideation. First, there was little engagement with the app. While 68% downloaded and completed onboarding of the app, only 36% of the intervention group engaged with at least one activity in the first module (Understand Your Thoughts; see Table 3). This reflects that participants in the intervention group were not exposed to the full anticipated benefit of BrighterSide. Indeed, only 51% of participants found the app had met their needs. This may be an artifact of the nonprescriptive approach to the intervention. Participants were able to access content in any order they wished, or, if they preferred, use a guided option. If participants were initially exposed to the first module and did not find it helpful, they may have been less inclined to engage with the remaining modules, which may have been more relevant to them. In a similar vein, it could be argued that, given the identified relationship between thwarted belongingness and suicidal ideation [22,23], placing the onus on the participant to choose their method of interaction without the possibility of social connection (ie, no contact with another person) may have been counterintuitive to the aim of reducing suicidal ideation. This may reflect the finding that only half of participants felt the app met their needs, and an investigation into the usefulness of an app with or without a social connection aspect, and the impact of this on app engagement, should be investigated in future research. In any case, the lack of engagement with the app would anticipate a null effect.

Second, a high proportion of participants (453 of 550 participants, over 80%) had recently sought professional help for their mental health, and the strategies provided in the app may therefore have been already known to participants. Third, the trial recruitment period occurred during a period of public health protections in the COVID-19 pandemic. While rates of suicidal ideation were high but stable during the period of this study [24], broader public mental health support during this period may have had a confounding effect on the trial. Finally, despite the extensive co-design process to engage people with lived experience and clinicians, the final intervention may not have achieved an optimal balance between therapeutic content and user engagement. Given the null effect of the BrighterSide app on suicidal ideation, the remainder of this discussion will evaluate the differences between BrighterSide and other digital interventions that have also aimed to reduce suicidal ideation.

Psychological Distress

Participants in the BrighterSide group reported significantly lower distress at 6 weeks than did those in the control group. While this difference was not maintained at 12 weeks, it might indicate that use of the app provided useful tools to navigate psychological distress in the short term. This is noteworthy given that high psychological distress is evidenced to be related to high reports of suicidal ideation and suicide attempts [25]. Ameliorating psychological distress may therefore act as a protective factor, although this is not directly captured by the data. Given that this study indicates the two are not comorbid, further investigation into the relationship between psychological distress and suicidal ideation, with greater power, is warranted.

Comparison of Therapeutic Models

The previous finding that the digital intervention Living with Deadly Thoughts did not significantly reduce suicidal ideation between groups should be discussed in the context of this study [8]. While BrighterSide is modeled on this previous intervention, one of the key differences was the co-design process involved in developing the content for BrighterSide. Incorporating consumers in the design process of these interventions was one method proposed by Torous et al [11] to enhance engagement. Despite this, BrighterSide did not see substantial app engagement, and was rated by almost half of participants to have not met their needs. However, it is known that suicide prevention apps are useful in reducing suicide ideation—for example, the LifeBuoy app saw a significant difference in suicide ideation for the intervention group compared to an active control group [26]. We therefore consider the differences in the therapeutic content involved in these 2 models.

BrighterSide contains modules based on a mix of cognitive behavioral therapy, dialectical behavior therapy, acceptance and commitment therapy, and positive psychology. The amalgamation of different therapeutic elements for a brief intervention such as BrighterSide may have lacked enough adherence to a particular model to see any benefit. Instead, a greater effect on suicidal ideation might be achieved with modules that adhere to one therapeutic model, such as dialectical behavior therapy, since it has been shown to have a great effect on reducing suicidal behavior (see Ougrin et al [27] for a review). The LifeBuoy study implemented an intervention that followed a dialectical behavior therapy model and demonstrated a significant reduction in suicidal ideation when compared to a control group. They also allowed flexibility with module use, but their implementation was prescriptive, so that one module had to be completed in order to unlock the next. Additionally, if there were greater coherence within the modules, perhaps participants would be more inclined to engage meaningfully. Given these results, and the null findings for BrighterSide, it may be beneficial in future research to adhere to a single therapeutic model, such as dialectical behavior therapy, to both enhance delivery of skills and to enhance engagement among participants.

Limitations

The most considerable limitation in this study is the low engagement with the app itself, which inhibits the capacity to adequately assess the primary and secondary outcomes. While maintaining engagement with app use is a recognized issue in digital mental health [11], this study did take into consideration some factors to enhance usability via the co-design method. Additionally, the way the app conveyed information may not have been conducive to participants actually implementing the learned knowledge and skills from the modules, particularly given the brief nature of the app and the amalgamation of components from different therapeutic models. Future research should consider assessing the learnability of skills portrayed in the modules, to determine whether apps are able to implement behavioral change.

While the interviews with participants were mostly positive, interviews were conducted with participants who self-selected to participate in an interview after completing the 12-week follow-up. Therefore, the small number who self-selected were more likely to have actively engaged in the app, generally had greater motivation for improving mental health research, and may have already felt more positively about the app. Regardless, the interview outcomes were consistent with the results from the surveys on acceptability, functionality, and perceptions of BrighterSide.

Conclusion

This study aimed to investigate the ability of an app, BrighterSide, to reduce suicide ideation. While there were no between-condition effects for suicidal ideation, the severity of psychological distress was significantly reduced in the intervention condition after having access to the app for 6 weeks, relative to the control group. Further work may be required to optimally incorporate effective therapeutic content with engaging user design.

Acknowledgments

This research was supported by the Australian Government Department of Health–funded National Suicide Prevention Research Fund, managed by Suicide Prevention Australia. The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. We acknowledge that BrighterSide is a registered trademark, even though the ® symbol has been omitted from this paper.

Authors' Contributions

NJ, AT, NC, and ML contributed to project management; NJ, PB, and ML contributed to data analysis; NJ, PB, and ML contributed to data interpretation; NJ contributed to draft manuscript preparation and final manuscript preparation; MT, PB, QW, JRB, AT, SH, KH, JR, NC, HC, and ML contributed to study conception and design; MT, SH, KH, JR, NC, HC, and ML contributed to study resources; MT, PB, QW, JRB, AT, HC, and ML contributed to draft revisions; and AT, SH, and ML contributed to data curation. All authors reviewed the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Complete set of appendices.
[DOCX File, 42 KB - mental_v11i1e55528_app1.docx]

Multimedia Appendix 2
BrighterSide mobile app screenshots.
[DOCX File, 1089 KB - mental_v11i1e55528_app2.docx]

Checklist 1
CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Application and online Tele Health) checklist.
[PDF File, 7360 KB - mental_v11i1e55528_app3.pdf]

References

1. National study of mental health and wellbeing. Australian Bureau of Statistics. 2023. URL: https://www.abs.gov.au/statistics/health/mental-health/national-study-mental-health-and-wellbeing/latest-release [accessed 2024-03-04]
2. Nock MK, Borges G, Bromet EJ, et al. Cross-national prevalence and risk factors for suicidal Ideation, plans and attempts. Br J Psychiatry 2008 Feb;192(2):98-105. [doi: 10.1192/bjp.bp.107.040113] [Medline: 18245022]
3. 2022 annual report. Lifeline Australia. 2023. URL: https://www.lifeline.org.au/media/hikdmw5w/lifeline-annual-report-2022.pdf [accessed 2024-03-04]
4. Healthcare quarterly 2022. Bureau of Health Information. 2023. URL: https://www.bhi.nsw.gov.au/__data/assets/pdf_file/0005/825017/BHI_HQ50_JUL-SEP_2022_REPORT.pdf [accessed 2024-04-03]
5. Etzelmueller A, Vis C, Karyotaki E, et al. Effects of internet-based cognitive behavioral therapy in routine care for adults in treatment for depression and anxiety: systematic review and meta-analysis. J Med Internet Res 2020 Aug 31;22(8):e18100. [doi: 10.2196/18100] [Medline: 32865497]

6. Rohrbach PJ, Dingemans AE, Evers C, et al. Cost-effectiveness of internet interventions compared with treatment as usual for people with mental disorders: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2023 Jan 5;25:e38204. [doi: 10.2196/38204] [Medline: 36602854]
7. Torok M, Han J, Baker S, et al. Suicide prevention using self-guided digital interventions: a systematic review and meta-analysis of randomised controlled trials. Lancet Digit Health 2020 Jan;2(1):e25-e36. [doi: 10.1016/S2589-7500(19)30199-2] [Medline: 33328037]
8. van Spijker BA, Werner-Seidler A, Batterham PJ, et al. Effectiveness of a web-based self-help program for suicidal thinking in an Australian community sample: randomized controlled trial. J Med Internet Res 2018 Feb;20(2):e15. [doi: 10.2196/jmir.8595] [Medline: 29444769]
9. De Jaegere E, van Landschoot R, van Heeringen K, et al. The online treatment of suicidal ideation: a randomised controlled trial of an unguided web-based intervention. Behav Res Ther 2019 Aug;119:103406. [doi: 10.1016/j.brat.2019.05.003] [Medline: 31176889]
10. van Spijker BAJ, van Straten A, Kerkhof A. Effectiveness of online self-help for suicidal thoughts: results of a randomised controlled trial. PLoS One 2014;9(2):e90118. [doi: 10.1371/journal.pone.0090118] [Medline: 24587233]
11. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018 Aug;21(3):116-119. [doi: 10.1136/eb-2018-102891] [Medline: 29871870]
12. van Spijker BAJ, Batterham PJ, Calear AL, et al. The Suicidal Ideation Attributes Scale (SIDAS): community-based validation study of a new scale for the measurement of suicidal ideation. Suicide Life Threat Behav 2014 Aug;44(4):408-419. [doi: 10.1111/sltb.12084] [Medline: 24612048]
13. Carver CS. You want to measure coping but your protocol is too long: consider the brief cope. Int J Behav Med 1997;4(1):92-100. [doi: 10.1207/s15327558ijbm0401_6] [Medline: 16250744]
14. Batterham PJ, Poyser C, Gulliver A, Banfield M, Calear AL. Development and psychometric properties of the Functioning And Recovery Scale: a new measure to assess psychosocial functioning after a suicide attempt. Suicide Life Threat Behav 2020 Dec;50(6):1105-1114. [doi: 10.1111/sltb.12656] [Medline: 32744356]
15. Batterham PJ, Sunderland M, Carragher N, Calear AL, Mackinnon AJ, Slade T. The Distress Questionnaire-5: population screener for psychological distress was more accurate than the K6/K10. J Clin Epidemiol 2016 Mar;71:35-42. [doi: 10.1016/j.jclinepi.2015.10.005] [Medline: 26464194]
16. Rickwood DJ, Braithwaite VA. Social-psychological factors affecting help-seeking for emotional problems. Soc Sci Med 1994 Aug;39(4):563-572. [doi: 10.1016/0277-9536(94)90099-x] [Medline: 7973856]
17. Von Korff M, Crane PK, Alonso J, et al. Modified WHODAS-II provides valid measure of global disability but filter items increased skewness. J Clin Epidemiol 2008 Nov;61(11):1132-1143. [doi: 10.1016/j.jclinepi.2007.12.009] [Medline: 18619808]
18. Weiner BJ, Lewis CC, Stanick C, et al. Psychometric assessment of three newly developed implementation outcome measures. Implement Sci 2017 Aug;12(1):108. [doi: 10.1186/s13012-017-0635-3] [Medline: 28851459]
19. Henson P, Wisniewski H, Hollis C, Keshavan M, Torous J. Digital mental health apps and the therapeutic alliance: initial review. BJPsych Open 2019 Jan;5(1):e15. [doi: 10.1192/bjo.2018.86] [Medline: 30762511]
20. Braun V, Clarke V. Using thematic analysis in psychology. Qualita Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
21. Comtois KA, Mata-Greve F, Johnson M, Pullmann MD, Mosser B, Arean P. Effectiveness of mental health apps for distress during COVID-19 in US unemployed and essential workers: remote pragmatic randomized clinical trial. JMIR Mhealth Uhealth 2022 Nov 7;10(11):e41689. [doi: 10.2196/41689] [Medline: 36191176]
22. Chu C, Buchman-Schmitt JM, Stanley IH, et al. The interpersonal theory of suicide: a systematic review and meta-analysis of a decade of cross-national research. Psychol Bull 2017 Dec;143(12):1313-1345. [doi: 10.1037/bul0000123] [Medline: 29072480]
23. Joiner T. Why People Die by Suicide: Harvard University Press; 2005.
24. Batterham PJ, Calear AL, Shou Y, et al. Effects of the COVID-19 pandemic on suicidal ideation in a representative Australian population sample-longitudinal cohort study. J Affect Disord 2022 Mar;300:385-391. [doi: 10.1016/j.jad.2022.01.022] [Medline: 34995704]
25. Eskin M, Sun JM, Abuidhail J, et al. Suicidal behavior and psychological distress in university students: a 12-nation study. Arch Suicide Res 2016 Jul;20(3):369-388. [doi: 10.1080/13811118.2015.1054055] [Medline: 26954847]
26. Torok M, Han J, McGillivray L, et al. The effect of a therapeutic smartphone application on suicidal ideation in young adults: findings from a randomized controlled trial in Australia. PLoS Med 2022 May;19(5):e1003978. [doi: 10.1371/journal.pmed.1003978] [Medline: 35639672]
27. Ougrin D, Tranah T, Stahl D, Moran P, Asarnow JR. Therapeutic interventions for suicide attempts and self-harm in adolescents: systematic review and meta-analysis. J Am Acad Child Adolesc Psychiatry 2015 Feb;54(2):97-107. [doi: 10.1016/j.jaac.2014.10.009] [Medline: 25617250]

Abbreviations

ASHQ: Actual Health-Seeking Questionnaire
BDI: Black Dog Institute
Brief-COPE: Coping Orientation to Problems Experienced Inventory
CONSORT: Consolidated Standards of Reporting Trials
DQ5: Distress Questionnaire-5
DWAI: Digital Working Alliance Inventory
FRS: Functioning and Recovery Scale
IAM: Implementation Appropriateness Measure
MMRM: mixed model repeated measures
RCT: randomized controlled trial
SIDAS: Suicidal Ideation Attributes Scale
WHODAS: World Health Organization Disability Assessment Schedule

Edited by J Torous; submitted 23.01.24; this is a non–peer-reviewed article;accepted 31.01.24; published 18.03.24.

Please cite as:
Josifovski N, Torok M, Batterham P, Wong Q, Beames JR, Theobald A, Holland S, Huckvale K, Riley J, Cockayne N, Christensen H, Larsen M
Efficacy of BrighterSide, a Self-Guided App for Suicidal Ideation: Randomized Controlled Trial
JMIR Ment Health 2024;11:e55528
URL: https://mental.jmir.org/2024/1/e55528
doi:10.2196/55528

Original Paper

Veteran Experiences With an mHealth App to Support Measurement-Based Mental Health Care: Results From a Mixed Methods Evaluation

Robin T Higashi[1,2], PhD; Bella Etingen[1,2,3], PhD; Eric Richardson[2,4,5], PhD; Jennifer Palmer[2,4,6], MS, PhD; Mark S Zocchi[2,7], PhD; Felicia R Bixler[2,8], MS; Bridget Smith[2,8,9], PhD; Nicholas McMahon[2,7], MPH; Kathleen L Frisbee[2,10], MPH, PhD; John C Fortney[11,12], PhD; Carolyn Turvey[13,14,15], PhD; Jennifer Evans[16], PhD; Timothy P Hogan[1,2,7], PhD

[1]Peter O'Donnell Jr School of Public Health, University of Texas Southwestern Medical Center, Dallas, TX, United States
[2]eHealth Partnered Evaluation Initiative, Veterans Affairs Bedford Healthcare System, Bedford, MA, United States
[3]Research and Development Service, Dallas Veterans Affairs Medical Center, Dallas, TX, United States
[4]Center for Healthcare Organization and Implementation Research (CHOIR), Veterans Affairs Boston Healthcare System, Boston, MA, United States
[5]Department of Psychiatry, Harvard Medical School, Boston, MA, United States
[6]Section of General Internal Medicine, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, United States
[7]Center for Healthcare Organization and Implementation Research (CHOIR), Veterans Affairs Bedford Healthcare System, Bedford, MA, United States
[8]Center of Innovation for Complex Chronic Healthcare (CINCCH), Hines Veterans Affairs Hospital, Hines, IL, United States
[9]Northwestern University Feinberg School of Medicine, Chicago, IL, United States
[10]Office of Connected Care, Veterans Health Administration, US Department of Veterans Affairs, Washington, DC, United States
[11]Center of Innovation for Veteran-Centered and Value-Driven Care, Veterans Affairs Puget Sound Health Care System, Seattle, WA, United States
[12]Division of Population Health, Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, United States
[13]Center for Access & Delivery Research and Evaluation, Iowa City Veterans Affairs Health Care System, Iowa City, IA, United States
[14]Office of Rural Health, Veterans Rural Health Resource Center – Iowa City, Iowa City Veterans Affairs Health Care System, Iowa City, IA, United States
[15]Department of Psychiatry, University of Iowa, Iowa City, IA, United States
[16]Office of Mental Health and Suicide Prevention, US Department of Veterans Affairs, Washington, DC, United States

Corresponding Author:
Robin T Higashi, PhD
Peter O'Donnell Jr School of Public Health
University of Texas Southwestern Medical Center
5323 Harry Hines Blvd
Dallas, TX, 75390-9066
United States
Phone: 1 (214) 648 3645
Email: Robin.Higashi@UTSouthwestern.edu

Abstract

Background: Mental health conditions are highly prevalent among US veterans. The Veterans Health Administration (VHA) is committed to enhancing mental health care through the integration of measurement-based care (MBC) practices, guided by its Collect-Share-Act model. Incorporating the use of remote mobile apps may further support the implementation of MBC for mental health care.

Objective: This study aims to evaluate veteran experiences with Mental Health Checkup (MHC), a VHA mobile app to support remote MBC for mental health.

Methods: Our mixed methods sequential explanatory evaluation encompassed mailed surveys with veterans who used MHC and follow-up semistructured interviews with a subset of survey respondents. We analyzed survey data using descriptive statistics. We then compared responses between veterans who indicated having used MHC for ≥3 versus <3 months using χ^2 tests. We analyzed interview data using thematic analysis.

Results: We received 533 surveys (533/2631, for a 20% response rate) and completed 20 interviews. Findings from these data supported one another and highlighted 4 key themes. (1) The MHC app had positive impacts on care processes for veterans: a majority of MHC users overall, and a greater proportion who had used MHC for ≥3 months (versus <3 months), agreed or strongly agreed that using MHC helped them be more engaged in their health and health care (169/262, 65%), make decisions about their treatment (157/262, 60%), and set goals related to their health and health care (156/262, 60%). Similarly, interviewees described that visualizing progress through graphs of their assessment data over time motivated them to continue therapy and increased self-awareness. (2) A majority of respondents overall, and a greater proportion who had used MHC for ≥3 months (versus <3 months), agreed/strongly agreed that using MHC enhanced their communication (112/164, 68% versus 51/98, 52%; P=.009) and rapport (95/164, 58% versus 42/98, 43%; P=.02) with their VHA providers. Likewise, interviewees described how MHC helped focus therapy time and facilitated trust. (3) However, veterans also endorsed some challenges using MHC. Among respondents overall, these included difficulty understanding graphs of their assessment data (102/245, 42%), not receiving enough training on the app (73/259, 28%), and not being able to change responses to assessment questions (72/256, 28%). (4) Interviewees offered suggestions for improving the app (eg, facilitating ease of log-in, offering additional reminder features) and for increasing adoption (eg, marketing the app and its potential advantages for veterans receiving mental health care).

Conclusions: Although experiences with the MHC app varied, veterans were positive overall about its use. Veterans described associations between the use of MHC and engagement in their own care, self-management, and interactions with their VHA mental health providers. Findings support the potential of MHC as a technology capable of supporting the VHA's Collect-Share-Act model of MBC.

(JMIR Ment Health 2024;11:e54007) doi:10.2196/54007

KEYWORDS

measurement-based care; mobile health app; mental health; veteran; mHealth; support; mixed-methods evaluation; digital health

Introduction

Veterans are impacted by mental health conditions at disproportionately higher rates in comparison to nonveteran US adults [1]. Evidence further suggests that veterans are motivated to use technology to enhance their health [2,3], and respond positively to digital health interventions, including those that address their mental health needs [4-6]. Currently, there are at least a dozen mental and behavioral health apps on VAMobile, the Veterans Health Administration (VHA) mobile health app store, including those targeting support for veterans with depression, posttraumatic stress disorder (PTSD), anxiety, and insomnia. Studies evaluating app effectiveness have shown positive outcomes; for example, the VHA's PTSD Coach and Virtual Hope Box, two of their most popular mental health apps, showed significant improvements in symptom reduction and coping ability among app users versus control patients [7-9]. Thus, incorporating the use of mobile apps may further support the implementation of measurement-based care (MBC) for mental health care.

MBC is the systematic evaluation of patient symptoms using data collected from patients, including patient-reported outcome (PRO) measures, to inform health care and treatment decision-making [10,11]. In the context of mental health care, the integration of MBC practices has been demonstrated to enhance patient satisfaction and mental health outcomes [12-14]. The American Psychological Association endorses routine monitoring of symptoms, such as the use of PROs in MBC, as part of the delivery of evidence-based psychological practice [15].

The VHA, the largest integrated health care system in the United States [16], provides comprehensive mental health services across its nationwide system of care. The VHA launched its own MBC initiative in 2015 [17]. The VHA's MBC initiative espouses a Collect-Share-Act model, in which veteran PRO data are collected, shared between veterans and VHA care team members, and used to inform treatment and facilitate shared decision-making [18]. The VHA's Collect-Share-Act model was developed in light of research indicating that simply collecting assessment data from patients without facilitating patient-provider discussion of assessment results often fails to improve PROs such as symptom severity and feelings of well-being [19]. In addition to its Collect-Share-Act model, the VHA has also invested in various technology platforms intended to support remote measurement-based care (R-MBC).

R-MBC refers to the collection of PROs outside of the clinical encounter, typically with support from technologies including but not limited to secure messaging through a patient portal or mobile health apps. R-MBC is envisioned as a means to extend the reach of MBC by remotely tracking, monitoring, and asynchronously communicating symptoms to providers and facilitating patient self-management. Asynchronous R-MBC addresses a key barrier to MBC, disruption of clinic visit workflow due to the time it takes to administer and score assessments. Although recent systematic reviews have described mixed impacts of R-MBC [20], previous trials of interventions that have included R-MBC as part of effective mental health treatment have demonstrated improvements in clinical outcomes (eg, response to therapy and depression remission, adherence to medication, and satisfaction with depression treatment) in comparison to usual care [21,22].

Mental Health Checkup (MHC) is a mobile health app that was developed by the VHA Office of Connected Care to support R-MBC in VHA mental health services. Veterans can log on to MHC on the web using approved log-in credentials. Using MHC, veterans are assigned assessments by their VHA mental health providers that they may complete outside of a clinical encounter on a smartphone, tablet, or personal computer at intervals specified and assigned by providers. When a provider assigns assessments to a veteran in MHC, they receive an auto-generated email with a link that they can follow to complete the assessments, and reminders from the app when they have assigned assessments that they need to but have not yet completed. The app also includes a data visualization feature, wherein veterans can view graphs of their assessment results over time to monitor symptoms and progress. Additional app features allow mental health providers to receive alerts on veteran assessment scores, as well as notifications when veterans have completed assessments in between their appointments.

Despite the potential of technologies like the MHC app to facilitate R-MBC, few studies to date have described patient experiences using them. We conducted a mixed methods evaluation to assess veteran perspectives on and experiences with using the VHA MHC app, including their perspectives on its usability and applicability to their mental health care, and suggestions for enhancing the app.

Methods

Design

Our evaluation used a sequential-explanatory design comprising mailed surveys and semistructured telephone interviews with a national sample of veterans who were users of VHA's MHC app. "Users" were defined as veterans who had ≥2 unique log-ins to the MHC app to account for those veterans who may have logged in once to explore the app but not used it.

Recruitment and Data Collection

Surveys

We developed a survey with input from VHA operational leadership and mental health providers, including some involved in facilitating VHA's MBC initiative. The survey was designed to address the potential impacts of the MHC app on the facets of MBC that comprise VHA's Collect, Share, Act model of MBC, as well as other important constructs highlighted by stakeholders. Survey topics included frequency of use, preferences, barriers, and satisfaction with using the MHC app. A veteran member of our evaluation team reviewed survey question items for clarity and readability. We initiated the first survey mailing to 2690 veterans in June 2021. We initiated a follow-up mailing to nonresponders in July 2021 to facilitate response. Included in the survey mailings were a cover letter and a postage-paid reply envelope. We gave veterans the option to complete surveys by phone as preferred. Survey respondents were subsequently mailed a US $10 CVS (CVS Pharmacy, Inc) gift card in appreciation of their time.

Semistructured Interviews

Consistent with data integration through "building" in mixed methods [23], we recruited a subset of survey respondents to participate in follow-up semistructured telephone interviews intended to elaborate on survey responses. We purposefully stratified the interview sample to reflect variation in the following survey responses: app use and experiences with the app, including number of unique log-ins, app satisfaction, types of assessments completed in the app, age, and gender. We developed a semistructured interview guide with feedback from VHA operational leadership and VHA mental health providers and questions that broadly mapped to the domains of the well-established Non-Adoption, Abandonment, and Challenges to the Scaleup, Spread, and Sustainability (NASSS) of Health and Care Technologies Framework [24].

Interviews assessed veteran experiences with completing assessments in the app and using the app's other features, as well as the perceived impacts of app use on a veteran's care. Specifically, we asked veterans to describe (1) the condition for which they were using the MHC app, and the mental health care that they receive; (2) their previous experiences with MBC and how using MHC compares (eg, Can you please tell me about if and how you completed VA mental health care assessments before you began using MHC?); (3) their current experiences with R-MBC (eg, Are you still currently using the MHC app to complete assessments assigned to you by your VA mental health providers?); (4) their experience learning to use and using the MHC app (eg, What challenges have you encountered using MHC? What do you think is valuable about MHC?); (5) their perceived impact of MHC use on their mental health care, self-management, and outcomes (eg, How has completing assessments through MHC impacted your experiences with your VA mental health treatment?); and (6) their suggestions for improving MHC/R-MBC (eg, What changes could be made to make using the MHC app easier for veterans?).

Two evaluation team members with qualitative expertise (JP and FRB) conducted all interviews by telephone between August 2021 and September 2021, which coincided with the COVID-19 pandemic. Interviews lasted an average of 42 minutes (range 17-76 minutes). Each interview was audio recorded with the veteran's permission and then professionally transcribed verbatim. Veterans who participated in an interview were mailed a US $20 CVS gift card in appreciation of their time.

Data Analysis

Surveys

We characterized responses to the survey using descriptive statistics (frequencies and proportions) and compared survey responses between veterans who indicated on their survey that they used the app for ≥3 months with veterans who used the app for 3 months using the chi-square test statistic. Only veterans who indicated that they received and completed an assessment in MHC and also responded to the survey question regarding the duration of MHC use were included in the analyses. Stata (version 17; StataCorp LLC) was used for analysis of the survey data.

Semistructured Interviews

We analyzed interview data using thematic analysis facilitated by NVivo (version 12.0; QSR International). Two qualitatively trained investigators (RTH and JP) initially drafted a codebook using a deductively driven approach informed by topics in the interview guide. They jointly coded the first 25% of transcripts (5/20 interviews), iteratively refining the codes and definitions in an inductive manner. They then finalized the codebook and independently coded all the remaining transcripts. Any cross-analyst discrepancies in coding decisions were resolved through discussions in weekly meetings. We reviewed and synthesized findings in thematic reports, noting the frequency and emphasis of veteran comments, and selected exemplary quotes. These synthesized reports were then reviewed and discussed with the larger evaluation team for interpretation and integration with the survey findings.

We used a number of established methods [25] that strengthen the rigor and credibility of our findings, including using a sequential-explanatory design that compared quantitative and qualitative data, triangulating interpretations through discussions among a team of qualitatively trained doctoral-level analysts, and incorporating both confirmatory and contrasting cases of emergent themes.

Ethical Considerations

All evaluation procedures were reviewed by relevant VHA institutional review boards and designated as program evaluation for quality improvement purposes, exempting it from further oversight (VHA Handbook 1058.05).

Results

Overview

We received 533 surveys (533/2631 surveys, with 59 surveys excluded from the denominator because they were returned due to incorrect addresses, for a 20% response rate), of which 271 respondents indicated they completed an MHC assessment. As noted above, veterans who also responded to the question about their duration of MHC use (n=268) were included in our analyses.

Demographic and Veteran Characteristics

Veteran sociodemographic data are shown in Table 1. Survey respondents were predominantly male (167/259, 65%), White (183/266, 69%), and aged 55 years or younger (160/266, 60.2%). Veterans who had used MHC for ≥3 months did not differ in sociodemographic characteristics from veterans who used MHC for <3 months.

We also completed 20 semistructured interviews. Interviewed participants were mostly White (14/20, 70%), male (13/20, 65%), lived between 16 and 60 minutes from their nearest VHA facility (15/20, 75%), had completed a bachelor's degree or higher (13/20, 65%), and were diverse in age (6/20, 30% aged ≤45 years; 9/20, 45% aged between 46 and 65 years; and 5/20, 25% aged 66 years and older). More than half (12/19, 63%) received between 90% and 100% of their VHA care remotely in the prior year, with 16% (n=3) receiving 0%-10% and 16% (n=3) receiving 75%-89% of VHA care remotely.

Table 1. Sociodemographic characteristics of surveyed Mental Health Checkup (MHC) app users by duration of use (n=268).

	Total, n (%)	Duration of MHC use		*P* value
		Less than 3 months, n (%)	3 months or longer, n (%)	
Age (years; n=266)				.54
<25	1 (0.4)	0 (0)	1 (0.6)	
25-35	31 (11.7)	13 (13.1)	18 (10.8)	
36-45	61 (22.9)	28 (28.3)	33 (19.8)	
46-55	67 (25.2)	20 (20.2)	47 (28.1)	
56-65	63 (23.7)	21 (21.2)	42 (25.1)	
66-75	37 (13.9)	15 (15.2)	22 (13.2)	
76 or older	6 (2.3)	2 (2)	4 (2.4)	
Sex (n=259)				.28
Male	167 (64.5)	61 (61.6)	106 (66.3)	
Female	88 (34)	35 (35.4)	53 (33.1)	
Decline to answer	4 (1.5)	3 (3)	1 (0.6)	
Race (n=266)				
American Indian or Alaskan Native	11 (4.1)	3 (3)	8 (4.8)	.49
Asian	11 (4.1)	3 (3)	8 (4.8)	.49
Black or African American	48 (18)	19 (19.2)	29 (17.4)	.71
Native Hawaiian or other Pacific Islander	6 (2.3)	2 (2)	4 (2.4)	.84
White	183 (68.8)	65 (65.7)	118 (70.7)	.40
Other	16 (6)	8 (8.1)	8 (4.8)	.28
Decline to answer	14 (5.3)	7 (7.1)	7 (4.2)	.31
Ethnicity (n=263)				.70
Yes, Hispanic or Latino	39 (14.8)	14 (14.1)	25 (15.2)	
No, not Hispanic or Latino	207 (78.7)	77 (77.8)	130 (79.3)	
Decline to answer	17 (6.5)	8 (8.1)	9 (5.5)	
Relationship status (n=266)				.93
Married or civil union	154 (57.9)	61 (61.6)	93 (55.7)	
Engaged or in a relationship	17 (6.4)	5 (5.1)	12 (7.2)	
Single, never married	35 (13.2)	12 (12.1)	23 (13.8)	
Separated	11 (4.1)	4 (4)	7 (4.2)	
Divorced	40 (15)	15 (15.2)	25 (15)	
Widowed	5 (1.9)	1 (1)	4 (2.4)	
Decline to answer	4 (1.5)	1 (1)	3 (1.8)	
Education (n=266)				.29
Some high school	1 (0.4)	1 (1)	0 (0)	
High school graduate or equivalent (eg, GED[a])	21 (7.9)	9 (9.1)	12 (7.2)	
Some college or vocational School	72 (27.1)	26 (26.3)	46 (27.5)	
Associate degree	36 (13.5)	8 (8.1)	28 (16.8)	
Bachelor degree	62 (23.3)	28 (28.3)	34 (20.4)	
Master degree	59 (22.2)	23 (23.2)	36 (21.6)	
Professional school degree (eg, MD, DDC, and JD)	6 (2.3)	2 (2)	4 (2.4)	
Other doctoral degree (eg, PhD and EdD)	4 (1.5)	0 (0)	4 (2.4)	

	Total, n (%)	Duration of MHC use		*P* value
		Less than 3 months, n (%)	3 months or longer, n (%)	
Decline to answer	5 (1.9)	2 (2)	3 (1.8)	
Travel time to VHA[b] (n=265)				.73
<15 minutes	31 (11.7)	11 (11.2)	20 (12)	
16-30 minutes	85 (32.1)	29 (29.6)	56 (33.5)	
31-60 minutes	102 (38.5)	43 (43.9)	59 (35.3)	
61-120 minutes	38 (14.3)	12 (12.2)	26 (15.6)	
>120 minutes	9 (3.4)	3 (3.1)	6 (3.6)	
Proportion of VHA care that was received remotely in the prior year? (n=242)				.14
0-10	11 (4.5)	2 (2.2)	9 (6)	
11-49	13 (5.4)	2 (2.2)	11 (7.3)	
50-74	41 (16.9)	20 (22)	21 (13.9)	
75-89	43 (17.8)	15 (16.5)	28 (18.5)	
90-100	134 (55.4)	52 (57.1)	82 (54.3)	

[a]GED: General Education Development.

[b]VHA: Veterans Health Administration.

Key Findings

Overview

Findings from the integrated quantitative and qualitative data sets supported one another and highlighted 4 key themes [23]. Specifically, (1) the use of the MHC app had positive impacts on care processes for veterans, including increased engagement in mental health care and self-management and (2) enhanced veteran-provider communication and rapport. However, (3) veterans also reported a variety of challenges using MHC. Last, (4) interviewees offered suggestions for improving the app's future functionality and adoption.

Theme #1: Enhancing Self-Management and Engagement in Mental Health Care

Veterans reported completing a variety of assessments in the MHC app: PTSD (209/267, 78%), depression (191/267, 72%), and anxiety (180/267, 67%) were most common; others included assessments related to sleeping habits (125/267, 47%), quality of life (102/267, 38%), alcohol use (75/267, 28%), and physical health (68/267, 26%). Approximately two-thirds (169/268) of survey respondents indicated that they had been using MHC for ≥3 months. Veterans using the app for ≥3 months rated their overall satisfaction with the app similarly to those who used MHC for <3 months. However, a greater proportion of veterans using the app for ≥3 months endorsed several benefits of using the app (Table 2). A greater proportion of veterans who used MHC for ≥3 months also reported agreement that MHC helped to improve their VHA mental health care when compared with those who used it for <3 months (101/163, 62% versus 44/98, 45% agreed or strongly agreed, respectively; *P*=.007).

The MHC app offers veterans the ability to view their assessment data in graphs over time. While only 39% (n=104) of survey respondents reported viewing graphs of their assessment results, those who did described how the visualizations helped them gauge personal progress.

> *We were able to use that graph also in talking about, like, "Hey, during this time frame we were really peaking in your therapy." You could see where your severity was high and how it's come down and leveled out. So, I think that was also a good visual to be able to use as well.* [veteran 0522]

For some, the visualizations of their assessment data were an encouraging factor to continue therapy.

> *I really wanted to give it an opportunity to see if I could make any progress because it's been a challenge for so long. Then after four or five weeks we looked back at the scores and where I was and how far I had come. Even though it wasn't by much, it was enough to give me more encouragement to stay in therapy to keep me going, because she [provider] would go back, and she would review the graph and you could see in the graph where you're making progress.* [Veteran 0367]

Seeing their assessment scores also enhanced some veterans' self-awareness. As one veteran said, "it lets you know whether you're in the 'green level' or whether you need to seek more help." Others remarked:

> *It really helps me. It sure helps my therapist to see those numbers, but it helps me go, "Gosh, yeah, we're still here? I'm still stuck?" It definitely helps you if you're stuck to realize that you're still stuck.* [Veteran 0230]

> *I can look at past surveys [ie, self-assessments] and, noticing that I have higher scores on some weeks, I think back, okay, what was going on during that week that may have caused my scores to be higher? Then*

I can go back and reflect on those and say, "Okay, what could have been done differently, if anything?" [Veteran 0367]

Table 2. Impact of Mental Health Checkup (MHC) app on veteran self-management and engagement (n=268).

	Total, n (%)	Duration of MHC use		*P* value
		Less than 3 months, n (%)	3 months or longer, n (%)	
Overall satisfaction with MHC (n=262)				.22
Satisfied/strongly satisfied	189 (72.1)	65 (67.7)	124 (74.7)	
Neutral/somewhat dissatisfied/very dissatisfied	73 (27.9)	31 (32.3)	42 (25.3)	
Using MHC has helped with improving health and health care engagement (n=262)				.03
Agree/strongly agree	169 (64.5)	55 (56.1)	114 (69.5)	
Neutral/disagree/strongly disagree	93 (35.5)	43 (43.9)	50 (30.5)	
Understanding of one's health condition (n=262)				.02
Agree/strongly agree	171 (65.3)	55 (56.1)	116 (70.7)	
Neutral/disagree/strongly disagree	91 (34.7)	43 (43.9)	48 (29.3)	
Health management (n=262)				.01
Agree/strongly agree	154 (58.8)	48 (49)	106 (64.6)	
Neutral/disagree/strongly disagree	108 (41.2)	50 (51)	58 (35.4)	
Health-related goal-setting (n=262)				.06
Agree/strongly agree	156 (59.5)	51 (52)	105 (64)	
Neutral/disagree/strongly disagree	106 (40.5)	47 (48)	59 (36)	
Health-related goal-achievement (n=262)				.03
Agree/strongly agree	133 (50.8)	41 (41.8)	92 (56.1)	
Neutral/disagree/strongly disagree	129 (49.2)	57 (58.2)	72 (43.9)	
Completion of assessments (n=262)				.50
Agree/strongly agree	196 (74.8)	71 (72.4)	125 (76.2)	
Neutral/disagree/strongly disagree	66 (25.2)	27 (27.6)	39 (23.8)	
More frequently completing assessments (n=262)				.10
Agree/strongly agree	174 (66.4)	59 (60.2)	115 (70.1)	
Neutral/disagree/strongly disagree	88 (33.6)	39 (39.8)	49 (29.9)	
Improving VHA[a] mental health care (n=261)				.007
Agree/strongly agree	145 (55.6)	44 (44.9)	101 (62)	
Neutral/disagree/strongly disagree	116 (44.4)	54 (55.1)	62 (38)	

[a]VHA: Veterans Health Administration.

Theme #2: Enhancing Veteran-Provider Communication and Rapport

A greater proportion of veterans who had used MHC for ≥3 months (versus <3 months) agreed that using the app helped them better communicate (112/164, 68% versus 51/98, 52% agreed/strongly agreed; *P*=.009) and improve rapport (95/164, 58% versus 42/98, 43% agreed/strongly agreed; *P*=.02) with their VHA providers (Table 3).

Consistent with the survey findings, in interviews, veterans reported that using MHC facilitated communication with their providers, particularly initiating or directing therapy conversations. One veteran reported that MHC allowed their provider to "open the book on me." Others said:

It kind of gives them [providers] a ballpark idea and it gives them an opportunity of where they may start in asking questions and maybe where to open up the conversation, especially since I'm very private and unless I come to the table with a specific topic, I may not be one to open very easily. So, he [provider] may need some kind of information to help him pull information out of me. And this [sharing the Mental Health Checkup scores] can help him do that. [Veteran 0522]

It makes you open up because once you accept...this is what's happening to you, this is the way you've been feeling, it's easier when she [provider] has it in front of her and she's talking to me. That app helped *me to face exactly what she was trying to get me to face because it was there – it was happening.* [Veteran 0066]

Table 3. Impact of Mental Health Checkup (MHC) on veteran-provider interactions (n=268).

	Total, n (%)	Duration of MHC use		*P* value
		Less than 3 months, n (%)	3 months or longer, n (%)	
Using MHC has helped with communicating with VHA[a] providers (n=262)				.009
Agree/strongly agree	163 (62.2)	51 (52)	112 (68.3)	
Neutral/disagree/strongly disagree	99 (37.8)	47 (48)	52 (31.7)	
VHA providers understanding how veterans are doing in between visits (n=262)				.51
Agree/strongly agree	201 (76.7)	73 (74.5)	128 (78)	
Neutral/disagree/strongly disagree	61 (23.3)	25 (25.5)	36 (22)	
Improving rapport with veterans' providers (n=262)				.02
Agree/strongly agree	137 (52.3)	42 (42.9)	95 (57.9)	
Neutral/disagree/strongly disagree	125 (47.7)	56 (57.1)	69 (42.1)	
Making treatment decisions with veterans' VHA providers (n=262)				.14
Agree/strongly agree	157 (59.9)	53 (54.1)	104 (63.4)	
Neutral/disagree/strongly disagree	105 (40.1)	45 (45.9)	60 (36.6)	

[a]VHA: Veterans Health Administration.

Other veterans explained that sharing assessment scores from MHC improved communication with their providers because it helped them focus and structure their therapy time. As one veteran stated, the assessment scores help his provider "work more on what she knows I'm in need of and what we're supposed to be going into." Others said:

> *We didn't have to waste time at the beginning of my appointment going over questions about "How are you doing?". It was already done, so you could get into things, where let's say you said that you're really depressed, you know, you don't have to go into asking that question to begin with.* [Veteran 0501]

> *I guess it gave me the feeling that it made it more structured, which is something I like. I like structure.* [veteran 0505]

Importantly, discussing assessment findings also facilitated rapport and trust.

> *Psychologists in the past were just like, "Oh yeah, just fill this out" and, you know, they never talked about it with me. I didn't know what I was filling out.... This [provider] actually does care. That kind of gave me a breath of fresh air, like "Oh okay, he takes his job very seriously and cares about me and my health".... Now we know what the underlining issue is, and I think, without the app, we probably never would have done that.* [Veteran 0077]

> *It has made me more comfortable and trusting, where we [veteran and provider] decide together what's needed, for adjustment or to stay the same. So, it's like I don't have that stress of worrying about what's gonna happen to me because I know whatever we come up with.... It's gonna be something that is really gonna benefit me and help me.* [Veteran 0111]

Theme #3: Challenges Using MHC

While veterans described how the use of the MHC app had positive impacts on their self-management and engagement in care, as well as their interactions with their providers, both the survey and interview findings indicated that MHC posed several challenges for veterans.

Survey findings indicated that veterans who used MHC for <3 months reported almost all challenges at a similar frequency as veterans who used the app for ≥3 months (Table 4). However, a greater proportion of veterans who had used the MHC app for <3 months reported that not being able to answer assessment questions as well as they could when they were at a VHA facility was a somewhat serious or serious challenge (32/98, 33% versus 30/166, 18%; *P*=.007). Despite the challenges reported, 78% (n=207) of veterans expressed that they were comfortable using the app, and over 41% (n=108) described not having trouble using the MHC app. When veterans encountered trouble using MHC (N=154), the most frequently reported actions they took were securely messaging (84/154, 55%) or calling their VHA providers (57/154, 37%), and to a lesser extent, calling the VHA Help Desk for technical support (29/154, 19%).

Table 4. Veteran-reported challenges when using the Mental Health Checkup (MHC) app by duration of use (n=268).

	Total, n (%)	Duration of MHC use		*P* value
		Less than 3 months, n (%)	3 months or longer, n (%)	
Unable to access assessments (n=264)				.63
Somewhat/serious challenge	69 (26.1)	27 (27.8)	42 (25.1)	
Not a challenge	195 (73.9)	70 (72.2)	125 (74.9)	
All needed assessments not in MHC (n=248)				.46
Somewhat/serious challenge	51 (20.6)	21 (23.1)	30 (19.1)	
Not a challenge	197 (79.4)	70 (76.9)	127 (80.9)	
Unable to correct erroneous responses (n=256)				.25
Somewhat/serious challenge	72 (28.1)	31 (32.3)	41 (25.6)	
Not a challenge	184 (71.9)	65 (67.7)	119 (74.4)	
Unable to save answers (n=257)				.25
Somewhat/serious challenge	52 (20.2)	23 (24)	29 (18)	
Not a challenge	205 (79.8)	73 (76)	132 (82)	
Difficulty understanding graphs of assessment scores (n=245)				.11
Somewhat/serious challenge	102 (41.6)	43 (48.3)	59 (37.8)	
Not a challenge	143 (58.4)	46 (51.7)	97 (62.2)	
Inability of providers to see entered answers (n=246)				.35
Somewhat/serious challenge	39 (15.9)	17 (18.7)	22 (14.2)	
Not a challenge	207 (84.1)	74 (81.3)	133 (85.8)	
Providers and veterans do not talk about assessment scores (n=260)				.97
Somewhat/serious challenge	70 (26.9)	26 (26.8)	44 (27)	
Not a challenge	190 (73.1)	71 (73.2)	119 (73)	
Providers do not use assessment scores to inform my care (n=247)				.60
Somewhat/serious challenge	63 (25.5)	26 (27.4)	37 (24.3)	
Not a challenge	184 (74.5)	69 (72.6)	115 (75.7)	
Inability to answer questions as well as when at the VHA[a] (n=264)				.007
Somewhat/serious challenge	62 (23.5)	32 (32.7)	30 (18.1)	
Not a challenge	202 (76.5)	66 (67.3)	136 (81.9)	
Lacking a private place to complete assessments (n=263)				.97
Somewhat/serious challenge	19 (7.2)	7 (7.1)	12 (7.3)	
Not a challenge	244 (92.8)	91 (92.9)	153 (92.7)	
Insufficient training about MHC (n=259)				.64
Somewhat/serious challenge	73 (28.2)	29 (29.9)	44 (27.2)	
Not a challenge	186 (71.8)	68 (70.1)	118 (72.8)	
Providers did not explain how they wanted MHC used (n=258)				.18
Somewhat/serious challenge	65 (25.2)	29 (29.9)	36 (22.4)	
Not a challenge	193 (74.8)	68 (70.1)	125 (77.6)	
Not enough help available for troubleshooting with MHC (n=254)				.56
Somewhat/serious challenge	64 (25.2)	22 (23.2)	42 (26.4)	
Not a challenge	190 (74.8)	73 (76.8)	117 (73.6)	
Limited veteran access to a mobile/computing device (n=263)				.50
Somewhat/serious challenge	23 (8.7)	7 (7.2)	16 (9.6)	

	Total, n (%)	Duration of MHC use		P value
		Less than 3 months, n (%)	3 months or longer, n (%)	
Not a challenge	240 (91.3)	90 (92.8)	150 (90.4)	
Limited veteran access to the Internet or sufficient Wi-Fi capacity (n=264)				.39
Somewhat/serious challenge	33 (12.5)	10 (10.2)	23 (13.9)	
Not a challenge	231 (87.5)	88 (89.8)	143 (86.1)	
Concerns about privacy and security of responses entered into MHC (n=263)				.49
Somewhat/serious challenge	66 (25.1)	22 (22.7)	44 (26.5)	
Not a challenge	197 (74.9)	75 (77.3)	122 (73.5)	

[a]VHA: Veterans Health Administration.

Veterans described similar challenges within the follow-up semistructured interviews. Notably, many of the challenges they discussed were not unique to the MHC app; rather, they pertained to uncertainty about the purpose of MBC; emotional burden related to completing assessments; and challenges related to PRO assessments more generally.

One veteran, confused about the impact of his scores on his overall mental health care, noted his uncertainty about the purpose of measurement: “I wasn’t sure what the results or effect was gonna be.”

Some veterans also reported emotional burdens related to completing assessments. For example, completing assessments could be emotionally triggering.

> *I hate doing them because it just brings up, you know, everything I’m trying to avoid.... It just reminds me like I’m not getting any better or still have these issues.* [Veteran 0077]

Emotional burden was also experienced when a veteran felt anxiety related to the assessments: feeling “pressure” about not scoring well or “failing the program” if they did not score well and the implications this might have for their provider:

> *...when you don’t see progress...it almost like nullifies the work you’re doing, right, so you’re like, oh,.... I’m failing at this program...and I’m also failing the provider that I’m working with because now he can’t show his provider or his boss that he’s being successful with me, so that adds another level – another layer of – of something to work through. ... It’s another layer of pressure. ... it was always hanging out there, right, you know, like almost a sense of dread, like, okay, I gotta show some dots moving up, you know?* [Veteran 0041]

Last, some veterans experienced challenges with PRO assessments themselves. First, a veteran found assessing their emotional status “... hard to quantify.... There’s no definition of what severe means... ” (Veteran 0041)

Assessments could also be challenging given competing demands on a veteran’s attention and time:

> *I think the problem is more me than you guys or your app, you know, in my lack of mindfulness and my forgetfulness and my just lifestyle don’t make it super-easy to carve out time to give it the attention that it deserves.* [Veteran 0047]

Another veteran pointed out how completing the same type of assessment over time was “tedious...just because it’s the same questions every time.” (Veteran 0087)

Veterans found inadequate instructions about how to use the app to be a further challenge in completing assessments.

Theme #4: Suggestions for Improving MHC

In addition to sharing comments about the challenges they faced, interviews also provided an ideal opportunity for veterans to share their thoughts about how to overcome some of the aforementioned challenges and how to improve the functionality of the MHC app. Technical suggestions included reducing the number of steps to log into MHC to just a username and password, and including a reminder function (eg, an alarm) or a link to one’s electronic calendar to support the completion of assessments and, by extension, sustained engagement. Other suggestions pertained to the assessments themselves. veterans described the potential utility of defining scale terms (eg, does “frequently” mean 3 times/week or 5 times/week?), reordering assessment items so they do not feel redundant, and reducing the time frames over which veterans are asked to reflect upon and recall the symptoms they are experiencing. Veterans also wanted providers to talk to them about how assessments “fit into the big picture” of their mental health care. Finally, some veterans offered their thoughts about ways to support the broader implementation of the MHC app, including marketing campaigns within and beyond clinical settings, providing individual training opportunities and further technical support, and ensuring awareness of the MHC app and its consistent use in practice among VHA mental health providers.

Discussion

Principal Findings

Survey and interview findings revealed that, although veteran experiences varied, use of MHC, an app designed to support R-MBC for mental health care in the VHA, appeared to be associated with several self-reported increases in mental health care benefits, particularly for veterans who used the app for ≥3 months. Compared with veterans using MHC for <3 months, veterans who used MHC for ≥3 months were more likely to report that MHC helped them be more engaged in treatment, manage their health, and communicate with their VHA providers. These findings were echoed among assertions from interviewed veterans. At the same time, veterans also identified several usability challenges and other issues related to MBC more generally and the use of technologies to support R-MBC, as well as opportunities to improve upon MHC specifically.

These findings extend previous work that has shown that MBC improves communication between providers and patients [11,26,27]. While veterans in our sample also reported improved communication with providers, a unique feature of our findings is that veterans further expressed a sense of enhanced rapport, engagement, and self-management of their conditions. The specific features offered by technology platforms that are intended to support R-MBC can also reflect different philosophies about the use of patient health information to support clinical care.

Previous literature has suggested that MBC approaches may be used to bolster patient self-management and engagement in care [28]. As noted above, and unlike some other existing R-MBC platforms, the MHC app offers veterans the ability to review their own assessment scores and to view graphical representations of them over time. Our results similarly suggest that these are meaningful features that help at least some veterans to be more self-aware, motivated to continue their treatment, and stay in touch with their own progress. It is possible that such features could facilitate veteran symptom self-management and engagement in mental health care, and perhaps should be included in other R-MBC technology platforms.

Additionally, whereas most studies to date have focused on provider perspectives toward R-MBC [29-31], ours is one of a few that has examined patient perspectives and experiences with the processes of R-MBC and the kinds of technologies that can be used to support them, an equally important factor for uptake and implementation [32,33]. Finally, our evaluation's use of semistructured interviews enabled veterans to characterize the breadth of their experiences in their own words and to offer open-ended suggestions, rather than being restricted to choosing from among predetermined survey response options. Such feedback may be particularly helpful for enhancing MBC and R-MBC processes and the technologies used to support them, in that veterans have the opportunity to share, in narrative form, their thoughts on app enhancement strategies and ways to best incorporate veteran perspectives into MBC processes.

Our qualitative work yielded additional findings new to the field; that is, veterans offered valuable suggestions that could enhance the usability and future implementation of the MHC app as well as other possible R-MBC technology platforms among veterans and their VHA mental health providers. Suggestions included offering new ways to remind veterans to complete the assessments they have been assigned as well as offering support to help veterans understand the meaning of the graphical representations of their assessment results within the app. This may help veterans stay more engaged between the time they complete an assessment and their next appointment with their VHA mental health provider.

Challenges related to log-in validation and other usability issues underscored the potential benefit of having readily available training and technical support to enhance veteran uptake and continued use of the MHC app. It is also noteworthy that some veterans offered feedback that suggested they might benefit from a greater understanding of the goals of MBC more generally, which, in turn, might facilitate their completion of assessments and sharing of assessment data. Specifically, veterans suggested that it may be helpful for providers to explain what the PRO assessments they are asked to complete are intended to measure, why these assessments were assigned, and how the assessments are structured so that they are easier to navigate.

Taken together, our findings speak to the potential of the MHC app as a platform that can support VHA's Collect-Share-Act model of MBC. Completing assessments through the MHC app (ie, the "collecting" stage of the Collect-Share-Act model) facilitated veterans' abilities to express their emotions, which some described to be otherwise challenging. Additionally, the ability to view graphs of their assessment results helped some veterans gauge their personal progress and was a source of encouragement to continue with their therapy for others, as was the feedback that some veterans received from their mental health providers (ie, the "sharing" stage of Collect-Share-Act), especially for those who felt "stuck" or lacking in motivation. Veterans also felt that receiving feedback about their assessment scores played an important role in their willingness to use MHC given that, for some veterans, completing mental health assessments through a technology like the MHC app could be potentially upsetting. Finally, veterans expressed satisfaction with data being used during clinical appointments (ie, the "acting" stage of Collect-Share-Act) to inform treatment decisions such as how best to continue care. In sum, this evaluation of MHC underscored that each phase of the Collect-Share-Act model serves a critical role in supporting MBC for mental health care to help meet veteran mental health care needs.

Limitations

Our surveys and interviews captured the self-reported behaviors and attitudes of veterans who have used the MHC app; thus, these methods may be affected by common problems of self-reported data including recall, desirability, and other types of bias. Because we conducted data collection during the COVID-19 pandemic, a large proportion (134/242, 55%) of the veterans in this evaluation reported that 90%-100% of their care was virtual over the year preceding participation. Attitudes toward the MHC app may have differed if this evaluation was conducted during a time when care at a distance using technology was not considered "standard" practice. It is also possible that interviews conducted via telephone may have restricted the incorporation of nonverbal cues that may have enriched our analysis. Additionally, because our survey and interview samples comprised veterans who had at least some preceding experience using the MHC app and, by extension, were established users of VHA mental health care, our findings may not translate to those veterans who are newer to VHA mental health services. Future work to evaluate the effectiveness of MHC should qualitatively assess the experiences of veterans who have used MHC for longer durations to better understand their perspectives toward the usability of the app, benefits, challenges, and suggestions for enhancements. Future work may also benefit from a more rigorous design, such as randomization and less reliance on self-reported data, to evaluate the relationships between MHC use and clinical outcomes. Finally, like the general population, digital disparities exist in the veteran population [34,35]; technical training and targeted assistance to older, rural-residing, and less technologically literate veterans will be key to the success of any R-MBC app, including MHC.

Conclusions

Among a national sample of veterans, those who used the VHA's MHC app for ≥3 months were more likely to report better engagement and self-management of their condition (ie, understanding their condition, managing their health, and achieving their health and health care goals), increased communication and rapport with their VHA providers, and improved aspects of their mental health care compared with veterans who used the app for <3 months. Despite some challenges with usability, the overwhelming majority reported that they were comfortable using the app. Suggestions from interviewed veterans included streamlining the log-in process, implementing enhanced training and technical assistance, reminder tools, orientation to assessment design, and enhanced marketing to veterans and providers. These findings suggest that MHC may be a valuable tool to enhance veteran mental health care in accordance with the VHA's Collect-Share-Act model and should be further evaluated to assess its effectiveness in enhancing clinical outcomes.

Acknowledgments

This work was supported by the US Department of Veterans Affairs; Office of Connected Care; and Office of Research and Development, Health Services Research and Development Service, Quality Enhancement Research Initiative Program (PEC 15-470; principal investigator: TPH). The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

Conflicts of Interest

None declared.

References

1. Trivedi RB, Post EP, Sun H, Pomerantz A, Saxon AJ, Piette JD, et al. Prevalence, comorbidity, and prognosis of mental health among US veterans. Am J Public Health 2015;105(12):2564-2569. [doi: 10.2105/AJPH.2015.302836] [Medline: 26474009]
2. Houston TK, Volkman JE, Feng H, Nazi KM, Shimada SL, Fox S. Veteran internet use and engagement with health information online. Mil Med 2013;178(4):394-400 [FREE Full text] [doi: 10.7205/MILMED-D-12-00377] [Medline: 23707823]
3. Jiang CY, Strohbehn GW, Dedinsky RM, Raupp SM, Pannecouk BM, Yentz SE, et al. Teleoncology for veterans: high patient satisfaction coupled with positive financial and environmental impacts. JCO Oncol Pract 2021;17(9):e1362-e1374 [FREE Full text] [doi: 10.1200/OP.21.00317] [Medline: 34406817]
4. Smith AH, Touchett H, Chen P, Fletcher T, Arney J, Hogan J, et al. Patient satisfaction with a coach-guided, technology-based mental health treatment: qualitative interview study and theme analysis. JMIR Ment Health 2024;11:e50977 [FREE Full text] [doi: 10.2196/50977] [Medline: 38306167]
5. Fleuty K, Almond MK. Remote access therapy for veterans with psychological problems: current state of the art. Mil Med 2020;185(7-8):e1046-e1050 [FREE Full text] [doi: 10.1093/milmed/usaa020] [Medline: 32060551]
6. Fletcher TL, Amspoker AB, Wassef M, Hogan JB, Helm A, Jackson C, et al. Increasing access to care for trauma-exposed rural veterans: a mixed methods outcome evaluation of a web-based skills training program with telehealth-delivered coaching. J Rural Health 2022;38(4):740-747. [doi: 10.1111/jrh.12628] [Medline: 34648188]
7. Kuhn E, Kanuri N, Hoffman JE, Garvert DW, Ruzek JI, Taylor CB. A randomized controlled trial of a smartphone app for posttraumatic stress disorder symptoms. J Consult Clin Psychol 2017;85(3):267-273. [doi: 10.1037/ccp0000163] [Medline: 28221061]
8. Possemato K, Kuhn E, Johnson E, Hoffman JE, Owen JE, Kanuri N, et al. Using PTSD coach in primary care with and without clinician support: a pilot randomized controlled trial. Gen Hosp Psychiatry 2016;38:94-98. [doi: 10.1016/j.genhosppsych.2015.09.005] [Medline: 26589765]
9. Bush NE, Smolenski DJ, Denneson LM, Williams HB, Thomas EK, Dobscha SK. A virtual hope box: randomized controlled trial of a smartphone app for emotional regulation and coping with distress. Psychiatr Serv 2017;68(4):330-336 [FREE Full text] [doi: 10.1176/appi.ps.201600283] [Medline: 27842473]
10. Lewis CC, Boyd M, Puspitasari A, Navarro E, Howard J, Kassab H, et al. Implementing measurement-based care in behavioral health: a review. JAMA Psychiatry 2019;76(3):324-335 [FREE Full text] [doi: 10.1001/jamapsychiatry.2018.3329] [Medline: 30566197]

11. Scott K, Lewis CC. Using measurement-based care to enhance any treatment. Cogn Behav Pract 2015;22(1):49-59 [FREE Full text] [doi: 10.1016/j.cbpra.2014.01.010] [Medline: 27330267]
12. Dowrick C, Leydon GM, McBride A, Howe A, Burgess H, Clarke P, et al. Patients' and doctors' views on depression severity questionnaires incentivised in UK quality and outcomes framework: qualitative study. BMJ 2009;338:b663 [FREE Full text] [doi: 10.1136/bmj.b663] [Medline: 19299474]
13. Goldstein LA, Gibbons MBC, Thompson SM, Scott K, Heintz L, Green P, et al. Outcome assessment via handheld computer in community mental health: consumer satisfaction and reliability. J Behav Health Serv Res 2011;38(3):414-423 [FREE Full text] [doi: 10.1007/s11414-010-9229-4] [Medline: 21107916]
14. Zhu M, Hong RH, Yang T, Yang X, Wang X, Liu J, et al. The efficacy of measurement-based care for depressive disorders: systematic review and meta-analysis of randomized controlled trials. J Clin Psychiatry 2021;82(5):21r14034 [FREE Full text] [doi: 10.4088/JCP.21r14034] [Medline: 34587377]
15. Boswell JF, Hepner KA, Lysell K, Rothrock NE, Bott N, Childs AW, et al. The need for a measurement-based care professional practice guideline. Psychotherapy (Chic) 2023;60(1):1-16. [doi: 10.1037/pst0000439] [Medline: 35771518]
16. About VHA. U.S. Department of Veterans Affairs. 2022. URL: https://www.va.gov/health/aboutVHA.asp [accessed 2022-09-08]
17. Mental health services. USA.gov. URL: https://www.benefits.gov/benefit/4747 [accessed 2022-09-09]
18. Peterson K, Anderson J, Bourne D. VA evidence-based synthesis program reports. In: Evidence Brief: Use of Patient Reported Outcome Measures for Measurement Based Care in Mental Health Shared Decision-Making. Washington, DC: Department of Veterans Affairs (US); 2018.
19. Krägeloh CU, Czuba KJ, Billington DR, Kersten P, Siegert RJ. Using feedback from patient-reported outcome measures in mental health services: a scoping study and typology. Psychiatr Serv 2015;66(3):224-241 [FREE Full text] [doi: 10.1176/appi.ps.201400141] [Medline: 25727110]
20. Goldberg SB, Buck B, Raphaely S, Fortney JC. Measuring psychiatric symptoms remotely: a systematic review of remote measurement-based care. Curr Psychiatry Rep 2018;20(10):81. [doi: 10.1007/s11920-018-0958-z] [Medline: 30155749]
21. Yeung AS, Jing Y, Brenneman SK, Chang TE, Baer L, Hebden T, et al. Clinical Outcomes in Measurement-Based Treatment (COMET): a trial of depression monitoring and feedback to primary care physicians. Depress Anxiety 2012;29(10):865-873. [doi: 10.1002/da.21983] [Medline: 22807244]
22. Simon GE, Ralston JD, Savarino J, Pabiniak C, Wentzel C, Operskalski BH. Randomized trial of depression follow-up care by online messaging. J Gen Intern Med 2011;26(7):698-704 [FREE Full text] [doi: 10.1007/s11606-011-1679-8] [Medline: 21384219]
23. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs-principles and practices. Health Serv Res 2013;48(6 Pt 2):2134-2156 [FREE Full text] [doi: 10.1111/1475-6773.12117] [Medline: 24279835]
24. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
25. Patton MQ. Qualitative Research and Evaluation Methods. 3rd Edition. Thousand Oaks, CA: Sage Publications; 2002.
26. Torous J, Powell AC, Rodriguez-Villa E. Health information technology resources to support measurement-based care. Child Adolesc Psychiatr Clin N Am 2020;29(4):763-773. [doi: 10.1016/j.chc.2020.06.011] [Medline: 32891375]
27. Resnick SG, Oehlert ME, Hoff RA, Kearney LK. Measurement-based care and psychological assessment: using measurement to enhance psychological treatment. Psychol Serv 2020;17(3):233-237 [FREE Full text] [doi: 10.1037/ser0000491] [Medline: 32881578]
28. Ferrari M, Pawliuk N, Pope M, MacDonald K, Boruff J, Shah J, et al. A scoping review of measures used in early intervention services for psychosis. Psychiatr Serv 2023;74(5):523-533. [doi: 10.1176/appi.ps.202100506] [Medline: 36321318]
29. Oslin DW, Hoff R, Mignogna J, Resnick SG. Provider attitudes and experience with measurement-based mental health care in the VA implementation project. Psychiatr Serv 2019;70(2):135-138 [FREE Full text] [doi: 10.1176/appi.ps.201800228] [Medline: 30373495]
30. Tauscher JS, Cohn EB, Johnson TR, Diteman KD, Ries RK, Atkins DC, et al. What do clinicians want? understanding frontline addiction treatment clinicians' preferences and priorities to improve the design of measurement-based care technology. Addict Sci Clin Pract 2021;16(1):38 [FREE Full text] [doi: 10.1186/s13722-021-00247-5] [Medline: 34130724]
31. Goldstein DA, Meyers K, Endsley M, Zerth EO. Measurement-based care implementation in a veterans affairs primary care-mental health integration program. Psychol Serv 2020;17(3):323-331. [doi: 10.1037/ser0000370] [Medline: 31318239]
32. Katz IR, Resnick S, Hoff R. Associations between patient experience and clinical outcomes in general mental health clinics: findings from the veterans outcomes assessment survey. Psychiatry Res 2021;295:113554. [doi: 10.1016/j.psychres.2020.113554] [Medline: 33183768]
33. Almklov E, Afari N, Floto E, Lindamer L, Hurst S, Pittman JOE. Post-9/11 veteran satisfaction with the VA eScreening program. Mil Med 2020;185(3-4):519-529 [FREE Full text] [doi: 10.1093/milmed/usz324] [Medline: 32060553]
34. Luger TM, Hogan TP, Richardson LM, Cioffari-Bailiff L, Harvey K, Houston TK. Older veteran digital disparities: examining the potential for solutions within social networks. J Med Internet Res 2016;18(11):e296 [FREE Full text] [doi: 10.2196/jmir.6385] [Medline: 27881361]

35. Swed O, Sheehan CM, Butler JS. The digital divide and veterans' health: differences in self-reported health by internet usage. Armed Forces Soc 2018;46(2):238-258. [doi: 10.1177/0095327x18809069]

Abbreviations

MBC: measurement-based care
MHC: Mental Health Checkup
NASSS: Non-Adoption, Abandonment, and Challenges to the Scaleup, Spread, and Sustainability
PRO: patient-reported outcome
PTSD: posttraumatic stress disorder
R-MBC: remote measurement-based care
VHA: Veterans Health Administration

Edited by J Torous; submitted 31.10.23; peer-reviewed by O Cohen Elimelech, K Hallgren; comments to author 03.01.24; revised version received 28.02.24; accepted 15.03.24; published 10.05.24.

Please cite as:
Higashi RT, Etingen B, Richardson E, Palmer J, Zocchi MS, Bixler FR, Smith B, McMahon N, Frisbee KL, Fortney JC, Turvey C, Evans J, Hogan TP
Veteran Experiences With an mHealth App to Support Measurement-Based Mental Health Care: Results From a Mixed Methods Evaluation
JMIR Ment Health 2024;11:e54007
URL: https://mental.jmir.org/2024/1/e54007
doi:10.2196/54007
PMID:38728684

Original Paper

Web-Based Therapist-Guided Mindfulness-Based Cognitive Behavioral Therapy for Body Dysmorphic Disorder: Pilot Randomized Controlled Trial

Camrie Kerry[1], MSc; Prabhdeep Mann[1], BSc; Nazanin Babaei[2], MSc; Joel Katz[3], PhD; Meysam Pirbaglou[1], PhD; Paul Ritvo[1], PhD

[1]School of Kinesiology and Health Sciences, York University, Toronto, ON, Canada
[2]Department of Psychology, University of Victoria, Victoria, BC, Canada
[3]Department of Psychology, York University, Toronto, ON, Canada

Corresponding Author:
Camrie Kerry, MSc
School of Kinesiology and Health Sciences
York University
4700 Keele St
Toronto, ON, M3J 1P3
Canada
Phone: 1 416 580 8021
Email: camrie@yorku.ca

Abstract

Background: Internet-based cognitive behavioral therapy (CBT) and stand-alone mindfulness meditation interventions are gaining empirical support for a wide variety of mental health conditions. In this study, we test the efficacy of web-based therapist-guided mindfulness-based cognitive behavioral therapy (CBT-M) for body dysmorphic disorder (BDD), a psychiatric disorder characterized by preoccupations with perceived defects in appearance.

Objective: This study aims to determine whether CBT-M for BDD delivered on the web is feasible and acceptable and whether mindfulness meditation adds to CBT treatment effects for BDD.

Methods: In this 8-week, 2-arm, parallel pilot randomized controlled trial, n=28 adults (aged between 18 and 55 years) were randomly allocated to an experimental group (web-based therapist-guided CBT-M) or a control group (web-based therapist-guided CBT). Study retention, accrual, and intervention adherence were assessed, along with self-report measures for BDD, depression, anxiety, and pain intensity taken at baseline and postintervention.

Results: This study was feasible to implement and deemed acceptable by participants. After 8 weeks, significant improvements were found on all outcome measures for both treatment groups, and large between-group effect sizes favoring CBT-M were found for BDD symptom severity (d=–0.96), depression (d=–1.06), pain severity (d=–1.12), and pain interference (d=–1.28). However, linear mixed models demonstrated no significant differences between the groups over 8 weeks.

Conclusions: The results suggest that mindfulness meditation may add to beneficial web-based CBT treatment effects for BDD. An adequately powered randomized control trial of web-based CBT-M is warranted.

Trial Registration: ClinicalTrials.gov NCT05402475, http://clinicaltrials.gov/ct2/show/NCT05402475

(JMIR Ment Health 2024;11:e55283) doi:10.2196/55283

KEYWORDS

body dysmorphic disorder; BDD; dysmorphophobia; obsessive-compulsive and related disorders; OCD; internet-delivered cognitive behavior therapy; iCBT; cognitive behavior therapy; mindfulness-based cognitive therapy; mindfulness; eMental health; randomized controlled trial

Introduction

Background

Body dysmorphic disorder (BDD) is a psychiatric disorder marked by excessive preoccupations with perceived physical defects that are slight or imperceptible to others [1]. Individuals with BDD perform repetitive behaviors such as analyzing perceived defects in reflective surfaces, seeking reassurance from others, excessive grooming, and camouflaging with oversized clothing or makeup [2]. Preoccupations with appearance and associated behaviors often lead to significant distress and functional impairment, including difficulties attending work or school [3] that render many (nearly 30%) housebound [4,5]. Furthermore, suicidal ideation and suicide attempts are common in BDD-affected populations [6], with completed suicide rates nearly 45 times higher than the general population's [7], reflecting the debilitating morbidity of BDD.

BDD symptoms typically arise during adolescence and impact 1.7% to 2.4% of the general population over the life span [6-9]. While population-based studies indicate a higher proportion of BDD-affected individuals are female (nearly 60%), demographics, body areas of concern, symptom severity, clinical features, and impairment appear similar between genders [2,8,9]. However, these and other BDD-related data may be approximate, as BDD is understudied, underdiagnosed, and undertreated [6], partly because individuals reluctantly disclose BDD symptoms due to feelings of shame [10]. The reluctance to disclose may lead BDD-affected individuals to seek out cosmetic enhancements at an observed prevalence of 6% to 15% [11-13].

Individuals affected by BDD typically do not benefit from surgical treatment, while cognitive behavioral therapy (CBT) offers an efficacious treatment alternative [14]. Mindfulness meditation has also been associated with reductions in body image distress through procedures that support appearance acceptance and present-moment awareness [15]. Despite the potential of multiple treatment options, several factors may impede access to BDD treatment, including poor insight, lower socioeconomic status, shame, geographical barriers, and psychotherapeutic ambivalence [16-18].

Surveys have identified that only 17.4% of individuals who actively sought psychotherapy services for BDD received CBT, despite its empirical support [5,17]. Internet-based CBT for BDD may increase treatment access, with research investigating the efficacy of internet-based CBT (iCBT) for BDD identified as encouraging. In Sweden, Enander et al [19] conducted a 12-week randomized controlled trial (RCT) comparing iCBT for BDD with web-based supportive therapy. Results revealed a between-group effect size of d=0.95 (95% CI 0.52 to 1.38) after treatment, indicating substantially reduced BDD outcomes for the intervention group when compared to supportive therapy. In 2020, Wilhelm et al [20] developed and pilot-tested a CBT digital service, marking the first smartphone-delivered individual CBT treatment for BDD. Although substantial symptom reduction was reported after the 12-week open trial (d=2.60), the efficacy indications must be interpreted cautiously as this study was an underpowered, nonrandomized trial (n=10 participants) in which depression outcomes did not meaningfully reduce. Ultimately, these aforementioned factors underscore the importance of expanding BDD research to identify appropriate and accessible clinical interventions.

While iCBT interventions address multiple BDD treatment barriers and demonstrate impressive empirical support, there are readily identified needs for improvement. First, CBT interventions for BDD appear to minimally reduce comorbid depression [14] which is unfortunate given that individuals with BDD have high rates of suicidal ideation, functional impairment, and major depressive disorder [2,3,6]. There is evidence from existing web-based therapist-guided RCTs that mindfulness-based CBT (CBT-M) interventions reduce depression symptoms [21,22]. Although these trials did not include BDD participants, they demonstrate that accessible and effective treatments exist for depressed subpopulations. While previous web-based CBT for BDD protocols have incorporated mindfulness [23], CBT-M interventions deliberately emphasize mindfulness practice linked to techniques for engaging in self-compassion [24], self-acceptance, and nonjudgement [25], demonstrating significant promise for reducing body image distress [15]. This is essential to consider for those affected by BDD, given their vulnerability to negative self-appearance fixations. Findings from a study focused on the short-term effectiveness of targeting intrusive appearance-linked thoughts through mindfulness meditation identified that positive affect increased in BDD individuals compared to healthy controls [15]. As mindfulness techniques implicitly and explicitly emphasize relaxation, nonjudgement, and nonreactivity [26], participants may develop a strengthened propensity to engage in CBT practices by identifying maladaptive beliefs that drive BDD and are found to be modifiable.

Aims of the Study

While the efficacy of web-based therapist-guided CBT-M for BDD is heretofore unknown, reductions in BDD symptoms already achieved in web-based CBT and mindfulness meditation interventions suggest a combined treatment approach is an important investigative priority. Furthermore, the current trial isolates mindfulness components to better understand their treatment effects.

Curiously, despite self-injurious behavior patterns such as skin-picking, excess exercise, suicide attempts [2] and cosmetic surgery engagement [27], research investigating physical pain in BDD appears sparse. In extreme cases, self-mutilation, self-surgery [2] and requests for healthy limb amputations [28] have been reported in individuals affected by BDD. This notable gap in pain investigation was additionally prioritized in this study.

We anticipate that web-based intervention implementation will be feasible, indicated by rates of accrual, retention, and attendance at therapist-guided calls, and acceptability will be demonstrated by responses on the NexJ Program Experience survey. Furthermore, we hypothesize that the CBT-M group participants will reveal greater posttreatment reductions in BDD, depression, anxiety, and pain according to between-group effect sizes, and preliminary efficacy for both groups will be demonstrated by quantitative outcomes (symptom reductions) assessed at baseline and postintervention.

Methods

Participants

Recruitment was undertaken through advertisements on web-based platforms (Facebook [Meta Platforms, Inc], Reddit [Reddit, Inc], and CloudResearch [Prime Research Solutions LLC]), and direct referrals were accepted from the Centre for Addiction and Mental Health (Toronto, Ontario).

Inclusion Criteria

US or Canadian residents aged between 18 and 55 years; fluent in English; having smartphone access; BDD diagnosis based on Body Dysmorphic Disorder Questionnaire (BDDQ) responses [29] must answer "yes" to the following questions: (1) Are you worried about how you look? (2) Do you think about your appearance problems a lot and wish you could think about them less? Must answer "yes" to any of the following questions: (1) How has this problem with how you look affected your life? Has it often upset you a lot? (2) Has it often gotten in the way of doing things with friends, dating, your relationships with people, or your social activities? (3) Has it caused you any problems with school, work, or other activities? (4) Are there things you avoid because of how you look? Eligible participants must additionally indicate that they spend ≥1 hour each day thinking about how they look.

Exclusion Criteria

Individuals whose self-reports are congruent with the *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]) criteria for: eating disorder, bipolar disorder, borderline personality disorder, schizophrenia (or other primary psychotic disorder) or severe substance abuse disorder or addiction; disclosure of imminent intent or attempted suicide in the past 6 months; concurrent psychological treatment; and no smartphone access.

Sample Size

Following Julious' [30] pilot recommendations, 24 participants were allocated to 2 comparison groups, and 28 participants were enrolled for a final recruitment target of 24, assuming 15% (4/28) attrition.

Screening

Participants were screened on the web with Survey Monkey (Survey Monkey, Inc), a HIPAA (Health Insurance Portability and Accountability Act)-compliant platform for safeguarding data collection. The BDDQ [29] was used to detect BDD in combination with a psychological history interview.

Data Collection

Participants completed a web-based consent form and 4 questionnaires: the Body Dysmorphic Disorder-Symptom Scale [31], the Patient Health Questionnaire-9 (PHQ-9) [32], the Generalized Anxiety Disorder-7 Scale (GAD-7) [33], and the Brief Pain Inventory (BPI) [34]. The web-based assessment used a SurveyMonkey link with the additional safeguard of a unique study ID number.

Upon questionnaire completion, participants provided phone numbers for receipt of therapist-guided calls associated with the NexJ Connected Wellness platform. Self-reported psychometric data were collected from participants at baseline (T1) and postintervention (T2). Participants were compensated US $30 for intervention participation and measurement completion.

Randomization

A 1:1 ratio randomization schedule was used with randomly selected block sizes of 4 and 2 treatment arm allocations (CBT-M or CBT) based on a randomization sequence generator. Participants were allocated to treatment groups after providing consent and before baseline measurement.

Interventions

After participants were randomly allocated to treatment, the principal investigator (Dr Paul Ritvo) assigned participants to student-therapists for weekly guided calls. Participants were additionally connected to the NexJ Connected Wellness platform to access BDD-related content and to enable secure message exchanges with their therapist.

CBT-M participants received 24/7 accessible CBT and mindfulness content through NexJ Connected Wellness. The content was based on 2 previous internet-based CBT-M RCTs with distressed students and individuals diagnosed with major depressive disorder [21,35]. Altogether, 8 workbook chapters focused on (1) perfectionism, unchangeable physical features, media influences and comparisons; (2) automatic negative thoughts, repetitive behaviors, and mindful acceptance; (3) mindful nonreactivity, cognitive restructuring, and behavior modification; (4) body-based assumptions and befriending your body; (5) overcoming avoidance, and internal versus external body image; (6) diaphragmatic breathing, loving-kindness meditation, and cultivating compassion to others and self; (7) bringing health from most preferred body parts to least preferred body parts, shifting to a neutral focus, and adjusting assumptions; and (8) objective self-talk, shifting attentional focus, and continuing the intervention independently.

While the CBT-only group did not receive mindfulness meditation content (mindful acceptance, mindful nonreactivity, loving-kindness meditation, compassion to others and self, and shifting attentional focus), both groups were provided with 8 modules and engaged in 1 hour of self-directed CBT-M or CBT content per week. The CBT-M group was provided with ~135 minutes of mindfulness audio to select from.

Client-centered therapist calls were conducted for 60 minutes per week over 8 weeks, which involved discussing module content and progress. All study therapists attended group training and 1-on-1 supervision sessions weekly.

Measures

The primary feasibility outcomes were measured by an accrual rate (total number of enrolled participants divided by the number of months recruitment occurred), retention (percentage of participants who completed the 8-week intervention from randomization to completion of postintervention measures) and adherence (percentage of participants attending weekly and scheduled therapist-guided calls). The primary acceptability outcome was measured by the NexJ Program Experience survey, a 7-item questionnaire (1-5 rating scale) developed by NexJ Health Inc wherein higher scores represent greater participant satisfaction. The Body Dysmorphic Disorder Symptom Scale (BDD-SS), a reliable and valid self-report questionnaire was used to examine the severity of BDD symptoms [31]. This study uses the 7 severity items, with higher scores representing greater severity (of symptoms), up to a sum score of 70.

Secondary outcomes for depression, anxiety and pain were measured using the PHQ-9 [32], the GAD-7 [33], and the BPI Short Form [34], respectively. The PHQ-9 is a reliable and valid 9-item self-report questionnaire for depression screening and severity, with a sensitivity of 88% and specificity of 88% [32]. The GAD-7 is a valid measure for screening and assessing the severity of generalized anxiety disorder, with a sensitivity of 89% and a specificity of 82%, indicating good reliability [33]. Lastly, the BPI (Short Form) is a reliable self-report questionnaire for physical pain experiences, using 4 pain severity items and 7 pain interference items. The BPI has demonstrated good to excellent validity and reliability [34].

Statistical Analyses

SPSS Statistics (version 28.0; IBM Corp) for Windows was used to analyze self-reported data. Numeric variables were presented as means and SDs, and categorical data were presented as frequencies and percentages.

Independent sample 2-tailed *t* tests and the Fisher exact test for all numeric and categorical variables were used to detect differences between study groups at baseline and between study completers and dropouts to determine whether missing postintervention measurements were considered missing at random. Linear mixed model (LMM) analyses for repeated measures were used in an intention-to-treat approach. An unstructured restricted maximum likelihood approach was used for the main analysis ("Results" section). Fixed effects, including group and time, and their interaction (group × time) were evaluated. Cohen *d* within-group and between-group effect sizes were evaluated with means, SDs, and correlations calculated for all participants with completed data (n=18).

Pearson *r* bivariate correlations were computed to assess the relationship between BDD-SS and BPI scores.

Ethical Considerations

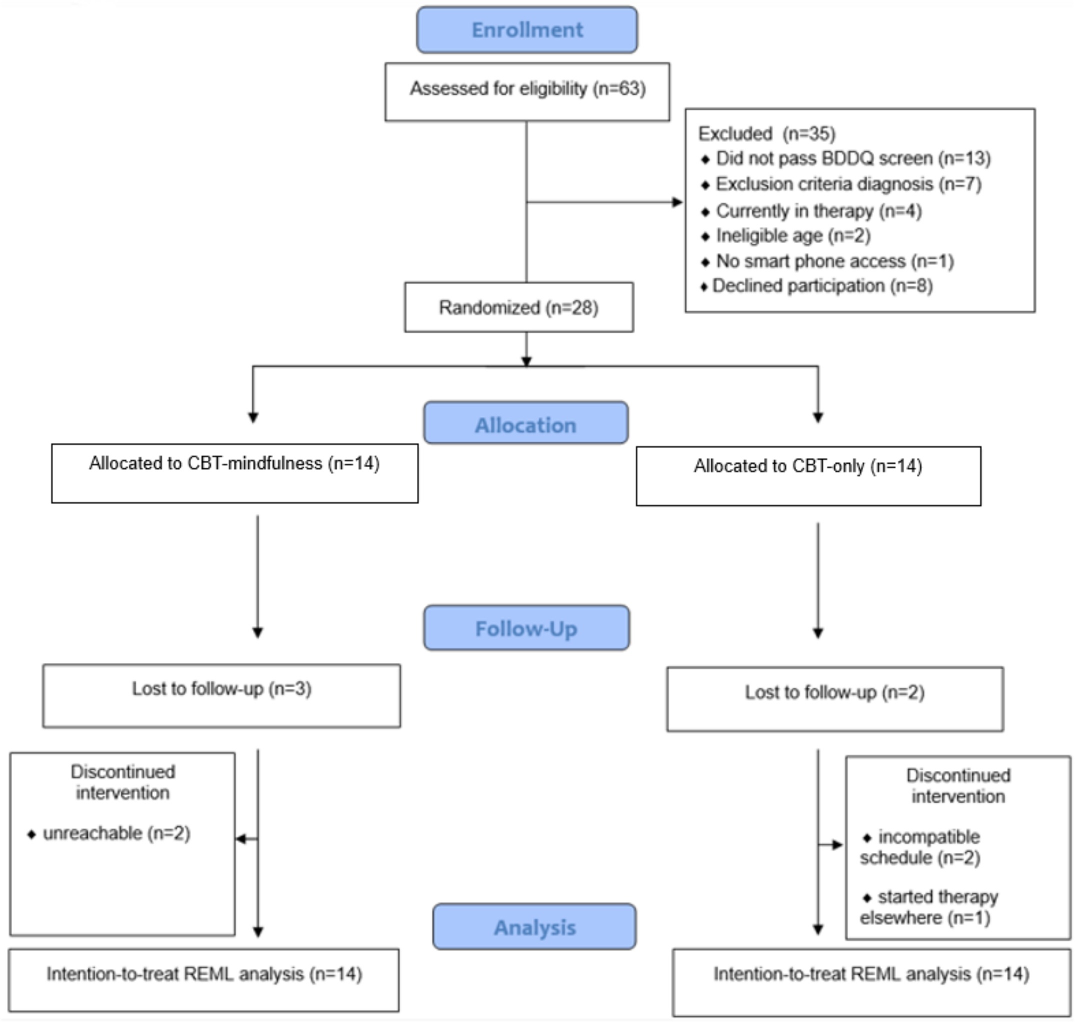
This 8-week 2-arm pilot RCT used a web-based software platform, NexJ Connected Wellness (NexJ Health Inc) and was approved by York University Research and Ethics and registered with ClinicalTrials.gov (NCT05402475).

Results

Participant Flow

The CONSORT flow diagram (Figure 1) illustrates the recruitment and flow of 63 adults (aged between 18 and 55 years) who were screened and interviewed to determine eligibility (September 2022 to February 2023). Of those interviewed, 35 adults were ineligible because: they did not pass BDDQ screening (n=13), were excluded due to psychiatric diagnosis (n=7), current therapy engagement (n=4), ineligible age (n=2), lack of smart phone access (n=1), or decided not to participate (n=8). A total of 28 adults met inclusion criteria and provided informed consent to participate. Demographic and psychological characteristics for both CBT-M and CBT groups are presented in Table 1.

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) participant flow diagram. BDDQ: Body Dysmorphic Disorder Questionnaire; CBT: cognitive behavioral therapy; REML: restricted maximum likelihood.



XSL•FO
RenderX

Table 1. Demographic and psychological characteristics for both groups at baseline (N=28).

Variable	CBT-M[a]	CBT[b]	*P* value
Age (years), mean (SD)	31.14 (7.81)	34.43 (11.33)	.93
Gender, n (%)			.38
Female	7 (50)	10 (71.4)	
Male	5 (35.7)	4 (28.6)	
Other	2 (14.3)	0 (0)	
Ethnicity, n (%)			.53
White	8 (57.1)	7 (50)	
Black	2 (14.3)	3 (21.4)	
South Asian	2 (14.3)	1 (7.1)	
East Asian	0 (0)	2 (14.3)	
Latin American	0 (0)	1 (7.1)	
Multiethnic	2 (14.3)	0 (0)	
Highest level of education, n (%)			.58
High school	3 (21.4)	1 (7.1)	
College	1 (7.1)	3 (21.4)	
Bachelor's degree	8 (57.1)	6 (42.9)	
Master's degree	2 (14.3)	3 (21.4)	
Other	0 (0)	1 (7.1)	
Marital status, n (%)			.70
Married or common-law	7 (50)	5 (35.7)	
Single	7 (50)	9 (64.3)	
Psychological variables, mean (SD)			
BDD-SS[c] (Severity)	35.50 (10.47)	41.64 (11.24)	.15
PHQ-9[d]	10.07 (3.69)	11.36 (5.44)	.47
GAD-7[e]	8.36 (4.34)	8.57 (4.13)	.90
BPI[f] (severity 0-10)	2.57 (2.2)	2.66 (2.29)	.92
BPI (interference 0-10)	2.26 (2.32)	3.41 (3.27)	.29

[a]CBT-M: mindfulness-based CBT.
[b]CBT: cognitive behavioral therapy.
[c]BDD-SS: Body Dysmorphic Disorder Symptom Scale.
[d]PHQ-9: Patient Health Questionnaire-9.
[e]GAD-7: Generalized Anxiety Disorder-7.
[f]BPI: Brief Pain Inventory.

Intervention Feasibility

A total of 28 eligible participants enrolled during the 5 months of recruitment (September 2022-February 2023) for an accrual rate of 5.6 participants per month. A total of 10 enrollees did not complete the 8-week intervention for a 64% (18/28) retention rate.

All 18 retained participants completed 8 scheduled counseling calls, although 5 participants had to reschedule 1 call (n=5-CBT) and another 5 participants had to reschedule 2 calls (n=2-CBT-M and n=3-CBT). About 44% (8/18) of participants maintained attendance for scheduled weekly calls throughout the study.

Intervention Acceptability

All 18 participants who completed the intervention provided program satisfaction ratings at 8 weeks. Participants rated their experience from 1-5 on the NexJ Program Experience survey (Table 2). The mean score for the question "How would you rate your overall program experience?" was 4.67 (SD 0.59), where 1 indicated "poor" and 5 indicated "excellent." In total, 72% (13/18) of participants identified their overall experience in the study was "excellent" while 22% (4/18) of participants indicated their experience was "very good," and one participant (0.06%; 1/18) indicated their experience was "good." Other acceptance indications are found in Table 2.

Table 2. Results from the NexJ Program Experience survey postintervention. Satisfaction scores ranged from 1-5.

Satisfaction metric	CBT-M[a], mean (SD)	CBT[b], mean (SD)	Mean (SD)
How would you rate your overall program experience?	4.78 (0.67)	4.56 (0.53)	4.67 (0.59)
On average, how much time did you spend on the program each week?	3.89 (0.78)	3.78 (0.83)	3.83 (0.79)
To what extent did the program meet your needs?	4.56 (1.01)	4.44 (0.88)	4.5 (0.92)
How would you rate the ease of using our platform?	3.89 (1.27)	4.67 (0.50)	4.28 (1.02)
Please indicate your agreement with the following statement: The information in the modules helped me work towards my mental health goals	3.33 (1.23)	3.22 (1.20)	3.28 (1.18)
Overall, my experience with my therapist was:	5 (0)	4.78 (0.67)	4.89 (0.47)
How likely are you to continue using our platform over the next 6 months?	3 (1.5)	3.67 (1.12)	3.33 (1.33)

[a]CBT-M: mindfulness-based cognitive behavioral therapy.

[b]CBT: cognitive behavioral therapy.

Preliminary Efficacy

Overview

As shown in Table 1, independent sample *t* tests and Fisher exact tests revealed no significant differences between groups on baseline measures for all assessed variables. This suggests that randomization allocation resulted in reasonably equivalent treatment groups. In addition, the main analysis did not include covariates within the statistical model. Independent sample *t* tests and Fisher exact tests revealed no significant differences at baseline in dropout patterns (those who stayed in vs dropped out): age (*P*=.41), gender (*P*=.84), education (*P*=.89), ethnicity (*P*=.25), marital status (*P*=.69), BDD-SS score (*P*=.82), PHQ-9 score (*P*=.55), GAD-7 score (*P*=.67), BPI (severity) score (*P*=.95), or BPI (interference) score (*P*=.84). As a result, missing postintervention measurements (n=10) were considered missing at random. Unstructured restricted maximum likelihood data (Table 3) and Cohen *d* effect sizes for each psychometric outcome, along with correlation coefficients between BDD and pain, are discussed below.

Table 3. Results from linear mixed model (LMM) analysis for changes in outcomes from baseline to postintervention between and within intervention groups (intention-to-treat using restricted maximum likelihood).

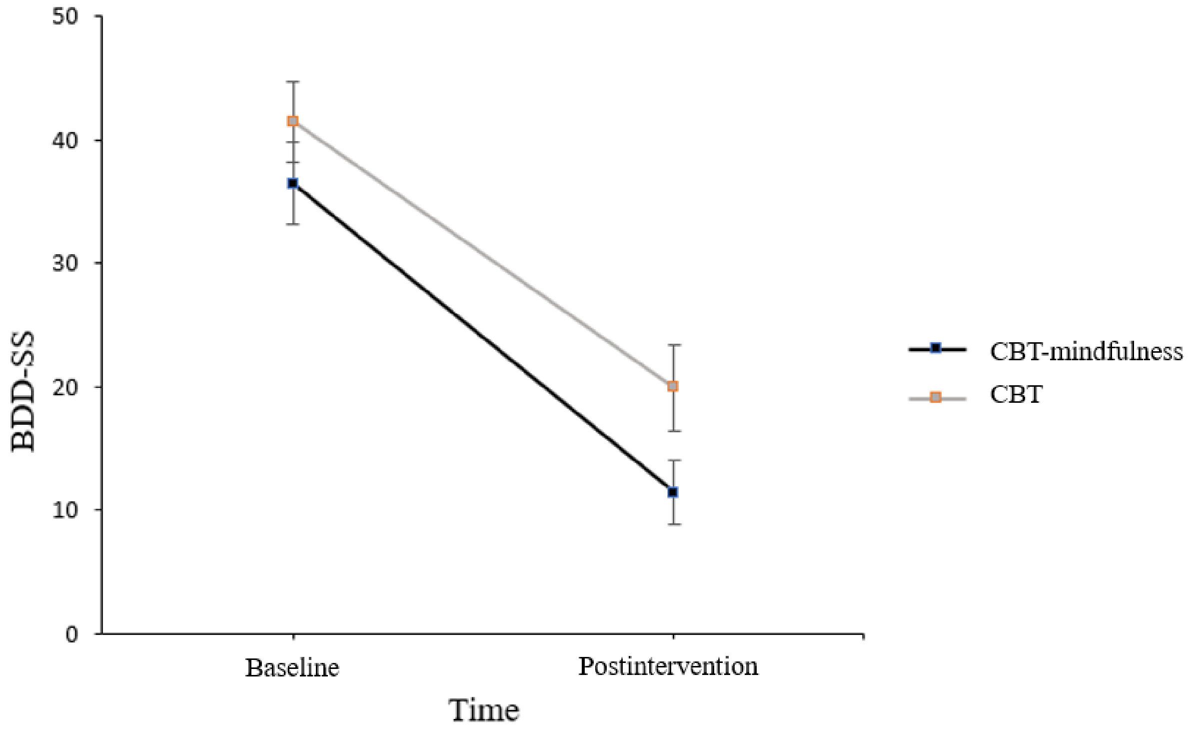
Outcomes	CBT-M[a]	CBT[b]	*d* (between-groups)	Group		Time		Group×*time*	
				F test (*df*)	*P* value	*F* test (*df*)	*P* value	*F* test (*df*)	*P* value
BDD-SS[c] (Severity)				5.12 (1, 25.63)	.03	76.93 (1, 22.24)	<.001	0.24 (1, 22.24)	.63
Baseline	36.44 (12.39)	41.44 (12.2)	—[d]						
8-weeks	11.44 (7.84)	19.89 (10.61)	–0.96						
d (within-groups)	–2.41	–1.87	—						
PHQ-9[e]				2.52 (1, 26.42)	.13	32.8 (1, 24.26)	<.001	0.72 (1, 24.26)	.40
Baseline	10.22 (4.3)	12 (5.79)	—						
8-weeks	4.44 (2.3)	7.33 (3.39)	–1.06						
d (within-groups)	–1.61	–0.91	—						
GAD-7[f]				0.001 (1, 23.64)	.98	7.11 (1, 19.53)	.02	0.05 (1, 19.53)	.83
Baseline	8.89 (4.54)	8.56 (4.59)	—						
8-weeks	5.67 (5.83)	5.22 (2.64)	–0.11						
d (within-groups)	–0.61	–0.87	—						
BPI[g] (Severity)				1.76 (1, 25.26)	.20	7.80 (1, 24.18)	.01	1.87 (1, 24.18)	.18
Baseline	2.57 (2.2)	2.66 (2.29)	—						
8-weeks	0.53 (0.76)	1.97 (1.77)	–1.12						
d (within-groups)	–1.20	–0.33	—						
BPI (Interference)				3.63 (1, 25.95)	.07	15.04 (1, 25.42)	.001	0.08 (1, 25.42)	.78
Baseline	2.26 (2.32)	3.41 (3.27)	—						
8-weeks	0.16 (0.29)	1.60 (1.66)	–1.28						
d (within-groups)	–1.34	–0.62	—						

[a]CBT-M: mindfulness-based cognitive behavioral therapy.
[b]CBT: cognitive behavioral therapy.
[c]BDD-SS: Body Dysmorphic Disorder Symptom Scale.
[d]Not available.
[e]PHQ-9: Patient Health Questionnaire-9.
[f]GAD-7: Generalized Anxiety Disorder-7.
[g]BPI: Brief Pain Inventory.

BDD-SS

LMM analysis revealed a statistically significant main effect for group ($F_{1,25.63}$=5.12, P=.03); a statistically significant main effect for time ($F_{1,22.24}$=76.93, P<.001); but no significant group by time interaction ($F_{1,22.24}$=0.24, P=.63). This suggests no statistically significant difference in BDD-SS change scores between the CBT-M and CBT groups from baseline to postintervention (Figure 2). A between- and within-groups evaluation of Cohen *d* effect sizes indicated a between-group effect size at 8 weeks of d=–0.96, and within-group effect sizes of d=–2.41 for CBT-M and d=–1.87 for CBT.

Figure 2. Body Dysmorphic Disorder Symptom Scale (BDD-SS) change from baseline to postintervention. Error bars represent standard error. CBT: cognitive behavioral therapy.



PHQ-9

Results from LMM analysis revealed a nonsignificant group by time interaction ($F_{1,24.26}$=0.72, *P*=.40). In addition, no statistically significant main effects were found for group ($F_{1,26.42}$=2.52, *P*=.13), although main effects for time did reveal statistical significance ($F_{1,24.26}$=32.79, *P*<.001). Cohen *d* effect sizes revealed a between-group effect size of *d*=–1.06 at 8 weeks. The within-group effect size for CBT-M was *d*=–1.61 and *d*=–0.91 for CBT.

GAD-7

A statistically significant main effect for time ($F_{1,19.53}$=7.11, *P*=.02) was found in the LMM analysis for GAD-7 scores. However, the main effects for group ($F_{1,23.64}$=0.001, *P*=.98), and the group by time interaction ($F_{1,19.53}$=0.05, *P*=.83) were not statistically significant. The Cohen *d* between-group effect size at 8 weeks was *d*=–0.11. The within-group effect size for CBT-M was *d*=–0.61 and *d*=–0.87 for CBT.

BPI (Severity)

LMM analysis revealed a statistically significant main effect for time ($F_{1,24.18}$=7.80, *P*=.01) for BPI (severity). The main effects for group ($F_{1,25.26}$=1.76, *P*=.20) and the group by time interaction ($F_{1,24.18}$=1.87, *P*=.18) were not statistically significant. Cohen *d* effect sizes were calculated for between-group and within-group changes, revealing a between-group effect size at 8 weeks of *d*=–1.12, and within-group effect sizes of *d*=–1.20 for CBT-M and *d*=–0.33 for CBT.

A positive correlation between BDD-SS and BPI (severity) at baseline, *r*=0.29, was not statistically significant (*P*=.14).

BPI (Interference)

LMM analysis revealed a nonstatistically significant main effect for group ($F_{1,25.95}$=3.63, *P*=.07), and group by time interaction ($F_{1,25.42}$=0.08, *P*=.78) for pain interference; however, the main effect for time ($F_{1,25.42}$=15.04, *P*=.001) was statistically significant. Cohen *d* evaluation of effect sizes revealed a between-group effect size of *d*=–1.28 at 8 weeks and within-group effect sizes of *d*=–1.34 for CBT-M and *d*=–0.62 for CBT.

A statistically significant, positive correlation was found between BDD-SS and BPI (interference) at baseline (*r*=0.56, *P*=.002).

Discussion

Overview

In this pilot RCT, the feasibility, acceptability, and relative efficacy of a web-based therapist-guided CBT-M intervention were compared with a web-based therapist-guided CBT-only intervention. Data indicate that, altogether, 28 participants were enrolled within 5 months of recruitment (September 2022 to February 2023) despite the inclinations of BDD-affected individuals to seek nonpsychiatric treatment at high rates [27]. These data challenge past data suggesting that individuals with BDD may be difficult to recruit for psychological treatment studies and support advocacy for more intervention access [14]. Adherence to scheduled phone-based counseling calls was another positive indicator, as 100% of study completers attended all 8 scheduled calls.

A moderate retention rate of 64% (18/28) suggests the RCT was implementable on the web with room for improvement.

Effective implementation of the RCT was further indicated by satisfaction responses on the NexJ Program Experience survey. Most participants (13/18, 72%) indicated their program experience was "excellent," and a good experience (or better) was reported by all study completers. Most participants (13/18, 72%) felt as though "almost all needs" were met, and participants found the NexJ Connected Wellness platform easy to use with a mean score of 4.28 (SD 1.02) out of an optimal score of 5.

Satisfaction ratings on module content were mixed, with 44% (8/18) indicating a neutral response, with 39% (7/18) finding modules helpful, and 17% (3/18) finding them not helpful. Overall discrepancies between satisfaction and low module ratings have been attributed to participant distress and the inconvenience of homework assignments in the CBT framework [36]. Findings also indicate a linkage between ambivalence about treatment tasks [37] and homework noncompliance, a challenging issue in CBT delivery [38]. Beyond the data presented above, additional data pertaining to participant use of module content were not obtained. However, 94% (17/18) of participants reported an excellent experience with their therapist during counseling calls, and only one indicated a "good" experience. These data align with other web-based CBT trials, amplifying the benefits of therapist support [19]. Of course, the use of an interactive platform with counseling calls makes it difficult to tease apart the benefits of web-based module access from the benefits of counseling calls.

Although no significant between-group differences from baseline to postintervention were observed on the BDD-SS, a statistically significant reduction in symptom severity from baseline to postintervention across both groups was found, as demonstrated by the LMM analysis. This supports our hypothesis that web-based therapist-guided CBT-M treatment for BDD will demonstrate preliminary efficacy.

Given the pilot sample, emphasis was placed on the between-group effect sizes for the proposed hypothesis that web-based therapist-guided CBT-M would be more effective than CBT alone. A between-group effect size of d=–0.96 at postintervention revealed that mindfulness meditation may contribute to beneficial treatment outcomes for BDD. Furthermore, very large within-group effect sizes were found for both CBT-M (d=–2.41), and CBT (d=–1.87) participants. These findings parallel results from a smartphone-based CBT for BDD RCT, where a within-group effect size of d=–2.26 (95% CI –2.93 to –1.58) was observed at 12 weeks [23]. This study achieved similar reductions in BDD after 8 weeks compared to Wilhelm et al's 12-week intervention [23]. In combination with a web-based approach, a shorter intervention may prove cost-effective while reducing waitlists and clinician or therapist time [39]. These findings are relevant to the dearth of trained BDD clinicians and limited access to treatment [40].

Additional scales indicated comorbid symptom reductions. Accordingly, the large within-group effect sizes for web-based therapist-guided CBT (d=–0.91) and web-based therapist-guided CBT-M (d=–1.61), along with the statistically significant time effect, support the hypothesis that web-based therapist-guided CBT-M would reduce depression symptoms. Although large effect sizes were observed in both groups, web-based therapist-guided CBT-M was associated with larger effects. A large between-group effect size (d=–1.06) was also found, suggesting that mindfulness meditation in combination with CBT contributes to reductions in BDD-related depression. The effect size observed exceeds findings from 5 previous CBT for BDD RCTs reported in a previous meta-analysis [14], which revealed a moderate overall effect size (d=–0.49, 95% CI −0.76 to −0.22) for depression symptom reductions. Given that BDD and MDD are highly comorbid and have longitudinal associations [41], web-based therapist-guided CBT-M's preliminary effectiveness for BDD-related depression is encouraging.

LMM analysis and Cohen's d between-group effect size (d=–0.11) for anxiety revealed that treatment effects did not statistically differ between groups, rejecting our hypothesis that web-based therapist-guided CBT-M would be more effective than CBT alone. Although significant symptom severity improvements were observed after 8 weeks on GAD-7 scores in both groups, only a moderate CBT-M within-group effect size was found (d=–0.61). These findings are comparable to GAD-7 within-group effect sizes observed in a 16-week CBT for BDD RCT, d=0.65 [42].

With limited insight into whether BDD and pain are associated, the BPI was used to investigate variations in pain interference and severity. As an increased risk of self-injurious behavior is prevalent (eg, excessive exercise, restrictive eating, skin-picking, cosmetic surgery, self-surgery, suicide attempts, alcohol or drug dependency, and steroid abuse) [2,43], Pearson r correlation coefficients were calculated to assess the relationship between BDD and pain. While a statistically significant, positive correlation was found between BDD and pain interference, a weaker positive correlation was found between BDD and pain severity. Although pain exploration is a notable gap in BDD research, there is some evidential support for associations between body dissatisfaction and pain in eating disorder populations. Specifically, researchers found that body dissatisfaction may induce a greater sensitivity to bodily pain [44]. As body dissatisfaction is a cornerstone for BDD diagnosis, this research supports the preliminary insights in our study. Furthermore, in another study, when healthy individuals were confronted with distorted images of their own bodies, pain perception increased [45].

These data suggest that web-based therapist-guided CBT-M for BDD can reduce pain interference and pain severity scores, as demonstrated by statistically significant time effects. Although LMM did not reveal significant between-group differences, large between-group effect sizes for pain interference (d=–1.28), and pain severity (d=–1.12) were observed. The within-group effect sizes for pain severity greatly differed between treatment groups, with a large effect (d=–1.20) for web-based therapist-guided CBT-M and a small effect (d=–0.33) for CBT alone. Moreover, within-group effect sizes notably differed for pain interference (CBT-M d=–1.34 and CBT d=–0.62). The large effect sizes observed in the web-based therapist-guided CBT-M group for both pain severity and interference may be due to better chronic pain management obtained through mindfulness meditation practice [46].

Limitations

Several limitations must be considered when interpreting this study's findings. Given the pilot nature of this RCT, the small sample size limited the capacities to test for significance between treatment groups. In addition, the BDDQ screening tool was used for inclusion rather than a clinician-administered diagnostic measure. This may have resulted in an unrepresentative sample.

Although CBT-M module content and therapist-guided calls emphasized mindfulness meditation, data pertaining to participants' historical meditation practices and active practice time throughout the 8-week intervention were not gathered, which limits understandings of the dose-response relationship [47]. Moreover, future CBT-M studies would optimally include a mindfulness measure such as the Five Facet Mindfulness Questionnaire [48] to establish linkage between potential mechanisms and BDD reductions.

Conclusion

In this pilot RCT, two 8-week web-based interventions were compared in the treatment of BDD. Both interventions used therapist-guided CBT, but only one intervention combined CBT with mindfulness meditation approaches. Given the high accrual and adherence rates, moderate retention rate, and overall participant program satisfaction, this RCT was acceptable and feasible to implement. Preliminary efficacy was demonstrated for both active treatment groups, with suggestions that mindfulness meditation could add to CBT treatment effects for BDD and comorbid symptoms. In addition, a relationship between BDD and pain may be present, which requires additional investigation. Given that individuals with BDD may be housebound and have high rates of suicidal ideation and depression, prompt access to effective treatment is imperative. This pilot trial provides promising insight into BDD and short-term, web-based–accessible treatment for individuals with BDD.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-eHealth checklist.
[PDF File (Adobe PDF File), 1185 KB - mental_v11i1e55283_app1.pdf]

References

1. American Psychiatric Association, American Psychiatric Association DSM-5 Task Force. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition: DSM-5™. Washington, D.C: American Psychiatric Publishing; 2013.
2. Phillips KA. The broken mirror: understanding and treating body dysmorphic disorder. New York, NY: Oxford University Press; 2005.
3. Phillips KA, Menard W, Fay C, Pagano ME. Psychosocial functioning and quality of life in body dysmorphic disorder. Compr Psychiatry 2005;46(4):254-260 [FREE Full text] [doi: 10.1016/j.comppsych.2004.10.004] [Medline: 16175755]
4. Phillips KA. Body dysmorphic disorder: recognizing and treating imagined ugliness. World Psychiatry 2004;3(1):12-17 [FREE Full text] [Medline: 16633443]
5. Bjornsson AS, Didie ER, Phillips KA. Body dysmorphic disorder. Dialogues in Clinical Neuroscience 2022;12(2):221-232 [FREE Full text] [doi: 10.31887/dcns.2010.12.2/abjornsson]
6. Buhlmann U, Glaesmer H, Mewes R, Fama JM, Wilhelm S, Brähler E, et al. Updates on the prevalence of body dysmorphic disorder: a population-based survey. Psychiatry Res 2010;178(1):171-175. [doi: 10.1016/j.psychres.2009.05.002] [Medline: 20452057]
7. Phillips KA, Menard W. Suicidality in body dysmorphic disorder: a prospective study. Am J Psychiatry 2006;163(7):1280-1282 [FREE Full text] [doi: 10.1176/ajp.2006.163.7.1280] [Medline: 16816236]
8. Koran LM, Abujaoude E, Large MD, Serpe RT. The prevalence of body dysmorphic disorder in the United States adult population. CNS Spectr 2008;13(4):316-322. [doi: 10.1017/s1092852900016436] [Medline: 18408651]
9. Rief W, Buhlmann U, Wilhelm S, Borkenhagen A, Brähler E. The prevalence of body dysmorphic disorder: a population-based survey. Psychol Med 2006;36(6):877-885. [doi: 10.1017/S0033291706007264] [Medline: 16515733]
10. Conroy M, Menard W, Fleming-Ives K, Modha P, Cerullo H, Phillips KA. Prevalence and clinical characteristics of body dysmorphic disorder in an adult inpatient setting. Gen Hosp Psychiatry 2008;30(1):67-72 [FREE Full text] [doi: 10.1016/j.genhosppsych.2007.09.004] [Medline: 18164943]
11. Phillips KA, Dufresne RG, Wilkel CS, Vittorio CC. Rate of body dysmorphic disorder in dermatology patients. J Am Acad Dermatol 2000;42(3):436-441. [doi: 10.1016/s0190-9622(00)90215-9] [Medline: 10688713]
12. Sarwer DB, Crerand CE. Psychological issues in patient outcomes. Facial Plast Surg 2002;18(2):125-133. [doi: 10.1055/s-2002-32203] [Medline: 12063660]
13. Carroll DH, Scahill L, Phillips KA. Current concepts in body dysmorphic disorder. Arch Psychiatr Nurs 2002;16(2):72-79. [doi: 10.1053/apnu.2002.32109] [Medline: 11925574]

14. Harrison A, Fernández de la Cruz L, Enander J, Radua J, Mataix-Cols D. Cognitive-behavioral therapy for body dysmorphic disorder: a systematic review and meta-analysis of randomized controlled trials. Clin Psychol Rev 2016;48:43-51. [doi: 10.1016/j.cpr.2016.05.007] [Medline: 27393916]
15. Hartmann AS, Thomas JJ, Greenberg JL, Rosenfield EH, Wilhelm S. Accept, distract, or reframe? An exploratory experimental comparison of strategies for coping with intrusive body image thoughts in anorexia nervosa and body dysmorphic disorder. Psychiatry Res 2015;225(3):643-650. [doi: 10.1016/j.psychres.2014.11.031] [Medline: 25530419]
16. Eisen JL, Phillips KA, Coles ME, Rasmussen SA. Insight in obsessive compulsive disorder and body dysmorphic disorder. Compr Psychiatry 2004;45(1):10-15 [FREE Full text] [doi: 10.1016/j.comppsych.2003.09.010] [Medline: 14671731]
17. Marques L, Weingarden HM, Leblanc NJ, Wilhelm S. Treatment utilization and barriers to treatment engagement among people with body dysmorphic symptoms. J Psychosom Res 2011;70(3):286-293. [doi: 10.1016/j.jpsychores.2010.10.002] [Medline: 21334500]
18. Cavanagh K. Geographic inequity in the availability of cognitive behavioural therapy in England and Wales: a 10-year update. Behav Cogn Psychother 2014;42(4):497-501. [doi: 10.1017/S1352465813000568] [Medline: 23829940]
19. Enander J, Andersson E, Mataix-Cols D, Lichtenstein L, Alström K, Andersson G, et al. Therapist guided internet based cognitive behavioural therapy for body dysmorphic disorder: single blind randomised controlled trial. BMJ 2016;352:i241 [FREE Full text] [doi: 10.1136/bmj.i241] [Medline: 26837684]
20. Wilhelm S, Weingarden H, Greenberg JL, McCoy TH, Ladis I, Summers BJ, et al. Development and pilot testing of a cognitive-behavioral therapy digital service for body dysmorphic disorder. Behav Ther 2020;51(1):15-26 [FREE Full text] [doi: 10.1016/j.beth.2019.03.007] [Medline: 32005333]
21. Ritvo P, Ahmad F, El Morr C, Pirbaglou M, Moineddin R, Team MVC. A mindfulness-based intervention for student depression, anxiety, and stress: randomized controlled trial. JMIR Ment Health 2021;8(1):e23491 [FREE Full text] [doi: 10.2196/23491] [Medline: 33264098]
22. Segal ZV, Dimidjian S, Beck A, Boggs JM, Vanderkruik R, Metcalf CA, et al. Outcomes of online mindfulness-based cognitive therapy for patients with residual depressive symptoms: a randomized clinical trial. JAMA Psychiatry 2020;77(6):563-573 [FREE Full text] [doi: 10.1001/jamapsychiatry.2019.4693] [Medline: 31995132]
23. Wilhelm S, Weingarden H, Greenberg JL, Hoeppner SS, Snorrason I, Bernstein EE, et al. Efficacy of app-based cognitive behavioral therapy for body dysmorphic disorder with coach support: initial randomized controlled clinical trial. Psychother Psychosom 2022;91(4):277-285 [FREE Full text] [doi: 10.1159/000524628] [Medline: 35588706]
24. Boellinghaus I, Jones FW, Hutton J. The role of mindfulness and loving-kindness meditation in cultivating self-compassion and other-focused concern in health care professionals. Mindfulness 2012;5(2):129-138 [FREE Full text] [doi: 10.1007/s12671-012-0158-6]
25. Keng SL, Smoski MJ, Robins CJ. Effects of mindfulness on psychological health: a review of empirical studies. Clin Psychol Rev 2011;31(6):1041-1056 [FREE Full text] [doi: 10.1016/j.cpr.2011.04.006] [Medline: 21802619]
26. Kabat-Zinn J. Mindfulness-based interventions in context: past, present, and future. Clin Psychol Sci Pract 2003;10:144-156. [doi: 10.1093/clipsy/bpg016]
27. Crerand CE, Franklin ME, Sarwer DB. Body dysmorphic disorder and cosmetic surgery. Plast Reconstr Surg 2006;118(7):167e-180e. [doi: 10.1097/01.prs.0000242500.28431.24] [Medline: 17102719]
28. Chan JKK, Jones SM, Heywood AJ. Body dysmorphia, self-mutilation and the reconstructive surgeon. J Plast Reconstr Aesthet Surg 2011;64(1):4-8 [FREE Full text] [doi: 10.1016/j.bjps.2010.03.029] [Medline: 20392680]
29. Phillips KA. Body Dysmorphic Disorder: Advances in Research and Clinical Practice. New York, NY: Oxford University Press; 2017.
30. Julious SA. Sample size of 12 per group rule of thumb for a pilot study. Pharmaceutical Statistics 2005;4(4):287-291. [doi: 10.1002/pst.185]
31. Wilhelm S, Greenberg JL, Rosenfield E, Kasarskis I, Blashill AJ. The body dysmorphic disorder symptom scale: development and preliminary validation of a self-report scale of symptom specific dysfunction. Body Image 2016;17:82-87 [FREE Full text] [doi: 10.1016/j.bodyim.2016.02.006] [Medline: 26971118]
32. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
33. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097 [FREE Full text] [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
34. Furler L. Validity and reliability of the pain questionnaire "brief pain inventory". a literature research. Pflege Z 2013;66(9):546-550. [Medline: 24137919]
35. El Morr C, Ritvo P, Ahmad F, Moineddin R, Team MVC. Effectiveness of an 8-week web-based mindfulness virtual community intervention for university students on symptoms of stress, anxiety, and depression: randomized controlled trial. JMIR Ment Health 2020;7(7):e18595 [FREE Full text] [doi: 10.2196/18595] [Medline: 32554380]
36. Tang W, Kreindler D. Supporting homework compliance in cognitive behavioural therapy: essential features of mobile apps. JMIR Ment Health 2017;4(2):e20 [FREE Full text] [doi: 10.2196/mental.5283] [Medline: 28596145]
37. Westra HA. Managing resistance in cognitive behavioural therapy: the application of motivational interviewing in mixed anxiety and depression. Cogn Behav Ther 2004;33(4):161-175. [doi: 10.1080/16506070410026426] [Medline: 15625790]

38. Huppert JD, Baker-Morissette SL. Beyond the manual: the insider's guide to panic control treatment. Cogn Behav Pract 2003;10(1):2-13. [doi: 10.1016/s1077-7229(03)80003-2]
39. Kavanagh M, Brouwer K, Lawrence P. Comparing the effectiveness of an 8-week and 12-week cognitive behavioural therapy group for bipolar affective disorder. Journal of Affective Disorders Reports 2021;6:100244 [FREE Full text] [doi: 10.1016/j.jadr.2021.100244]
40. Buhlmann U, Wilhelm S, Glaesmer H, Mewes R, Brähler E, Rief W. Perceived appearance-related teasing in body dysmorphic disorder: a population-based survey. International Journal of Cognitive Therapy 2011;4(4):342-348. [doi: 10.1521/ijct.2011.4.4.342]
41. Phillips KA, Stout RL. Associations in the longitudinal course of body dysmorphic disorder with major depression, obsessive-compulsive disorder, and social phobia. J Psychiatr Res 2006;40(4):360-369 [FREE Full text] [doi: 10.1016/j.jpsychires.2005.10.001] [Medline: 16309706]
42. Veale D, Anson M, Miles S, Pieta M, Costa A, Ellison N. Efficacy of cognitive behaviour therapy versus anxiety management for body dysmorphic disorder: a randomised controlled trial. Psychother Psychosom 2014;83(6):341-353 [FREE Full text] [doi: 10.1159/000360740] [Medline: 25323062]
43. Pope CG, Pope HG, Menard W, Fay C, Olivardia R, Phillips KA. Clinical features of muscle dysmorphia among males with body dysmorphic disorder. Body Image 2005;2(4):395-400 [FREE Full text] [doi: 10.1016/j.bodyim.2005.09.001] [Medline: 17075613]
44. Yamamotova A, Bulant J, Bocek V, Papezova H. Dissatisfaction with own body makes patients with eating disorders more sensitive to pain. J Pain Res 2017;10:1667-1675 [FREE Full text] [doi: 10.2147/JPR.S133425] [Medline: 28761371]
45. Osumi M, Imai R, Ueta K, Nobusako S, Morioka S. Negative body image associated with changes in the visual body appearance increases pain perception. PLoS One 2014;9(9):e107376 [FREE Full text] [doi: 10.1371/journal.pone.0107376] [Medline: 25210738]
46. Hilton L, Hempel S, Ewing BA, Apaydin E, Xenakis L, Newberry S, et al. Mindfulness meditation for chronic pain: systematic review and meta-analysis. Ann Behav Med 2017;51(2):199-213 [FREE Full text] [doi: 10.1007/s12160-016-9844-2] [Medline: 27658913]
47. Bowles NI, Davies JN, Van Dam NT. Dose-response relationship of reported lifetime meditation practice with mental health and wellbeing: a cross-sectional study. Mindfulness (N Y) 2022;13(10):2529-2546 [FREE Full text] [doi: 10.1007/s12671-022-01977-6] [Medline: 36193220]
48. Baer R, Gu J, Cavanagh K, Strauss C. Differential sensitivity of mindfulness questionnaires to change with treatment: a systematic review and meta-analysis. Psychol Assess 2019;31(10):1247-1263 [FREE Full text] [doi: 10.1037/pas0000744] [Medline: 31368738]

Abbreviations

BDD: body dysmorphic disorder
BDDQ: Body Dysmorphic Disorder Questionnaire
BDD-SS: Body Dysmorphic Disorder Symptom Scale
BPI: Brief Pain Inventory
CBT: cognitive behavioral therapy
CBT-M: mindfulness-based cognitive behavioral therapy
DSM-5: Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition)
GAD-7: Generalized Anxiety Disorder-7 Scale
HIPAA: Health Insurance Portability and Accountability Act
iCBT: internet-based cognitive behavioral therapy
LMM: linear mixed model
PHQ-9: Patient Health Questionnaire-9
RCT: randomized controlled trial

Edited by J Torous; submitted 10.12.23; peer-reviewed by E Bernstein; comments to author 08.01.24; revised version received 04.03.24; accepted 18.03.24; published 12.06.24.

Please cite as:
Kerry C, Mann P, Babaei N, Katz J, Pirbaglou M, Ritvo P
Web-Based Therapist-Guided Mindfulness-Based Cognitive Behavioral Therapy for Body Dysmorphic Disorder: Pilot Randomized Controlled Trial
JMIR Ment Health 2024;11:e55283
URL: https://mental.jmir.org/2024/1/e55283
doi:10.2196/55283
PMID:38865704

Original Paper

Acceptability and Engagement of a Smartphone-Delivered Interpretation Bias Intervention in a Sample of Black and Latinx Adults: Open Trial

IreLee Ferguson[1], BS; Grace George[2], MPA, PhD; Kevin O Narine[3], MA; Amari Turner[1], BS; Zelda McGhee[1], BS; Harris Bajwa[1], BA; Frances G Hart[1], BA; Sierra Carter[4], PhD; Courtney Beard[5], PhD

[1]Division of Depression and Anxiety Disorders, McLean Hospital, Belmont, MA, United States

[2]Department of Basic Neuroscience, McLean Hospital, Belmont, MA, United States

[3]Department of Clinical Psychology, William James College, Newton, MA, United States

[4]Department of Psychology, Georgia State University, Atlanta, GA, United States

[5]Department of Psychiatry, McLean Hospital/Harvard Medical School, Belmont, MA, United States

Corresponding Author:
Courtney Beard, PhD
Department of Psychiatry
McLean Hospital/Harvard Medical School
115 Mill St
Belmont, MA, 02478
United States
Phone: 1 617 855 3557
Email: cbeard@mclean.harvard.edu

Abstract

Background: Access to evidence-based interventions is urgently required, especially for individuals of minoritized identities who experience unique barriers to mental health care. Digital mental health interventions have the potential to increase accessibility. Previous pilot studies testing HabitWorks, a smartphone app providing an interpretation bias intervention, have found strong engagement and adherence for HabitWorks; however, previous trials' samples consisted of predominantly non-Hispanic, White individuals.

Objective: This study conducted an open trial of HabitWorks in a community sample of adults who identified as Black, Hispanic or Latinx, or both. This study aims to test safety, acceptability, and engagement with the HabitWorks app for Black and Latinx adults.

Methods: Black, Hispanic or Latinx adults (mean age 32.83, SD 11.06 y; 22/31, 71% women) who endorsed symptoms of anxiety or depression were asked to complete interpretation modification exercises via HabitWorks 3 times per week for 1 month. Interpretation bias and anxiety and depression symptoms were assessed at baseline and posttreatment assessments. Participants completed qualitative interviews to assess overall perceptions of HabitWorks.

Results: Of the 31 participants that downloaded the app, 27 (87%) used HabitWorks all 4 weeks. On average, participants completed 15.74 (SD 7.43) exercises out of the 12 prescribed, demonstrating high engagement. Acceptability ratings met all a priori benchmarks except for relevancy. Qualitative interviews also demonstrated high acceptability and few negative experiences. Significant improvements were found in interpretation style (t_{30}=2.29; *P*<.001), with a large effect size (Cohen *d*=1.53); anxiety symptoms (t_{30}=2.29; *P*=.03), with a small effect size (Cohen *d*=0.41); and depression symptoms (t_{30}=3.065; *P*=.005), with a medium effect size (Cohen *d*=0.55).

Conclusions: This study adds to the literature evaluating digital mental health interventions in Black and Latinx adults. Preliminary results further support a future controlled trial testing the effectiveness of HabitWorks as an intervention.

(JMIR Ment Health 2024;11:e56758) doi:10.2196/56758

KEYWORDS

interpretation bias; anxiety; depression; Black; Latinx; smartphone; mobile phone

Introduction

Background

The tendency to resolve ambiguity in a threatening or negative manner, that is, interpretation bias, has been associated with most emotional disorders [1]. HabitWorks is a personalized, transdiagnostic, smartphone-delivered intervention that targets this type of interpretation bias [2]. The primary feature of HabitWorks is a word-sentence association exercise that reinforces users for making benign interpretations and rejecting negative interpretations of ambiguous situations through repeated practice [3]. HabitWorks also includes several features expected to enhance user engagement, such as personalization of the ambiguous situations presented, personally scheduled notifications, performance feedback, level progression, in-app mood monitoring, and a diary feature. In addition, the exercises require only 5 minutes and can be completed at the user's convenience. HabitWorks showed good feasibility and acceptability in a pilot trial of adults receiving acute psychiatric care [2] and parents youth with anxiety [4], as well as excellent adherence and engagement [5].

This study aims to obtain feasibility and acceptability data about HabitWorks in a sample of adults identifying as Black, Hispanic, or Latinx. The rationale for conducting this pilot study was 2-fold. First, although digital mental health interventions (DMHIs) can overcome the most common barriers to receiving traditional mental health treatment (eg, availability of a provider, cost, transportation, and stigma) [6-9], the DMHI field has not yet realized its potential to increase access to evidence-based interventions. DMHIs are not reaching people who have typically not accessed mental health treatment; most data to date come from White women—the same demographic that is already best served by existing treatments [10,11]. For example, Ellis et al [12] recently reported that 97% of studies in a review of internet-delivered cognitive behavioral therapy either did not include participants of minoritized races and ethnicities or did not report on participant ethnoracial identities. This was further supported in another systematic review of internet-delivered cognitive behavioral therapy trials that found most studies did not report on race and ethnicity or included predominantly non-Hispanic White samples [13]. In a systematic review of culturally adapted DMHIs, only 4 studies examined Black American participants, and none of the studies examined anxiety in the Hispanic or Latinx communities [12]. DMHI developers and researchers must be more intentional to avoid exacerbating the health and access disparities they seek to address.

Second, it is unclear how minoritized groups respond to interpretation bias interventions. Evidence-based treatments may have harmful effects if applied in a one-size-fits-all manner and without attending to potential sociocultural influences. For example, it is generally recommended that therapists not apply cognitive restructuring for interpretations of racist experiences and, instead, focus on the internalized beliefs from these experiences [14]. Asking a client to question their interpretations about an identity-related experience could be extremely invalidating, exasperate distress, and contribute to mistrust of health care providers. Negative interpretations of ambiguous situations can be adaptive in minoritized groups (eg, Black individuals experiencing police brutality), and it is invalidating to suggest reappraising such interpretations. It is not yet known whether similar approaches that focus on cognitive reappraisal, such as interpretation bias interventions, are experienced as invalidating by minoritized individuals, even when not directly focused on racism-related interpretations.

During the development of HabitWorks, we attended to sociocultural identities in several ways. Although HabitWorks presents common situations related to cognitive distortions in anxiety and depression, these same situations could also bring up thoughts related to discrimination. Thus, we first completed a review of the Word Sentence Association Paradigm (WSAP) stimuli to minimize the risk of ambiguous situations bringing up experiences related to microaggressions or discrimination [15]. Two research assistants who identify with multiple minoritized identities examined 800 word-sentence pairs from various versions of the WSAP using the ADDRESSING framework [16]. ADDRESSING is an acronym for sociocultural identities including age, disability, diagnosis status, religion or spirituality, ethnicity or race, sexuality, socioeconomic status, Indigenous heritage, national origin, and gender. Each research assistant selected 1 sociocultural identity at a time and identified word-sentence pairs that strongly corresponded to potentially negative experiences related to that identity. Next, potentially problematic situations were either removed or modified to reduce association with identity-related negative experiences. Third, we collected qualitative feedback from participants via self-report exit questionnaires in all our pilot trials. In all studies, we specifically asked about any negative effects of the WSAP as well as cultural acceptability (eg, "Do you think your friends or family would want to use an app like this?"). Participants did not share any negative experiences related to the WSAP content related to identity-related experiences. Finally, we created a "HabitWorks in context" document in the app that describes the difference between anxiety-related negative interpretations and identity-related interpretations. For example, a person of a minoritized race or ethnicity may draw conclusions about the situation—"You have a job interview"—because of thoughts related to performance anxiety ("I'll appear nervous and they will think I'm weak") and racism ("They are going to judge me differently because of my race"). To provide participants with more context, we state that HabitWorks is designed to help people re-evaluate thoughts related to anxiety and depression. It is *not* the intention of HabitWorks to reframe interpretations related to discrimination because it is unhelpful, culturally insensitive, and invalidating to reframe thoughts related to racism, sexism, homophobia, ableism, or any other form of discrimination experienced.

Because modifying interpretations around discrimination-related experiences could be invalidating, we instead provided an extensive catalog of general and identity-specific resources to provide support in finding a therapist, mental health organizations, crisis lines, financial resources, identifying supportive social spaces, and ways to cope with discrimination.

While we hope these initial efforts resulted in an intervention that is safe and acceptable to minoritized groups, we currently do not have sufficient data to determine this. In our prior trials, we did not specifically ask about identity-related experiences in qualitative interviews. In addition, the interviewers were all White, which may impact participants' comfort in disclosing their experiences. In addition, our small pilot studies did not have a sufficient sample size of participants identifying as Black or Latinx; across the 3 pilot trials, only 2 participants identified as Black and only 3 identified as Hispanic or Latinx. Thus, it is currently unclear how a history of microaggressions and discrimination affect how users experience the HabitWorks app.

Objectives

This study aims to test safety, acceptability, and engagement with the HabitWorks app in a sample of Black and Latinx adults. On the basis of our prior pilot work, we hypothesized that HabitWorks would be feasible to deliver to a community sample of Black and Latinx adults and that participants would adhere to the recommended dosage. However, we were unsure whether HabitWorks would be safe and acceptable to Black and Latinx participants. To answer this question, we examined the rate of adverse outcomes and conducted qualitative interviews after participants used the app for 1 month. To address the limitations of our prior work, qualitative interviews were all conducted by a person of a minoritized racial or ethnic identity (although not necessarily Black or Latinx). In addition, interviews specifically probed whether HabitWorks brought up experiences of discrimination or was invalidating in any way. We compared safety, feasibility, acceptability, interpretation bias, and anxiety outcomes to a priori benchmarks selected based on prior studies [2,4,17,18]. These prior studies selected benchmarks based on previous interpretation bias modification studies in clinical settings and clinical judgment based on what would be clinically useful for a low-intensity smartphone intervention [4]. Given that this was a similar pilot trial, we used these same benchmarks. Benchmarks were selected before reviewing current data.

Methods

Participants

Participants were English-speaking adults aged ≥18 years and residing in the United States, who identified as Black, Hispanic, or Latinx. Additional inclusion criteria included access to an iPhone and at least mild anxiety symptoms (Generalized Anxiety Disorder [GAD]-2 score ≥3) or depression symptoms (Patient Health Questionnaire [PHQ]-2 score ≥3). Exclusion criteria included active mania or psychosis that would inhibit informed consent or completion of study procedures.

The study was posted for 4 months (November 2022 to February 2023) on Mass General Brigham's Rally website, a web-platform that advertises clinical research studies to the public. Participants were informed that the study was testing a smartphone app designed to reduce anxiety and depression. A total of 31 participants were included in analyses (Figure 1). Participants identified as primarily women (22/31, 71%), with their ages ranging from 18 to 61 (mean 32.83, SD 11.06) years, and most participants completed ≥4 years of college (18/31, 58%) and were employed full time (18/31, 58%; Table 1). Participants identified as Black (15/31, 48%), Latinx (13/31, 42%), and both Black and Latinx (3/31, 10%). Most participants resided in Massachusetts (24/31, 77%); however, we also enrolled 1 participant from each of the following states: Illinois, New York, Nebraska, Connecticut, Rhode Island, Tennessee, and South Carolina.

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram. Ineligible subgroups at the screening stage are not mutually exclusive. GAD-7: Generalized Anxiety Disorder; PHQ-8: Patient Health Questionnaire.

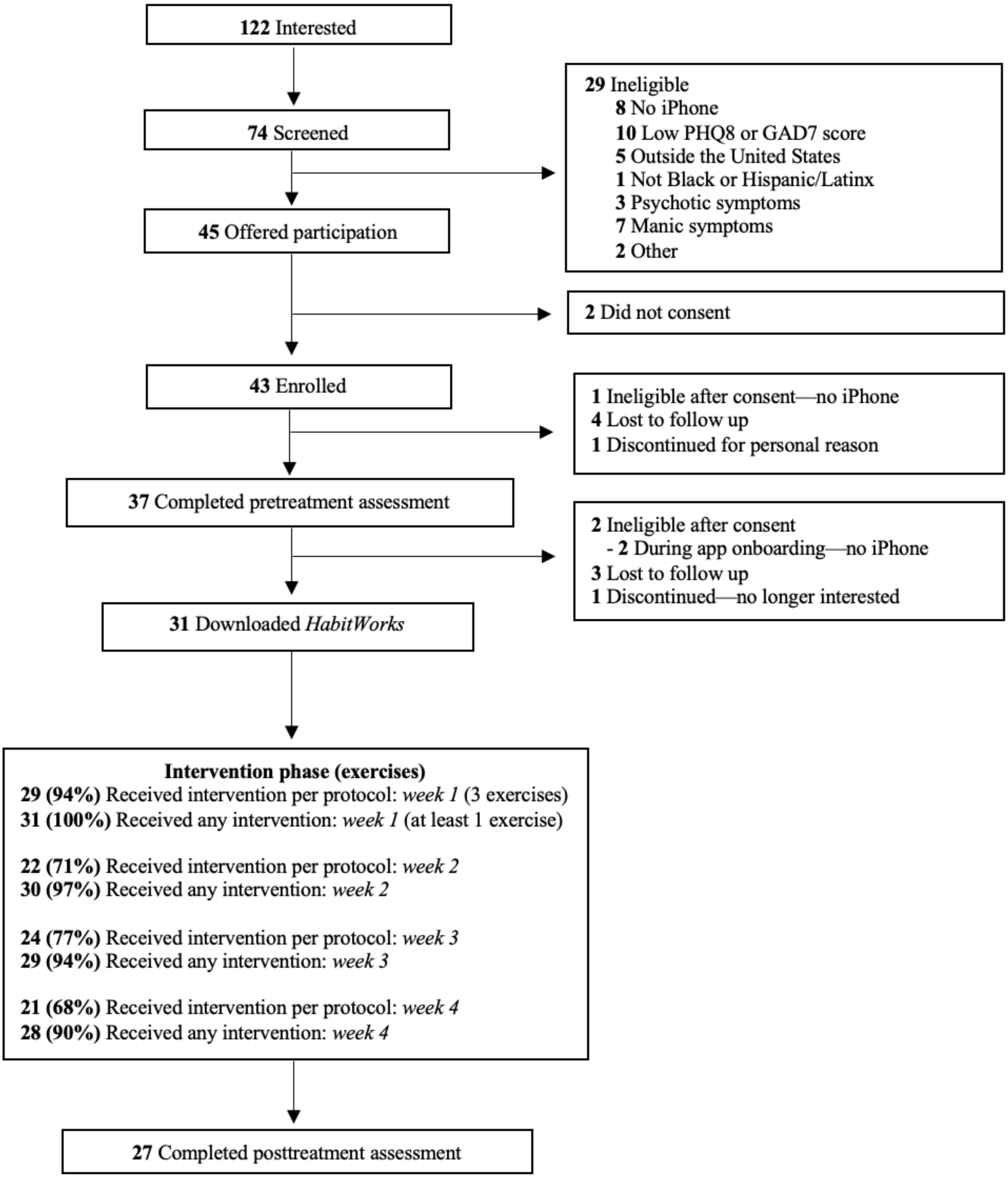


Table 1. Demographic characteristics of participants (N=31).

Characteristic	Participants
Age (y), mean (SD; range)	32.83 (11.06; 18-61)
Sex at birth, n (%)	
Female	23 (74)
Male	8 (26)
Intersex	0 (0)
Prefer not to answer	0 (0)
Gender, n (%)	
Cisgender, woman	22 (71)
Cisgender, man	7 (23)
Transgender, man	1 (3)
Transgender, woman	0 (0)
Nonbinary	0 (0)
Genderqueer	0 (0)
Agender	0 (0)
Not listed	0 (0)
Prefer not to answer	1 (3)
Sexual orientation, n (%)	
Asexual	0 (0)
Bisexual	3 (10)
Gay or lesbian	3 (10)
Heterosexual or straight	23 (74)
Pansexual	1 (3)
Queer	1 (3)
Not listed	0 (0)
Prefer not to answer	0 (0)
Education level, n (%)	
8th grade or less	0 (0)
Some high school	0 (0)
High school graduate or General Educational Development credential	1 (3)
Some college or associate's degree or trade school	12 (39)
4-year college graduate	10 (32)
Postcollege education	8 (26)
Prefer not to answer	0 (0)
Employment, n (%)	
Student	6 (19)
Student and employed part time	1 (3)
Not employed due to disability	1 (3)
Not employed—seeking job	2 (6)
Employed part time	2 (6)
Employed full time	18 (58)
Employed full time and part-time	1 (3)
Homemaker	0 (0)

Characteristic	Participants
Retired	0 (0)
Race and ethnicity, n (%)	
American Indian or Alaska Native	0 (0)
American Indian or Alaska Native, White, Hispanic or Latinx	1 (3)
Asian	0 (0)
Black or African American	12 (39)
Black or African American and Hispanic or Latinx	6 (19)
Black or African American and White	2 (6)
Hispanic or Latinx	5 (16)
Hispanic or Latinx, Native Hawaiian or Pacific Islander, and White	1 (3)
Hispanic or Latinx and White	4 (13)
Middle Eastern or North African	0 (0)
White	0 (0)
Do not know or unsure	0 (0)
Not listed	0 (0)
Prefer not to answer	0 (0)

Measures

Outcomes for the pilot study included safety, feasibility, acceptability, target engagement (interpretation bias), and anxiety and depression symptoms.

Safety

We tracked the number of participants who experienced any adverse outcomes, as well as clinical deterioration, defined as an increase of ≥5 points from the previous assessment on the GAD-7 or PHQ-8 during the 1-month treatment period.

Adherence

We prescribed 3 exercises per week for 4 weeks. We obtained exercise completion data from the app user statistics. In addition to the interpretation exercises, HabitWorks prompted participants to complete a self-report assessment of anxiety and depression symptoms and a diary entry weekly. Participants could also complete assessments of their symptoms and diary entries at any point during the treatment phase. We calculated the number of each completed.

Acceptability

First, we administered a 10-item self-report measure of participant satisfaction previously used in similar studies [2,4,18]. Administered during the posttreatment assessment time point, this exit questionnaire prompted participants to rate how helpful, relevant, user-friendly, and satisfying they found HabitWorks after 1 month of use on a Likert scale from 1 (*completely disagree*) to 7 (*completely agree*). In addition to the 5 Likert scale response items, the 5 remaining items had an open-ended response format (Multimedia Appendix 1).

Second, participants were asked to complete a semistructured qualitative interview following completion of the 1-month treatment phase via telephone. Interviewers were members of the research team (KON, GG, and HB). Before conducting interviews, all interviewers underwent comprehensive training, including an in-depth reading of relevant articles about qualitative interviews and observing a qualitative interview led by the principal investigator for a related study. The principal investigator developed the initial interview guide, which was then reviewed by experts in mental health in Black and Latinx populations with no further edits suggested. The interviewers then reviewed the guide and suggested some minor wording edits. Interview prompts asked about general impressions of HabitWorks; negative experiences due to the app, including whether it brought up thoughts related to discrimination; the effectiveness of the program; and suggestions for improvement. Interviews (n=20) ranged in duration from 7.68 to 60.65 (mean 19.38, SD 12.18) minutes.

We used a general inductive approach [19], a simple and straightforward procedure that is well suited to summarizing focused evaluation questions such as those in this study. Within the context of the general inductive approach, we applied a rapid analysis procedure, which is preferred for intervention development, implementation, and time-sensitive analyses needed to inform future waves of data collection [20]. Rapid analysis approaches have demonstrated comparable rigor to traditional qualitative analyses with less time and cost [21].

Four members of our team conducted the qualitative analysis (KON, FGH, ZM, and AT). We first created a template representing the domains of inquiry in the interview guide. Each coder independently read each transcript and recorded themes in the template. Each coder then reviewed the template several times to identify an initial set of themes across transcripts. They then met together to discuss their independently derived themes to reduce overlap and redundancy. They then reread transcripts or listened to interview recordings to ensure the existing categories captured all participant data relevant to our study aims as well as to identify contradictory points of view. They met again to discuss any revisions to the themes and decide on the ultimate structure of the data. Finally, the team met to discuss and reach consensus about the themes and structure.

Usability

We used the 10-item self-report System Usability Scale to examine participant ratings of the usability of HabitWorks [22]. Administered at the posttreatment assessment time point, the System Usability Scale asked participants to rate how usable (eg, “cumbersome,” “integrated,” and “easy”) they found the intervention, using a Likert scale ranging from 1 (*strongly disagree)* to 5 (*strongly agree)*. Technologies scoring >68 are classified as above average regarding usability [23]. This measure has exhibited excellent reliability and validity [23] and excellent internal consistency in the current sample (Cronbach α=0.94).

Credibility and Expectancy

After their first session of HabitWorks, participants were asked to complete the Credibility and Expectancy Questionnaire (CEQ) [24]. The CEQ is a widely used 6-item self-report measure with 2 factors: credibility (items 1 to 3, eg, “How logical does the therapy offered to you seem?”) and expectancy (items 4 to 6, eg, “How much improvement in your symptoms do you really feel will occur?”) designed to assess participants’ thoughts and feelings toward the intervention’s ability to reduce symptoms (ie, stress and anxiety). A rating scale of 1 (not at all) to 9 (completely), or 0% to 100%, is used for each question, depending on question content. The CEQ has demonstrated good test-retest reliability, adequate validity, and good internal consistency [24,25], including in the current sample (Cronbach α=0.84 for credibility and 0.85 for expectancy).

Interpretation Bias

As a manipulation check, we included an assessment version of the WSAP [26] at baseline and after treatment. The WSAP is a commonly used measure of interpretation bias with good internal consistency and test-retest reliability across clinical and nonclinical populations [3]. In the assessment version, no feedback is provided about accuracy of responses. The 50 word-sentence pairs were drawn from a previous study of interpretation bias and can be found on the internet [27,28]. Thus, stimuli were not personalized in the assessment version, meaning that all participants saw the same transdiagnostic set of 50 word-sentence pairs. Of note, although there is a chance that some word-sentence pairs were seen in both the intervention and assessment for some participants, overall, the WSAP assessment presented distinct stimuli from the intervention. Participants were asked to decide if a word and sentence were related. Half of the ambiguous sentences were paired with a word reflecting the negative or threat interpretation, and half were paired with a word reflecting a benign interpretation. For example, the word “bored” would reflect the negative or threat interpretation, while the word “tired” would reflect the benign interpretation for the sentence “Someone yawns while you’re talking.” Responses were coded as accurate when participants endorsed “yes—related” to benign trials and “no—not related” to negative trials. We calculated an interpretation bias score by averaging each participant’s accuracy score across the 50 trials; higher accuracy scores reflect less negative interpretation bias. As the WSAP assessment is a similar task to the intervention, we emphasize that this measure served as more of a manipulation check (ie, did participants learn the “correct” contingencies in the intervention and did this generalize to an assessment version of the task?) than a test of change in general interpretative style. The WSAP was administered via a customized link sent to each participant’s email at each time point.

Anxiety

Participants completed the GAD-7 item scale during in-app mood check-ins and at baseline and posttreatment assessment time points. The GAD-7 is a widely used 7-item self-report questionnaire that assesses symptoms of anxiety on a Likert scale ranging from 0 (*not at all*) to 3 (*nearly every day*), with higher scores indicating greater anxiety severity [29]. The GAD-7 has demonstrated good reliability and construct validity [30,31] and has been used as a transdiagnostic measure of anxiety in various clinical settings [30,32].

Depression

Participants completed the PHQ-8 item scale during in-app mood check-ins and at baseline and posttreatment assessment time points. The PHQ-9 is a widely used self-report questionnaire that assesses depression symptoms on a Likert scale ranging from 0 (*not at all*) to 3 (*nearly every day*), with higher scores indicating greater depression severity [33]. The PHQ-8 includes all items from the PHQ-9 except the ninth item assessing thoughts of death and harming oneself. The PHQ-9 has demonstrated good reliability and construct validity [34], and dropping item 9 of this questionnaire is common practice and has been shown not to affect reliability and validity [35].

Ethical Considerations

This study was approved by the Mass General Brigham Institutional Review Board (2022P001752). All participants provided informed consent at the beginning of the study and were able to opt out at any time. Privacy and confidentiality were maintained by deidentifying all data. Additionally, the HabitWorks app was designed to minimize risks of breach of privacy, confidentiality and data security. Participants were compensated up to US $100 for their participation. They received US $20 for completing the baseline and posttreatment assessments, US $30 for cellular data use related to the app, and US $30 for completing the feedback interview. Participants were not compensated for using the HabitWorks app, including the interpretation exercises.

Procedure

Study staff emailed potential participants a link to a brief screening survey to assess initial eligibility via REDCap (Research Electronic Data Capture; Vanderbilt University), a secure, web-based application designed to capture data for research studies [36]. If eligible, participants provided informed consent via REDCap. Once consented, participants were asked to complete a pretreatment assessment via REDCap. Following the pretreatment assessment, participants met with the research assistant via Zoom (Zoom Video Communications) to complete a 30-minute orientation session. During this session, the participants downloaded the HabitWorks app, watched instructional videos explaining the rationale of and how to use HabitWorks features, completed the HabitWorks personalization checklists, scheduled notifications to complete WSAP exercises, and practiced completing the first symptom survey and WSAP exercise in the app.

Participants were asked to use the HabitWorks smartphone app for 1 month. During this month, we asked them to complete the WSAP exercises 3 times per week, a symptom survey weekly, and a Habit diary entry weekly. The Habit diary encouraged participants to write about their progress by prompting them to write about instances when they jumped to a conclusion or noticed a change in their thinking behavior. Refer to the study by Beard et al [4] for an in-depth description of app features, and refer to the study by Beard et al [2] for a detailed description of the app development process.

The WSAP exercises took approximately 5 minutes to complete and included feedback that reinforced benign interpretations after each trial. Specifically, participants saw “Correct!” when they endorsed a benign interpretation or rejected a negative interpretation. They saw “Try again!” when they rejected a benign interpretation or endorsed a negative interpretation. Participants completed 50 trials during each scheduled exercise and 30 trials during any user-initiated bonus exercise. HabitWorks presented participants with personally relevant ambiguous situations from a pool of 714 potential word-sentence pairs, based on their responses to personalization checklists. Refer to the study by Beard et al [4] for more information on personalization.

A research assistant monitored adherence and emailed participants weekly. These weekly emails summarized the participant’s progress that week and offered encouragement. If the intervention was not completed that week, the email provided support related to low motivation and technological issues. In addition, the research assistant was available to meet with participants via Zoom or phone to provide technical support, although no participants asked to meet. After 1 month of app use, participants were sent a posttreatment assessment via REDCap. Once the posttreatment assessment was completed, participants were invited to complete a feedback interview.

Results

A Priori Benchmarks

A priori benchmarks and obtained data are presented in Table 2.

Table 2. Benchmarks.

Outcome	Target	Actual
Safety	No clinical deterioration or increase in suicidal ideation possibly related to study participation	0 participants reported worsening due to study participation
Feasibility (% eligible provided consent)	50% of eligible participants provide consent	43/45, 95%
Adherence to WSAP[a] exercises	50% complete sessions 3 times per week during the treatment phase	Week 1: 29/31, 93%; week 2: 29/31, 71%; week 3: 24/31, 77%; and week 4: 21/31, 68%
Adherence to study assessments	75% complete assessments	Baseline: 37/43, 86% and posttreatment: 27/31, 87%
Credibility	At least moderate credibility (5 on a scale of 1 to 9 for the CEQ[b] item “How logical?”)	Mean 6.19, SD 2.15
Expectancy	At least moderate expectancy (>50% out of 100% on the CEQ item “How much improvement?”)	Mean 45.56%, SD 24.2%
Satisfaction and acceptability	Average rating of “slightly agree” (5 on a scale of 1 to 7) on the exit questionnaire	Helpfulness: mean 5.19, SD 1.44; relevance: mean 4.78, SD 1.63; easy to use: mean 6.15, SD 1.23; satisfaction: mean 5.78, SD 1.50; and recommendation: mean 5.46, SD 1.63
Interpretation bias manipulation check	75% in healthy range (≥70% accuracy) on WSAP (note that 1/31, 3% were already in the healthy range at baseline)	Posttreatment: 23/31, 74%
Symptom reduction	50% in the “none to minimal” range on GAD-7[c]; effect size Cohen d >0.2	Posttreatment: 9/31, 29%; Cohen d=0.41

[a]WSAP: Word Sentence Association Paradigm.

[b]CEQ: Credibility and Expectancy Questionnaire.

[c]GAD-7: Generalized Anxiety Disorder-7.

Safety

During the 4-week treatment period, no adverse events were reported. Overall, 16% (5/31) of the participants were flagged for clinical deterioration; 4 (80%) of these 5 participants demonstrated clinical deterioration once, while 1 (20%) participant demonstrated it twice. None of these were deemed related to the HabitWorks intervention or research procedures.

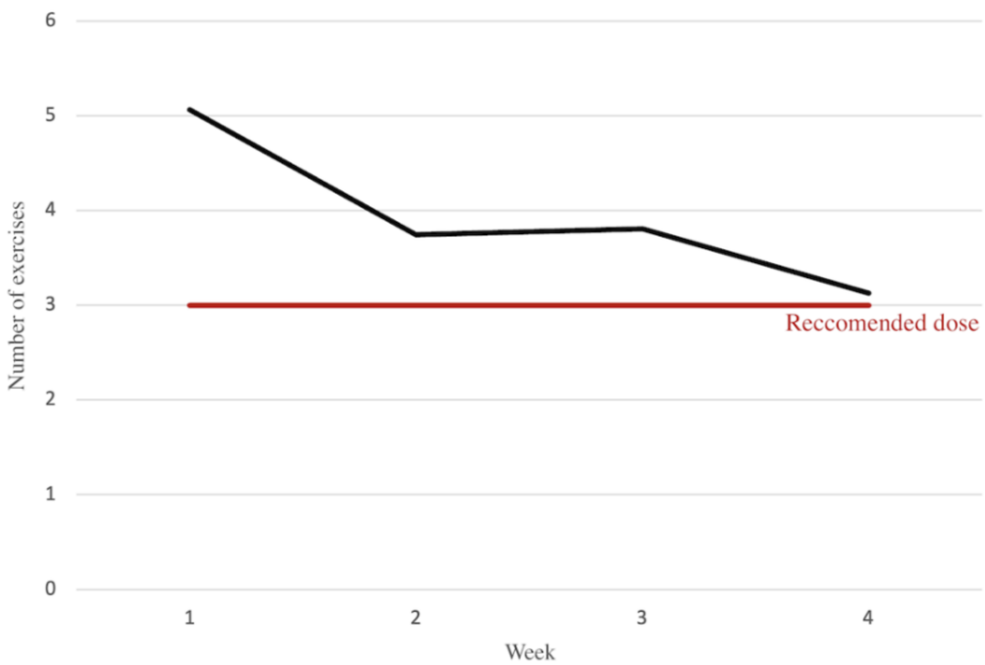
Feasibility

Of the 45 participants eligible for the study, 43 (95%) enrolled. Of the 43 enrolled participants, 31 (72%) downloaded the app. All recruitment, assessment, and intervention procedures were easily completed entirely remotely. Participants did not report any substantial issues with using the app.

Adherence

On average, participants completed 15.74 (SD 7.43) exercises in total out of the 12 prescribed over the 4 weeks. Of the 31 participants who onboarded to the app, 17 (55%) adhered to the prescribed intervention (3 exercises/wk) all 4 weeks, and 27 (87%) participants used the intervention all 4 weeks (Figure 2). Approximately two-thirds (21/31, 68%) of the participants completed at least 1 symptom survey every week of their participation. On average, participants completed 6.13 (SD 3.35) symptom surveys by the end of the 4 weeks. Overall, 29% (9/31) of the participants completed a diary entry each week, and 74% (23/31) of the participants completed at least 1 entry during their participation. On average, participants completed 3.55 (SD 3.77) entries out of the 4 prompted with highest completion rates occurring at the end of week 2.

Figure 2. Average number of Word Sentence Association Paradigm (WSAP) exercises completed in the treatment phase (N=31).



Acceptability

Acceptability ratings met our a priori benchmarks for satisfaction, perceived helpfulness, user-friendliness, and willingness to recommend the app to a friend (Table 2). A priori benchmarks for relevancy were not met.

Data from qualitative interviews also revealed high acceptability and few negative experiences (Multimedia Appendix 1). Regarding participants' general experience with HabitWorks, themes such as easy to use, intuitive, quick, and efficient emerged. People commented on several aspects of the app that they found helpful, such as the notifications, mood check-ins, exercises, and progress tracking graphs. Far fewer people discussed unhelpful aspects of HabitWorks, and these centered on the exercises feeling tedious or repetitive, the words flashing too quickly, and some instructional content being confusing. When asked about any negative experiences, most people denied any; a few people reported feeling briefly frustrated when the app told them they were incorrect about a situation, or they did not score as high as they wanted. When asked about any perceived changes since using HabitWorks, participants observed changes in their cognition, such as reappraising situations and not jumping to conclusions. Participants discussed increased awareness of their thought patterns and mood. They also noted changes in their emotions, such as feeling less anxious.

We were very interested in the personal relevance of the situations. Overall, participants reported that the situations presented in the exercises were frequently experienced and relatable to their daily life. However, a few participants (5/19, 26%) also noted specific situations that they did not find personally relevant. We also specifically asked participants whether the situations in the WSAP exercises triggered any thoughts of discrimination, racism, or microaggressions. Participants overwhelmingly did not think the WSAP brought up experiences of discrimination and denied any race-related distress while completing it. However, 2 participants did report that some of the situations could be related to xenophobia, social exclusion, and work-related discrimination.

Usability

On average, participants reported that HabitWorks had good usability (mean 67.60, SD 27.96; range 10.00-100.00), just approaching the cutoff of 68 as an indicator of above average usability [23].

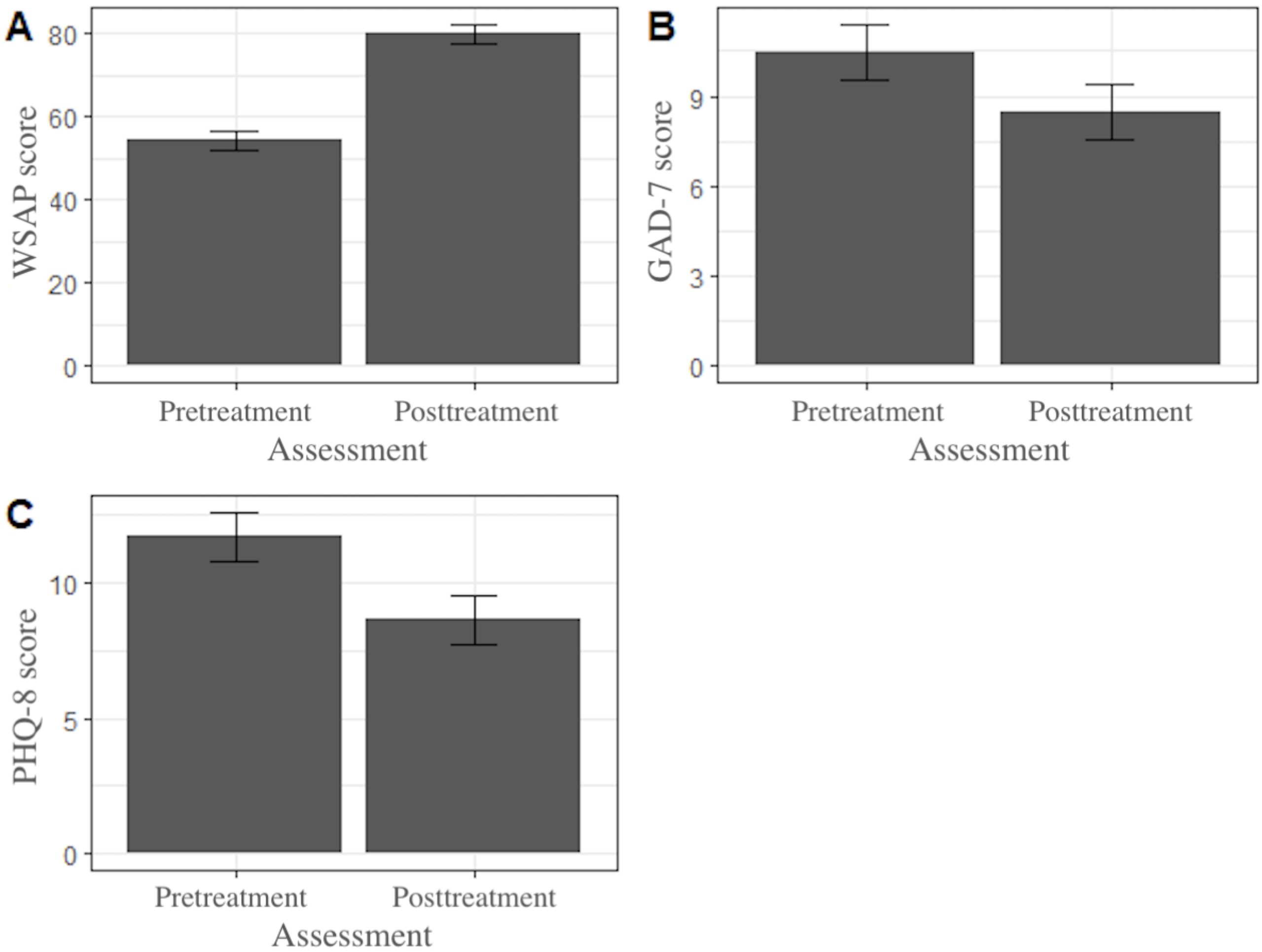
Credibility and Expectancy

On average, participants reported that HabitWorks was moderately credible (mean 6.13, SD 2.19; range 1 to 9). On average, participants reported that their expectation for percent improvement in their mental health symptoms was 43% (SD 22.77%; range 10%-100%) and did not meet our a priori benchmark.

Interpretation Bias

Five participants did not complete the WSAP at the posttreatment time point. Given the small pilot sample, we carried forward these participants' baseline scores for this time point. Our a priori benchmark for the assessment version of the WSAP was met (Table 2). At baseline, 3% (1/31) exceeded the healthy score threshold for accuracy (≥70% accuracy). By the posttreatment assessment, 74% (23/31) of the participants' interpretation bias scores on the WSAP task were in the healthy range. A paired samples 2-tailed *t* test revealed significant improvement in accuracy (t_{30}=2.29; *P*<.001; Figure 3). The effect size was large with a Cohen *d* of 1.53.

Figure 3. (A) Word Sentence Association Paradigm (WSAP), (B) Generalized Anxiety Disorder-7 (GAD-7), and (C) Patient Health Questionnaire-8 (PHQ-8) scores (N=31).



Anxiety and Depression Severity

Overall, 100% (31/31) of the participants completed the PHQ-8 and GAD-7 at baseline, and 87% (27/31) of the participants completed the PHQ-8 and GAD-7 at the posttreatment time point. We carried forward the last reported score on the PHQ-8 and GAD-7 for the 4 participants missing their postassessment. At baseline, the average GAD-7 score was in the "moderate" severity range (mean 10.47, SD 4.69). The average GAD-7 score reduced to the "mild" severity range at the posttreatment time point (mean 8.48, SD 5.66). A paired samples *t* test revealed a significant decrease in anxiety symptom severity (t_{30}=2.29; *P*=.03; Figure 3). The effect size was small with a Cohen *d* of 0.41. At baseline, the average PHQ-8 score was in the "moderate" severity range (mean 11.71, SD 4.53). The average PHQ-8 score reduced to the "mild" severity range (mean 8.65, SD 5.64). A paired samples *t* test revealed a decrease in depression symptom severity (t_{30}=3.065; *P*=.005; Figure 3). The effect size was medium with a Cohen *d* of 0.55.

Suggestions for Improvement

Suggestions to improve HabitWorks included ideas to improve exercises such as increasing the number of situations or adding the ability to change difficulty level. In addition, participants felt exercises could be better understood by giving participants an opportunity to explain their answers or offering explanations for incorrect answers. There were some participants (6/19, 32%) looking for enhancement of esthetic and (5/19, 26%)

improvements to the Habit diary. Participants also mentioned wanting app features that offered more support such as human support through the app (eg, coach), additional strategies for learning healthy habits and coping skills, and links to other resources. Overall, participants expressed satisfaction with the app and wanting to share it with others (Table 2).

Discussion

Summary

We tested the preliminary safety, feasibility, and acceptability of a smartphone-delivered interpretation bias intervention in a sample of adults who identify as Black, Hispanic or Latinx, or both. This study addresses a critical gap, given that prior studies of DMHIs and HabitWorks, specifically, have not had sufficient representation of Black or Latinx participants. HabitWorks met almost all a priori benchmarks for safety, feasibility, adherence, and acceptability.

Safety, Feasibility, and Acceptability

Overall, the evidence of safety, feasibility, and acceptability of HabitWorks for Black or Latinx adults was found to be strong and consistent with previous studies of HabitWorks comprising primarily non-Hispanic White participants [4,5]. No clinical deterioration occurred due to the use of the app. The a priori benchmark for the credibility of HabitWorks was met. Satisfactions ratings were high; participants found HabitWorks to be helpful and easy to use and would recommend the app to a friend. The a priori benchmarks not met were treatment outcome expectancies, as found in previous trials testing HabitWorks [4,5]. We suspect that expected improvement in mental health symptoms was low due to skepticism around DMHIs, as treatment-seeking individuals often initially feel that DMHIs are inferior in comparison to other mental health care options such as counseling and medications [37-40]. Future research should consider how education on the efficacy of DMHIs in comparison to other treatments could boost expectancy. For instance, we could educate users on the high satisfaction ratings of HabitWorks during the onboarding process to try to address low expectancy.

Despite personalizing the situations presented, our benchmark for personal relevancy was not met, and data from qualitative interviews also revealed that some participants did not find all the scenarios to be personally relevant. Engagement and adherence were good despite this, but suboptimal relevancy might explain the decrease in use of the app after the first week. This could partially be due to the limited variety in word-sentence pairs, which some participants mentioned during qualitative interviews. To achieve our relevance benchmark, we may need to increase the number of word-sentence pairs presented and provide more personalization questions to narrow the scope of scenarios provided within the exercises. It is also possible that incorporating participant-generated situations would improve relevancy as well as engagement. Despite the room for improvement in personalization of content, overall satisfaction was good.

It is also important to note that although expectancy for HabitWorks was low at baseline, overall satisfaction with the app across several domains was high at the posttreatment time point, as in previous studies of HabitWorks [4,5]. This was supported by qualitative data, which found that some participants wanted to continue to use the app beyond study participation. Therefore, experience using the app seemed to overcome initial skepticism and led to acceptability of the app allowing for higher engagement and adherence rates.

One of our main interests in this study was determining whether the intervention content was culturally sensitive and whether the exercises brought up experiences of racism or discrimination. Although we attended to sociocultural identities when developing HabitWorks, it was not culturally adapted for Black or Latinx communities. Despite the lack of specific cultural adaptation, HabitWorks demonstrated comparable acceptability to other culturally adapted DMHIs [12]. Overall, participants did not feel that HabitWorks asked them to reframe experiences of discrimination or caused them any race-related distress. When asked, only 2 participants reported that some situations brought up thoughts about xenophobia, feeling excluded, and work-related discrimination. Distress related to these situations was not reported and could have been attenuated by the previously mentioned app features such as the “HabitWorks in context” document and the ability for participants to remove word-sentence pairs they did not want to see again. Relatedly, as noted previously, we intentionally did not include any situations in HabitWorks that that could bring up identity-related stress. Thus, the current version of HabitWorks was not designed to address uncertainty and stress due to identity-related experiences. Because some minoritized individuals may have daily negative experiences related to their identity, there are many important research questions for future work. For example, it is possible that shifting participants’ general interpretive style might generalize to identity-relevant situations and provide some anxiety reduction in identity-related situations. Thus, future studies should include measures of identity-related stress. It is also possible that entirely different types of interventions are better suited to help people cope with ambiguous situations that bring on identity-based stress; thus, future studies might compare the effects of interpretation bias interventions combined with specific identity-based stress coping interventions. Overall, these findings suggest that HabitWorks shows promise as a culturally appropriate DMHI for Black and Latinx adults.

Adherence and Engagement

We asked participants to complete 3 WSAP exercises per week for 1 month. On average, participants exceeded this and completed 15.74 WSAP exercises in the 1-month period, with 54% (17/31) of the participants completing the prescribed in intervention all 4 weeks. Most people stop using mental health apps after 2 weeks [41]; however, HabitWorks use remained high even in week 4. As previously noted, some participants even expressed a desire to continue using HabitWorks after completing the study. Moreover, participants engaged with the other features of the app. On average, participants completed 6.13 symptom surveys in a 1-month period, with most participants completing at least 1 symptom survey every week, demonstrating high engagement with the symptom survey feature. Habit diary completion rates fell just short of the recommended dose of 1 per week. Qualitative data revealed mixed opinions of the Habit diary, as some participants found it helpful, while others did not and offered suggestions for its improvement (Multimedia Appendix 1). Importantly, we did not compensate participants for using the app, and most participants enrolled immediately before the winter holiday season. Thus, we find these adherence rates encouraging, especially considering the typical challenges of engaging users, particularly during this time of year. However, compensation for cellular data use and completion of the feedback interview may have acted as motivation for participants to complete the study.

A meta-analysis revealed that culturally adapted DMHIs have an average attrition rate of 42%, despite reporting high satisfaction [12]. In contrast, this pilot trial demonstrated high satisfaction and a low attrition rate, with 67% (21/31) of the participants meeting the recommended treatment dose by the end of the study. These adherence rates provide strong, preliminary evidence that this intervention has the potential to be an engaging treatment for anxiety and depression across races and ethnicities.

Interpretation Bias

We examined interpretation bias before and after treatment. On average, participants' interpretation bias significantly improved from baseline to the posttreatment time point with a large effect size. By the end of treatment, 74% (23/31) of the participants had moved in the healthy range, which is slightly lower than our a priori benchmark and what was found in other studies testing this type of intervention [2,4,18]. These findings are promising in that they demonstrate an ability for participants to learn from the correctional feedback provided by the WSAP exercises and apply it to new situations, but without additional measures of interpretation bias, we cannot demonstrate an improvement in interpretation bias that translates beyond the task itself. However, we can draw upon participants' perceived cognitive changes. In qualitative interviews, participants reported that HabitWorks increased their awareness of jumping to conclusions, and their ability to reappraise situations had improved with app use.

Anxiety and Depression Severity

We also examined anxiety and depression symptom severity before and after treatment. On average, participants' depression and anxiety symptom severity significantly decreased from baseline to the posttreatment time point, with small to medium effect sizes, respectively. Participants also expressed in qualitative interviews that they felt better and less anxious. Given that this study was an open trial without a control group, we cannot determine what drove this symptom improvement, as it could be due to HabitWorks or due to expectancy, attention from researchers, or other factors. However, this degree of symptom improvement is encouraging and suggests that further testing is warranted.

Limitations

Although this study addresses important limitations of prior studies testing DMHIs, we note several limitations that are inherent in a pilot open trial. First, this study was not designed to test efficacy and therefore included a small sample and did not include a comparison arm. Second, most of our sample was, relatively, highly educated and employed full time, limiting our understanding of the feasibility of HabitWorks for populations with lower income. Third, we prioritized qualitative data to gain a deeper understanding of participants' experience; however, results from this small sample may not generalize to the entire Black and Latinx populations. The Hispanic and Latinx communities are extremely diverse, and we were unable to examine outcomes for specific groups withing this community (eg, Mexican Americans and Puerto Ricans). Finally, at the time of this trial, HabitWorks was only available in English and in iOS, which also limits the generalizability. After securing additional funding, we have now developed the Android version, and creating a Spanish version will be essential for future evaluations of HabitWorks in the Hispanic community.

Conclusions

This study adds to the scarce literature evaluating DMHIs in Black and Latinx adults. These preliminary findings support the next stage of effectiveness testing in these populations. Future studies should compare HabitWorks to a credible control arm and include an independent measure of interpretation bias to determine whether improvement on the WSAP generalizes to other measures of interpretive style. In addition, HabitWorks should be investigated in other minoritized racial and ethnic groups that were not examined in this study, such as Asian and Asian Americans, to increase generalizability in our findings. Although none of our participants reported negative experiences due to using the app, it will be important to monitor safety in future studies with larger samples.

According to the 2020 census, only 5.08% of psychologists are African American or Black and 7.95% are Hispanic or Latinx, despite the US population identifying as 13.6% Black and 19.1% Hispanic or Latinx [42,43]. Moreover, waitlists for treatment services with a mental health provider across racial and ethnic identities are extremely long, hindering access to care. DMHIs, such as HabitWorks, have the potential to help bridge the gap between the demand for mental health services and low availability of providers, while addressing the most common barriers, such as cost, stigma, childcare, and time away from work. Our study provides evidence that HabitWorks may be a beneficial culturally sensitive and responsive intervention for minoritized racial and ethnic groups.

Acknowledgments

The authors would like to thank the National Institute of Mental Health for funding support (grant R01 MH129374). The authors would like to thank Laura Germine, Elizabeth Grinspoon, and Luke Scheuer for creating the web-based version of the Word Sentence Association Paradigm used for study assessments. The authors would also like to thank the Harvard Catalyst Community

Coalition for Equity Research for their feedback on the evaluation of HabitWorks and Dr Margarita Alegria for consulting with the authors on the study design and recruitment.

Authors' Contributions

IF contributed to the methodology, formal analysis, investigation, data curation, and writing the original draft. GG participated in the investigation and contributed to reviewing and editing the draft. KON contributed to the investigation, formal analysis, and reviewing and editing the draft. AT and ZM contributed to formal analysis. HB contributed to the study investigation. FGH contributed to formal analysis and reviewing and editing the draft. SC supervised the study and contributed to reviewing and editing the draft. CB contributed to conceptualization, methodology, validation, resources, writing the original draft, supervision, project administration, and funding acquisition.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Qualitative feedback themes.
[XLSX File (Microsoft Excel File), 15 KB - mental_v11i1e56758_app1.xlsx]

References

1. Hirsch CR, Meeten F, Krahé C, Reeder C. Resolving ambiguity in emotional disorders: the nature and role of interpretation biases. Annu Rev Clin Psychol 2016;12:281-305. [doi: 10.1146/annurev-clinpsy-021815-093436] [Medline: 27019398]
2. Beard C, Ramadurai R, McHugh RK, Pollak JP, Björgvinsson T. HabitWorks: development of a CBM-I smartphone app to augment and extend acute treatment. Behav Ther 2021 Mar;52(2):365-378 [FREE Full text] [doi: 10.1016/j.beth.2020.04.013] [Medline: 33622506]
3. Gonsalves M, Whittles RL, Weisberg RB, Beard C. A systematic review of the word sentence association paradigm (WSAP). J Behav Ther Exp Psychiatry 2019 Sep;64:133-148 [FREE Full text] [doi: 10.1016/j.jbtep.2019.04.003] [Medline: 31002979]
4. Beard C, Beckham E, Solomon A, Fenley AR, Pincus DB. A pilot feasibility open trial of an interpretation bias intervention for parents of anxious children. Cogn Behav Pract 2022 Nov;29(4):860-873 [FREE Full text] [doi: 10.1016/j.cbpra.2021.09.005] [Medline: 36506843]
5. Ramadurai R, Beckham E, McHugh RK, Björgvinsson T, Beard C. Operationalizing engagement with an interpretation bias smartphone app intervention: case series. JMIR Ment Health 2022 Aug 17;9(8):e33545 [FREE Full text] [doi: 10.2196/33545] [Medline: 35976196]
6. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019 Jul 22;21(7):e12869 [FREE Full text] [doi: 10.2196/12869] [Medline: 31333198]
7. Rai M, Vigod SN, Hensel JM. Barriers to office-based mental health care and interest in e-communication with providers: a survey study. JMIR Ment Health 2016 Aug 01;3(3):e35 [FREE Full text] [doi: 10.2196/mental.6068] [Medline: 27480108]
8. Sweetman J, Knapp P, Varley D, Woodhouse R, McMillan D, Coventry P. Barriers to attending initial psychological therapy service appointments for common mental health problems: a mixed-methods systematic review. J Affect Disord 2021 Apr 01;284:44-63. [doi: 10.1016/j.jad.2021.01.089] [Medline: 33582432]
9. Wolitzky-Taylor K, Fenwick K, Lengnick-Hall R, Grossman J, Bearman SK, Arch J, et al. A preliminary exploration of the barriers to delivering (and receiving) exposure-based cognitive behavioral therapy for anxiety disorders in adult community mental health settings. Community Ment Health J 2018 Oct;54(7):899-911 [FREE Full text] [doi: 10.1007/s10597-018-0252-x] [Medline: 29524078]
10. Alegría M, Alvarez K, Ishikawa RZ, DiMarzio K, McPeck S. Removing obstacles to eliminating racial and ethnic disparities in behavioral health care. Health Aff (Millwood) 2016 Jun 01;35(6):991-999 [FREE Full text] [doi: 10.1377/hlthaff.2016.0029] [Medline: 27269014]
11. Gan DZ, McGillivray L, Han J, Christensen H, Torok M. Effect of engagement with digital interventions on mental health outcomes: a systematic review and meta-analysis. Front Digit Health 2021 Nov 04;3:764079 [FREE Full text] [doi: 10.3389/fdgth.2021.764079] [Medline: 34806079]
12. Ellis DM, Draheim AA, Anderson PL. Culturally adapted digital mental health interventions for ethnic/racial minorities: a systematic review and meta-analysis. J Consult Clin Psychol 2022 Oct;90(10):717-733. [doi: 10.1037/ccp0000759] [Medline: 36227330]
13. De Jesús-Romero R, Holder-Dixon AR, Buss JF, Lorenzo-Luaces L. Race, ethnicity, and other cultural background factors in trials of internet-based cognitive behavioral therapy for depression: systematic review. J Med Internet Res 2024 Feb 01;26:e50780 [FREE Full text] [doi: 10.2196/50780] [Medline: 38300699]
14. Graham JR, Sorenson S, Hayes-Skelton SA. Enhancing the cultural sensitivity of cognitive behavioral interventions for anxiety in diverse populations. Behav Ther (N Y N Y) 2013 Jun;36(5):101-108 [FREE Full text] [Medline: 25392598]

15. Narine K, Salon YN, Beckham E, Beard C. Protocol for improving cultural sensitivity to assess and modify interpretation bias in diverse populations. In: Proceedings of the Association for Behavioral and Cognitive Therapies (ABCT) Annual Convention. 2021 Presented at: ABCT 2021; November 18-21, 2021; Virtual Event.
16. Hays PA. Addressing Cultural Complexities in Practice: Assessment, Diagnosis, and Therapy. Washington, DC: American Psychological Association; 2008.
17. Weisberg RB, Gonsalves MA, Ramadurai R, Braham H, Fuchs C, Beard C. Development of a cognitive bias modification intervention for anxiety disorders in primary care. Br J Clin Psychol 2022 Jan;61 Suppl 1:73-92. [doi: 10.1111/bjc.12281] [Medline: 33629751]
18. Beard C, Rifkin LS, Silverman AL, Björgvinsson T. Translating CBM-I into real-world settings: augmenting a CBT-based psychiatric hospital program. Behav Ther 2019 May;50(3):515-530. [doi: 10.1016/j.beth.2018.09.002] [Medline: 31030870]
19. Thomas DR. A general inductive approach for analyzing qualitative evaluation data. Am J Eval 2006;27(2):237-246. [doi: 10.1177/1098214005283748]
20. Gale RC, Wu J, Erhardt T, Bounthavong M, Reardon CM, Damschroder LJ, et al. Comparison of rapid vs in-depth qualitative analytic methods from a process evaluation of academic detailing in the veterans health administration. Implement Sci 2019 Feb 01;14(1):11 [FREE Full text] [doi: 10.1186/s13012-019-0853-y] [Medline: 30709368]
21. Nevedal AL, Reardon CM, Opra Widerquist MA, Jackson GL, Cutrona SL, White BS, et al. Rapid versus traditional qualitative analysis using the consolidated framework for implementation research (CFIR). Implement Sci 2021 Jul 02;16(1):67 [FREE Full text] [doi: 10.1186/s13012-021-01111-5] [Medline: 34215286]
22. Brooke J. SUS: a 'quick and dirty' usability scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B, editors. Usability Evaluation In Industry. London, UK: CRC Press; 1996.
23. Sauro J, Lewis JR. Quantifying the User Experience: Practical Statistics for User Research Second Edition. Burlington, MA: Morgan Kaufmann Publishers; 2016.
24. Devilly GJ, Borkovec TD. Psychometric properties of the credibility/expectancy questionnaire. J Behav Ther Exp Psychiatry 2000 Jun;31(2):73-86. [doi: 10.1016/s0005-7916(00)00012-4] [Medline: 11132119]
25. Nock MK, Ferriter C, Holmberg E. Parent beliefs about treatment credibility and effectiveness: assessment and relation to subsequent treatment participation. J Child Fam Stud 2006 Aug 25;16:27-38. [doi: 10.1007/s10826-006-9064-7]
26. Beard C, Amir N. Interpretation in social anxiety: when meaning precedes ambiguity. Cognit Ther Res 2009;33(4):406-415 [FREE Full text] [doi: 10.1007/s10608-009-9235-0] [Medline: 20046862]
27. Beard C, Rifkin LS, Björgvinsson T. Characteristics of interpretation bias and relationship with suicidality in a psychiatric hospital sample. J Affect Disord 2017 Jan 01;207:321-326. [doi: 10.1016/j.jad.2016.09.021] [Medline: 27741470]
28. Word Sentence Association Paradigm. Care Lab. URL: https://www.courtneybeardphd.com/wsap [accessed 2024-07-19]
29. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
30. Kroenke K, Spitzer RL, Williams JB, Monahan PO, Löwe B. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. Ann Intern Med 2007 Mar 06;146(5):317-325. [doi: 10.7326/0003-4819-146-5-200703060-00004] [Medline: 17339617]
31. Löwe B, Decker O, Müller S, Brähler E, Schellberg D, Herzog W, et al. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. Med Care 2008 Mar;46(3):266-274. [doi: 10.1097/MLR.0b013e318160d093] [Medline: 18388841]
32. Beard C, Björgvinsson T. Beyond generalized anxiety disorder: psychometric properties of the GAD-7 in a heterogeneous psychiatric sample. J Anxiety Disord 2014 Aug;28(6):547-552. [doi: 10.1016/j.janxdis.2014.06.002] [Medline: 24983795]
33. Kroenke K, Spitzer RL, Williams JB, Löwe B. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. Gen Hosp Psychiatry 2010;32(4):345-359. [doi: 10.1016/j.genhosppsych.2010.03.006] [Medline: 20633738]
34. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
35. Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr;114(1-3):163-173. [doi: 10.1016/j.jad.2008.06.026] [Medline: 18752852]
36. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
37. Clough B, Yousif C, Miles S, Stillerova S, Ganapathy A, Casey L. Understanding client engagement in digital mental health interventions: an investigation of the eTherapy Attitudes and Process Questionnaire. J Clin Psychol 2022 Sep;78(9):1785-1805 [FREE Full text] [doi: 10.1002/jclp.23342] [Medline: 35266551]
38. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018 Aug;21(3):116-119 [FREE Full text] [doi: 10.1136/eb-2018-102891] [Medline: 29871870]

39. Torous J, Wisniewski H, Liu G, Keshavan M. Mental health mobile phone app usage, concerns, and benefits among psychiatric outpatients: comparative survey study. JMIR Ment Health 2018 Nov 16;5(4):e11715 [FREE Full text] [doi: 10.2196/11715] [Medline: 30446484]
40. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021 Oct;20(3):318-335 [FREE Full text] [doi: 10.1002/wps.20883] [Medline: 34505369]
41. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
42. CWS data tool: demographics of the U.S. Psychology workforce. American Psychological Association. URL: https://www.apa.org/workforce/data-tools/demographics [accessed 2023-10-25]
43. U.S. Census Bureau QuickFacts: United States. United States Census Bureau. URL: https://www.census.gov/quickfacts/fact/table/US/PST045222 [accessed 2023-10-25]

Abbreviations

CEQ: Credibility and Expectancy Questionnaire
DMHI: digital mental health intervention
GAD: Generalized Anxiety Disorder
PHQ: Patient Health Questionnaire
REDCap: Research Electronic Data Capture
WSAP: Word Sentence Association Paradigm

Edited by J Torous; submitted 25.01.24; peer-reviewed by B Teachman, R De Jesús-Romero, A Ghosh; comments to author 05.03.24; revised version received 30.04.24; accepted 06.05.24; published 31.07.24.

Please cite as:
Ferguson I, George G, Narine KO, Turner A, McGhee Z, Bajwa H, Hart FG, Carter S, Beard C
Acceptability and Engagement of a Smartphone-Delivered Interpretation Bias Intervention in a Sample of Black and Latinx Adults: Open Trial
JMIR Ment Health 2024;11:e56758
URL: https://mental.jmir.org/2024/1/e56758
doi:10.2196/56758
PMID:39083330

Original Paper

Automated Real-Time Tool for Promoting Crisis Resource Use for Suicide Risk (ResourceBot): Development and Usability Study

Daniel DL Coppersmith[1], AM; Kate H Bentley[2], PhD; Evan M Kleiman[3], PhD; Adam C Jaroszewski[2], PhD; Merryn Daniel[1], BS; Matthew K Nock[1,2,4], PhD

[1]Department of Psychology, Harvard University, Cambridge, MA, United States

[2]Department of Psychiatry, Massachusetts General Hospital/Harvard Medical School, Boston, MA, United States

[3]Department of Psychology, Rutgers, The State University of New Jersey, Piscataway, NJ, United States

[4]Franciscan Children's Hospital, Brighton, MA, United States

Corresponding Author:
Daniel DL Coppersmith, AM
Department of Psychology
Harvard University
33 Kirkland St
Cambridge, MA, 02138
United States
Phone: 1 617 495 3800
Email: dcoppersmith@g.harvard.edu

Abstract

Background: Real-time monitoring captures information about suicidal thoughts and behaviors (STBs) as they occur and offers great promise to learn about STBs. However, this approach also introduces questions about how to monitor and respond to real-time information about STBs. Given the increasing use of real-time monitoring, there is a need for novel, effective, and scalable tools for responding to suicide risk in real time.

Objective: The goal of this study was to develop and test an automated tool (ResourceBot) that promotes the use of crisis services (eg, 988) in real time through a rule-based (ie, if-then) brief barrier reduction intervention.

Methods: ResourceBot was tested in a 2-week real-time monitoring study of 74 adults with recent suicidal thoughts.

Results: ResourceBot was deployed 221 times to 36 participants. There was high engagement with ResourceBot (ie, 87% of the time ResourceBot was deployed, a participant opened the tool and submitted a response to it), but zero participants reported using crisis services after engaging with ResourceBot. The most reported reasons for not using crisis services were beliefs that the resources would not help, wanting to handle things on one’s own, and the resources requiring too much time or effort. At the end of the study, participants rated ResourceBot with good usability (mean of 75.6 out of 100) and satisfaction (mean of 20.8 out of 32).

Conclusions: This study highlights both the possibilities and challenges of developing effective real-time interventions for suicide risk and areas for refinement in future work.

(JMIR Ment Health 2024;11:e58409) doi:10.2196/58409

KEYWORDS

suicidal thoughts; suicidal behaviors; ecological momentary assessment; crisis resources; real-time tool; self-report; psychoeducation; app

Introduction

Overview

Real-time monitoring methods—such as ecological momentary assessment (EMA)—capture fine-grained, “real-world” information about suicidal thoughts and behaviors (STBs) as they occur and thus have immense potential to advance our understanding of suicide [1,2]. The promise of real-time monitoring methods for STBs has been widely recognized, as evidenced in part by the recent proliferation of published studies using EMA to study STBs. A recent systematic review identified 45 articles that have used real-time monitoring methods to study STBs [3].

Collecting information about STBs in real time, however, poses important safety, ethical, and methodological concerns [4]. One complex ethical challenge is regarding how to monitor or respond to incoming information about STBs from suicidal or self-injuring individuals. For example, when participants submit a survey response indicating current suicidal intent that researchers can access in real time, should the study team intervene? How should the study team determine when an intervention is needed? What should the intervention involve?

A consensus statement (generated from a panel of 24 experts) on the ethical and safety practices for conducting real-time monitoring studies of individuals at risk for suicide and related behaviors was recently released [4]. There was a strong (about 94%) agreement that when participants provide a "high-risk" response, the study team should reach out to them directly to conduct a suicide risk assessment as soon as possible (within 12-24 hours for responses indicating "imminent" risk). An exception the panel noted, however, was anonymous studies where contact information for participants is not known. A systematic review of practices in 59 previous or ongoing digital monitoring studies of STBs [5], however, indicates a gap between this apparent consensus and reality, as just over half (58%) of studies reported monitoring and intervening upon incoming responses during the study. Thus, there remains a notable departure between expert consensus and real-world practices for responding to incoming data. The other most common safety practice identified in this review was automated notifications (eg, pop-up messages with crisis resources) triggered by specific survey responses, which was used in roughly half of the studies included.

Both common approaches of researchers intervening, and static pop-up messages have significant limitations. Static messages are easy for participants to habituate to or ignore, especially during periods of high distress. Human- (often clinician-led) active interventions (eg, calling participants) by the research team are resource-intensive and have the potential to cause undesired reactivity. If participants are aware the researchers will act if they provide a "high-risk" response, participants may underreport STBs (or stop responding to study surveys entirely) to avoid an unwanted intervention. A recent empirical investigation of this issue of reactivity found mixed support for reactivity to real-time interventions (in this case, phone outreach by the study team) [6]. If responding to incoming data on STBs does influence individuals' responding behavior, this could muddle the validity of the resultant study data. Another key limitation is the feasibility of monitoring and responding to incoming data, as this approach tends to require considerable staff, technology systems, and funding. The resources required for these safety protocols may partially explain the gap between expert consensus and real-world practices.

Given the increasing use of real-time monitoring methods [7], there is a need for novel, effective, and scalable tools for responding to suicide risk in real time. Recent advances in mobile technologies have the potential to facilitate automated, potentially highly efficient risk assessment strategies (ie, that do not require a clinician calling and may be less subject to reactivity) and deployment of specific types of notifications or alerts delivered directly to participants. Automated assessments and interventions could be faster and less burdensome for both participants and researchers. Automated tools (here, referring to systems that use rule-based [eg, if-then] logic, not those that use generative artificial intelligence; Blease and Torous [8]) might be more effective than human interventions because they can reach the participant faster than study staff. Recently developed rule-based automated assessments and interventions for STBs have shown promise. One such automated intervention consists of a brief, automated risk assessment and barrier reduction intervention (BRI) designed to increase the use of crisis resources [9]. The BRI component includes psychoeducation designed to reduce perceived barriers to using crisis resources by clearing up misconceptions on which these barriers may be based. A large-scale clinical trial found that this intervention was associated with a 23% increase in the use of crisis services [9]. A similar trial of an automated intervention also found that a brief automated intervention could increase the reported use of crisis resources [10]. The promise of this type of intervention for real-time monitoring is that it could guide participants to resources during high-risk situations. This type of intervention is consistent with recent calls for just-in-time adaptive interventions for suicide prevention, which are intended to provide the right type of support at the right time [11,12]. Therefore, providing ethical, scalable, and fast risk management during research studies.

This automated intervention, however, has only been tested at a single time point [9,10] and has never been tested in the context of participation in real time, longitudinal monitoring of those at risk for suicide. Understanding how this tool translates to this context is crucial given that one cannot assume engagement and efficacy generalize across contexts in digital health [13]. Engagement is a crucial first step given that digital interventions often suffer from low engagement and a quick drop-off in use over time. In short, to realize the potential of an automated digital intervention, participants must engage with the intervention. Therefore, it is crucial to first adapt this intervention for real-time monitoring. Second, it is important to understand the feasibility, acceptability, and utility of the intervention for real-time monitoring. The development of such a tool has the potential to improve the safety, scale, and breadth of real-time monitoring studies of STBs.

Study Aims and Outcomes

The overall project aims were (1) to adapt an evidence-based BRI that aims to increase the use of crisis resources for deployment in real time monitoring research and to (2) to test the feasibility, acceptability, and utility of this tool in a real-time monitoring study of suicidal thoughts. For the latter objective, the key outcomes of interest were as follows: (1) Do people engage with the BRI? (2) Do people contact crisis resources after the BRI? (3) What do people report about the acceptability and usability of the BRI?

Methods

Adapting BRI

To adapt the BRI [9], members of the author team first met to develop the workflow of the intervention as well as the text to be deployed. Over multiple meetings, the author team iterated on the workflow to be appropriate for the context of a real-time monitoring study. The main changes to be appropriate for real-time monitoring research included changing the beginning of the workflow to focus on the recent submission of a survey with self-report suicidal intent and for the text to be brief enough for viewing on mobile phones. The authors then worked with a graphic designer (MD) to name the BRI and develop images to pair with each text screen of the intervention. Images were added to promote engagement and to help differentiate the BRI from the base real-time survey questions. The BRI was named ResourceBot and images of the ResourceBot were generated for each screen. In total, there were 23 unique images or text slides generated. The ResourceBot was designed to be triggered after a participant submits a survey with elevated suicidal intent. The workflow of ResourceBot is: (1) confirm current distress (to ensure the participant-reported distress was not made in error), (2) offer resources, (3) identify barriers to using resources, and (4) provide psychoeducation to promote resource use. A general overview of the ResourceBot workflow is provided in Figure 1 and example slides of the tool are provided in Figures 2 and 3. All images of the ResourceBot and the decision logic are provided in Multimedia Appendix 1. ResourceBot was built and deployed directly in the Metricwire app (Metricwire Inc), which was also used for real-time monitoring surveys.

Figure 1. General overview of ResourceBot flow.

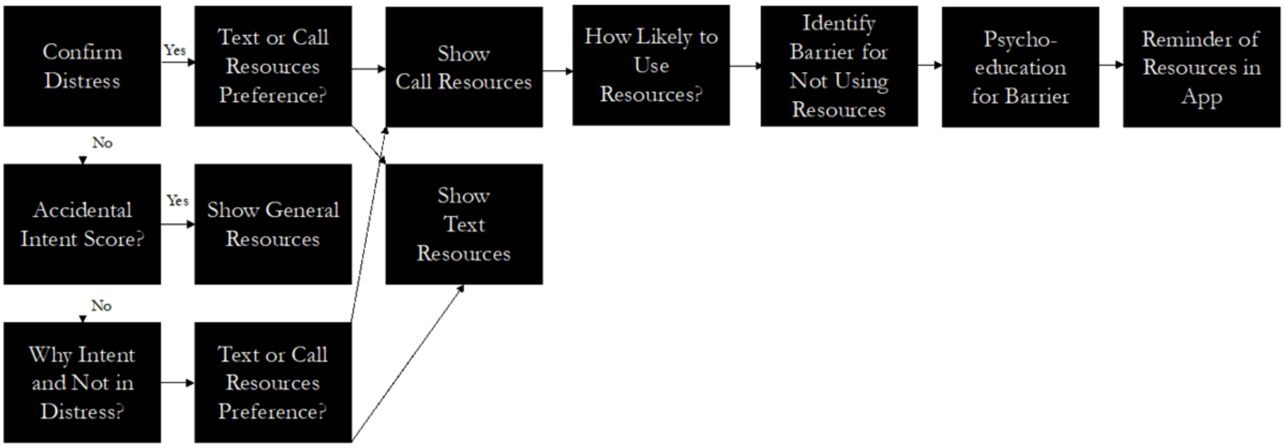


Figure 2. Example of ResourceBot offering resources. (A) Confirming current distress. (B) Asking about the type of resources to view. (C) Providing resources to call if participants selected that they wanted resources to call.

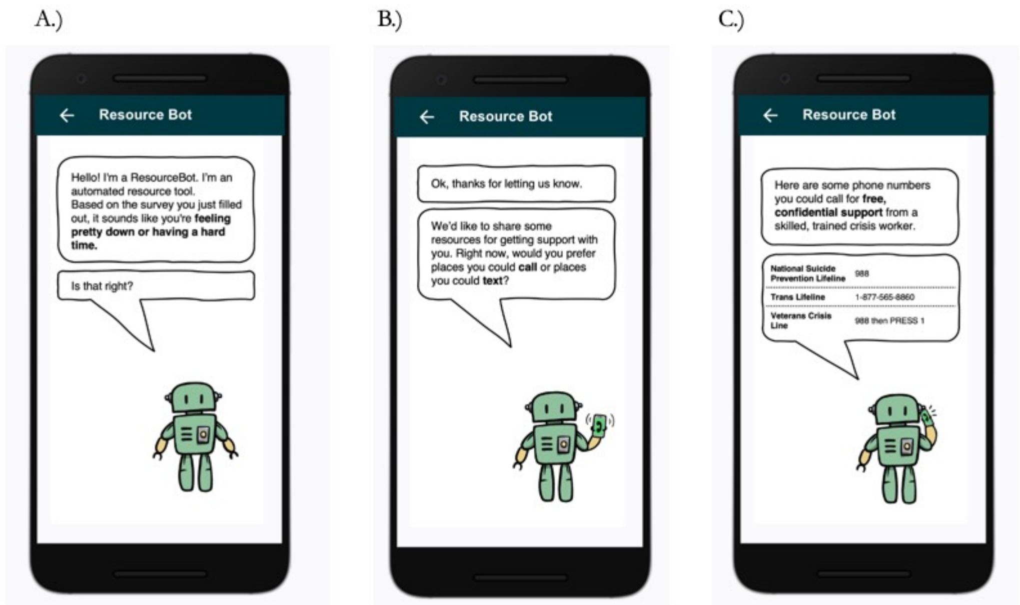
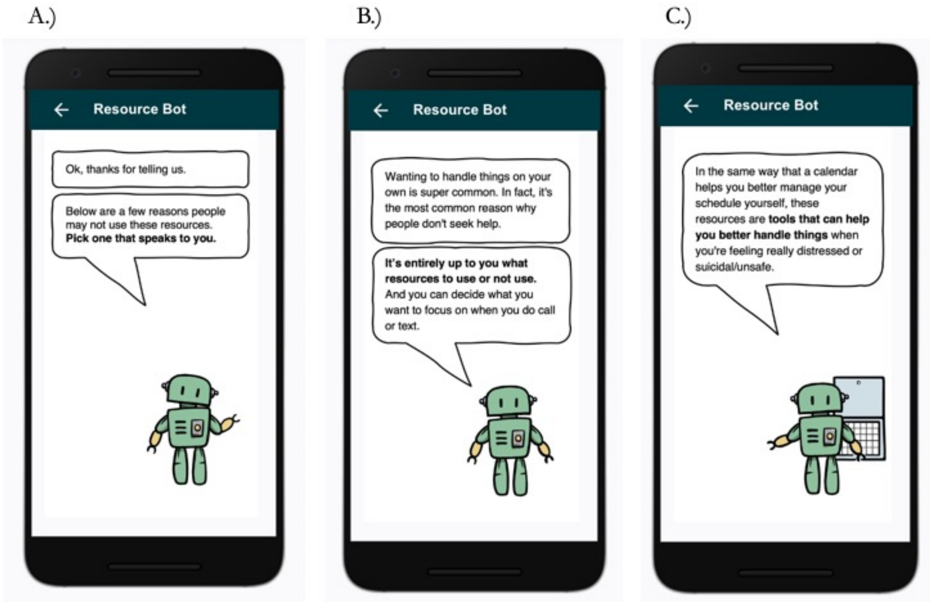


Figure 3. Barrier reduction intervention in ResourceBot. (A) Asking participants to select barriers to using resources. (B) The first psychoeducation slide shows if “I can handle it on my own” is selected. (C) The second psychoeducation slide shows if “I can handle it on my own” is selected.



As part of the development of the ResourceBot tool, we ran a pilot study with 8 participants to primarily determine when to trigger ResourceBot among other topics. In this pilot study, across 369 completed total surveys, the ResourceBot tool was only deployed twice to one participant. This one participant engaged with the ResourceBot and the crisis resources it provided. Based on the results of this pilot study, we lowered the threshold for triggering the ResourceBot from a suicidal intent rating on the EMA or daily survey of greater than 3 (out of 10) to greater than 1 (out of 10) for the main wave of data collection. This threshold was lowered so that a greater number of participants would be provided with the tool and able to provide feedback on it. Given that in the pilot, we found that the ResourceBot was successfully deployed, we proceeded to the main wave of data collection.

Participants

Participants were 74 adults who were recruited through the Prolific research platform. Prolific was selected for recruitment given it has been associated with high-quality data [14,15]. The demographics and clinical characteristics of the participants are provided in Table 1. The specific inclusion criteria for the study were suicidal thoughts in the past week, the ability to speak and write English fluently, access to an internet-capable smartphone, and living in the United States. To identify participants eligible for the study, a screening survey study was sent out to participants on the platform who lived in the United States, were fluent in English, had at least a 90% approval rating of past studies on Prolific, and endorsed a lifetime history of mental illness. We used the filter of lifetime history of mental illness to increase the prevalence of suicidal thinking in the population initially screened for recent suicidal thoughts. Consistent with recommendations for web-based data collection [16] and to help ensure quality data and attentive responding, suicidal thoughts were asked about in multiple ways on the screening survey (eg, binary lifetime presence of thoughts, ordinal recency of thoughts, and text entry number of days with thoughts). To be included participants needed to provide a consistent response pattern on the screening survey of endorsing the lifetime presence of suicidal thoughts, reporting the most recent time they had suicidal thoughts in the past week, and writing a number greater than zero for lifetime days with suicidal thoughts. Participants were compensated US $0.60 for completing the brief screening survey. Participants who met eligibility criteria based on the screening survey were then invited to the main study. To be included in the current analysis, participants had to complete at least 1 real-time survey.

Table 1. Participant demographics and clinical variables (n=74).

Demographics	Variables
Age (years), mean (range)	34.7 (20 to 62)
Sex assigned at birth, n	
Female	37
Male	37
Gender identity, n	
Female	31
Male	36
Genderqueer, non-binary, gender fluid	6
Other	1
Race, n	
White	59
Black	6
Asian	3
Multiracial	5
Other	1
Ethnicity, n	
Hispanic/Latino	7
Highest education level, n	
Less than high school	1
High school/GED[a]	11
Some college	22
2-year college degree	6
4-year college degree	30
Master degree	3
Professional degree	1
Lifetime suicide attempt, n	
Yes	46
Lifetime days with suicidal thoughts, median (SD)	550 (2451)
Patient Health Questionnaire-9, mean (SD)	18.0 (6.6)
Generalized Anxiety Disorder-7, mean (SD)	14.2 (5.6)
Psychotherapy history, n	
Lifetime use	59
Current use	22
Medication for mental health history, n	
Lifetime use	59
Current use	29

[a]GED: General Education Development.

Ethical Considerations

All study procedures were approved by the Harvard University Area institutional review board (protocol IRB22-0012; "Automated Real Time Safety and Monitoring Study"). All participants provided informed consent. Following the screening survey, eligible participants were sent the consent form and baseline survey. The informed consent form contained specific language about how real-time responses were not being monitored by the study team and automated messages would encourage resources.

Procedure

In the baseline survey, participants completed questionnaires assessing STBs, psychiatric symptoms, and mental health care history. At the end of the survey, participants were provided with instructions for downloading the Metricwire smartphone app and then confirmed that they had downloaded the app before submitting the baseline survey for approval. Participants were compensated US $10 for completing the baseline survey. Participants were then sent to the Prolific platform, an anonymous login for the Metricwire smartphone app.

The real-time monitoring period was 2 weeks long and began the day after participants logged into their anonymous accounts. Six surveys were sent each day. Five momentary surveys were sent semirandomly between 9 AM and 9 PM and spaced at least 90 minutes apart. One daily survey was sent at 8 PM each day. The momentary survey stayed open to complete for 1 hour and the daily survey for 2 hours. The last momentary survey of the day and the daily survey could overlap; ultimately, 11% percent of momentary surveys were submitted during the hours of 8 PM to 10 PM (after the daily survey was prompted). Participants were paid US $0.25 for each survey they completed. If participants completed 5 or more surveys in a day, they received a US $1.00 bonus for that day. All payments for the real-time surveys were sent the day after the real-time monitoring period was complete.

On the day after the last day of the real-time monitoring period, participants were sent an exit survey via MetricWire asking them about their experiences in the study. If a participant reported that they received the ResourceBot, they were then asked questions about the acceptability and usability of it. The exit survey stayed open to complete for 8 hours. For completing the exit survey, participants were paid US $3. With this payment structure, participants could earn up to US $48 in the study. All payments were sent through Prolific.

Baseline Assessment Surveys

In the baseline assessment, participants completed a self-report version of the Self-Injurious Thoughts and Behaviors Interview—Revised (SITBI-R; Fox et al [17]). The SITBI-R measures the presence, frequency, recency, and other aspects of self-injurious thoughts and behaviors. The SITBI-R has shown excellent reliability and validity [17]. Participants also completed the Patient Health Questionnaire-9 (PHQ-9; Kroenke et al [18]). The PHQ-9 is a widely used brief measure of the severity of symptoms of depression in the past 2 weeks. PHQ-9 scores range from 0 to 27. In this study, the PHQ-9 had excellent internal consistency (Cronbach α=0.89). The Generalized Anxiety Disorder 7-item (GAD-7; Spitzer et al [19]) was also administered at baseline. The GAD-7 is a brief measure of the severity of symptoms of anxiety in the past 2 weeks. GAD-7 scores range from 0 to 21. In this study, the GAD-7 had excellent internal consistency (Cronbach α=0.90).

We also measured participants' mental health treatment history and crisis service use with measures created for this study. For mental health history, participants were asked if they had ever received any form of mental health treatment from a professional. If so, they were asked if they had ever received talk therapy for mental health and if they were currently receiving talk therapy. Participants were also asked if they had ever been prescribed medications for mental health reasons and if they were currently being prescribed medications for mental health reasons. For crisis resources use, participants were asked if they had "ever called a suicide crisis lifeline (eg, 988 Suicide & Crisis Lifeline)" and if they had ever "texted a suicide crisis service (eg, Crisis Text Line)." If a participant endorsed using a crisis resource, they were asked how helpful it was on a 0 (not at all helpful) to 5 (very helpful) scale and how likely they were to use the resource in the future on a 0 (not at all likely) to 5 (very likely) scale. All participants were also asked, "Below are a few reasons that people may not call or text suicide crisis services. Do any of these speak to you as reasons why you wouldn't call or text a suicide crisis service in the future?" The reasons listed were: "I can handle things on my own," "Too much time/effort," "No professionals," "No police," "They won't help," and "None of these reasons."

EMA Items

The momentary and daily surveys contained multiple items on suicidal thinking, affective states, and cognitive processes. For the present analyses, the most relevant item is the suicidal intent item, which was used to trigger the ResourceBot. The exact item wording was "How strong is your intent to kill yourself right now? Intent = to what extent are you actually going to kill yourself." Participants rated this item on a 0 ("not at all") to 10 ("very strong") scale. This item has been used in previous real-time studies [20,21] and a similar item has shown predictive validity for suicidal behavior [22]. A daily version of the item was also included, "Today how strong was your intent to kill yourself? Intent = to what extent are you actually going to kill yourself." The daily item had the same scale and anchors as the momentary item. Suicidal intent was selected to trigger the ResourceBot because, in a consensus statement, it was identified as a key piece of information for determining real-time risk [4]. Furthermore, the level of suicidal intent has been used to determine interventions in other real-time risk protocols [5,6].

Another relevant item in the daily survey was an item on crisis resource use. Participants were specifically asked, "Today, did you use any crisis resources? For example, did you call 988?" with the response options of yes or no. If a participant selected yes, then a participant would be asked "What crisis resource did you use? For example, calling 988, texting crisis text line." If a participant selected yes, they would also be asked how helpful the resource was. These items were included to capture additional data on crisis resource use in case participants did not complete the ResourceBot follow-up survey.

Crisis Resources

At all times in the real-time monitoring period, there was an always-available list of resources they could open (ie, "resources survey") in the MetricWire app. The resources survey contained a list of the following resources: National Suicide Prevention Lifeline, Crisis Text Line, Trans Lifeline, Veterans Crisis Line Chat, Veterans Crisis Line, and Trevor Project Chat. A participant could select a resource from the list, which would take them to a page with more information on the resource and how to contact them. The page included a direct link to the webpage of the resource. Participants were told that "the research team may be able to see if you select a resource, but the team cannot see what you communicate or share with the support lines." These resources were selected because they offered support through different means of communication (eg, calling or texting) and support for different populations. Additionally, at the end of all real-time surveys, participants were reminded that the National Suicide Prevention Lifeline, Crisis Text Line, and their local emergency department were helpful resources.

ResourceBot Tool

The ResourceBot tool was built directly in MetricWire. When a participant submitted a survey with a suicidal intent rating greater than 1, it would trigger the deployment of the ResourceBot tool. This threshold of suicidal intent was lower than the threshold used in previous studies [5] because the intervention is lower intensity compared with other interventions (eg, clinician contact). A stop rule in place was that the ResourceBot tool would be only deployed once a day to limit the burden and increase engagement. The ResourceBot tool was sent immediately after the survey submission. If a participant did not open the ResourceBot survey, then a reminder notification was sent 5 minutes after the initial deployment. The ResourceBot survey stayed open for 4 hours. Once a participant opened the ResourceBot survey, a participant was guided through a protocol that (1) confirmed current distress, (2) offered crisis resources, (3) identified barriers to crisis resource use, and (4) provided psychosocial education on resources. An example interaction is provided in Figures 2 and 3. The tool is designed to overcome common concerns and misconceptions (ie, barriers) related to using crisis services, thereby increasing the use of these services.

ResourceBot Follow-Up Survey

One hour after the ResourceBot survey was submitted, a brief follow-up survey was sent. The survey asked if a participant used a resource since they were sent the ResourceBot. If yes, they were asked what resource they used and how helpful the resource was. If not, participants were asked why they did not use a resource. The response options for why they did not use a resource included: "Too much time/effort," "Didn't think it would help," "Felt better without them," "I handled it on my own," and "Other." Participants could enter more information into an open-ended text field if they selected "Other."

Exit Survey

At the end of the EMA period, participants were sent a survey asking them about their experiences in the study. All participants were sent questions about the overall study and participants who were sent the ResourceBot were sent additional questions about the ResourceBot. All participants were asked, "Did you feel comfortable answering the cellphone questions honestly?" and rated it on a scale of 0 (not at all) to 5 (very much). All participants were asked, "Did you receive the ResourceBot, which directed you to crisis services, in the study?" and shown a picture of the ResourceBot as a reminder. If a participant endorsed receiving the ResourceBot, they were sent the Client Satisfaction Questionnaire (CSQ; Larsen et al [23]) and the System Usability Scale (SUS; Lewis and Sauro [24]). The CSQ is an 8-item scale that produces a score from 8 to 32 with higher scores indicating greater satisfaction. In this study, the CSQ was used as a measure of acceptability and had excellent internal consistency (Cronbach α=0.96) The SUS is a 10-item scale that produces an original score of 0 to 40 which is then multiplied by 2.5 to create scaled scores of 0 to 100 with higher scores indicating that the user rates the system as more useable. In this study, the SUS had excellent internal consistency (Cronbach α=0.91).

Statistical Analysis

For all analyses, we focus on descriptive statistics. For the first aim of whether people engage with the ResourceBot, we focus on how often participants open and submit the ResourceBot survey. Although there are multiple ways to operationalize engagement [25], we highlight this simple definition of engagement for this first examination of ResourceBot. We also report on data provided within the ResourceBot survey, including the endorsed barriers to using crisis resources.

One factor that could have affected participants' engagement with the ResourceBot is the current level of suicidal intent. For example, a participant with higher levels of current intent (eg, 9 out of 10) and possibly greater risk may engage with the tool in a different way than a participant with lower levels of current intent (eg, 2 out of 10). To understand the relationship between the level of intent prior to the ResourceBot and engagement with the ResourceBot we ran additional analyses. We identified the momentary survey submitted closest in time (ie, the trigger survey) to the submitted ResourceBot survey. This resulted in momentary intent ratings for 181 of the 192 ResourceBot engagements; the 11 other engagements were triggered by a daily survey report. We focused on the 181 engagements for the subsequent descriptive analyses. The average time difference between the submitted momentary survey and the submitted ResourceBot survey was 7.6 minutes. We then categorized the momentary intent levels into low and high levels. We operationalized low as a score of 2, 3, or 4 and high as a score of greater than 5. This resulted in 92 low-intent ResourceBot engagements and 89 high-intent ResourceBot engagements. We present descriptive statistics on data within the ResourceBot encounter by momentary intent level.

Due to the potential for habituation to ResourceBot content with multiple deployments over time, we also isolated each participant's first submitted engagement with the ResourceBot and presented descriptive statistics on data within this first encounter with the ResourceBot.

For the second aim, if people contact crisis resources after engaging with ResourceBot, we focus on how often in the ResourceBot follow-up survey do people report using crisis resources. We also report on crisis resource use reported in the daily survey as well as the frequency of viewing the crisis resources. For the third aim, we report on the descriptive statistics on exit survey scores on acceptability and usability. We also report additional exit survey data on the honesty of responding. Together these analyses use multiple sources to comprehensively describe the feasibility, acceptability, and utility of a real-time crisis resource tool. All data analysis codes and results can be viewed on the Open Science Framework [26].

Results

Descriptive Statistics

Baseline data on lifetime use of crisis hotlines, experiences on crisis hotlines, and barriers to future use of crisis hotlines are provided in Table 2. Most participants (49/74, 66% for calling and 58/74, 78% for texting) had not used crisis hotlines in their life. Participants who had previously used crisis hotlines reported that they in general were not helpful (calling mean helpfulness=1.12 out of 5; texting mean helpfulness=1.19 out of 5). Participants reported on average they were not very likely to use hotlines in the future. The most frequently endorsed reason for not using crisis hotlines in the future was the belief that they would not help.

Table 2. Baseline crisis hotline lifetime histories.

Baseline histories	Variables
Lifetime called crisis line, n (%)	
Yes (Percent)	25 (34)
Lifetime texted crisis line, n (%)	
Yes (Percent)	16 (22)
Helpfulness of calling crisis line[a], mean (SD)	1.12 (1.54)
Helpfulness of texting call line[b], mean (SD)	1.19 (1.33)
How likely to call crisis line in the future[c], mean (SD)	1.14 (1.47)
How likely to text crisis line in the future[d], mean (SD)	1.35 (1.62)
Reasons for not using crisis lines in the future[e]	
I can handle things on my own	26
Too much time or effort	12
No professionals	12
No police	17
They would not help	49
None of these reasons	6

[a]Only answered by participants who answered they had called a crisis line (n=25).

[b]Only answered by participants who answered they had texted a crisis line (n=16).

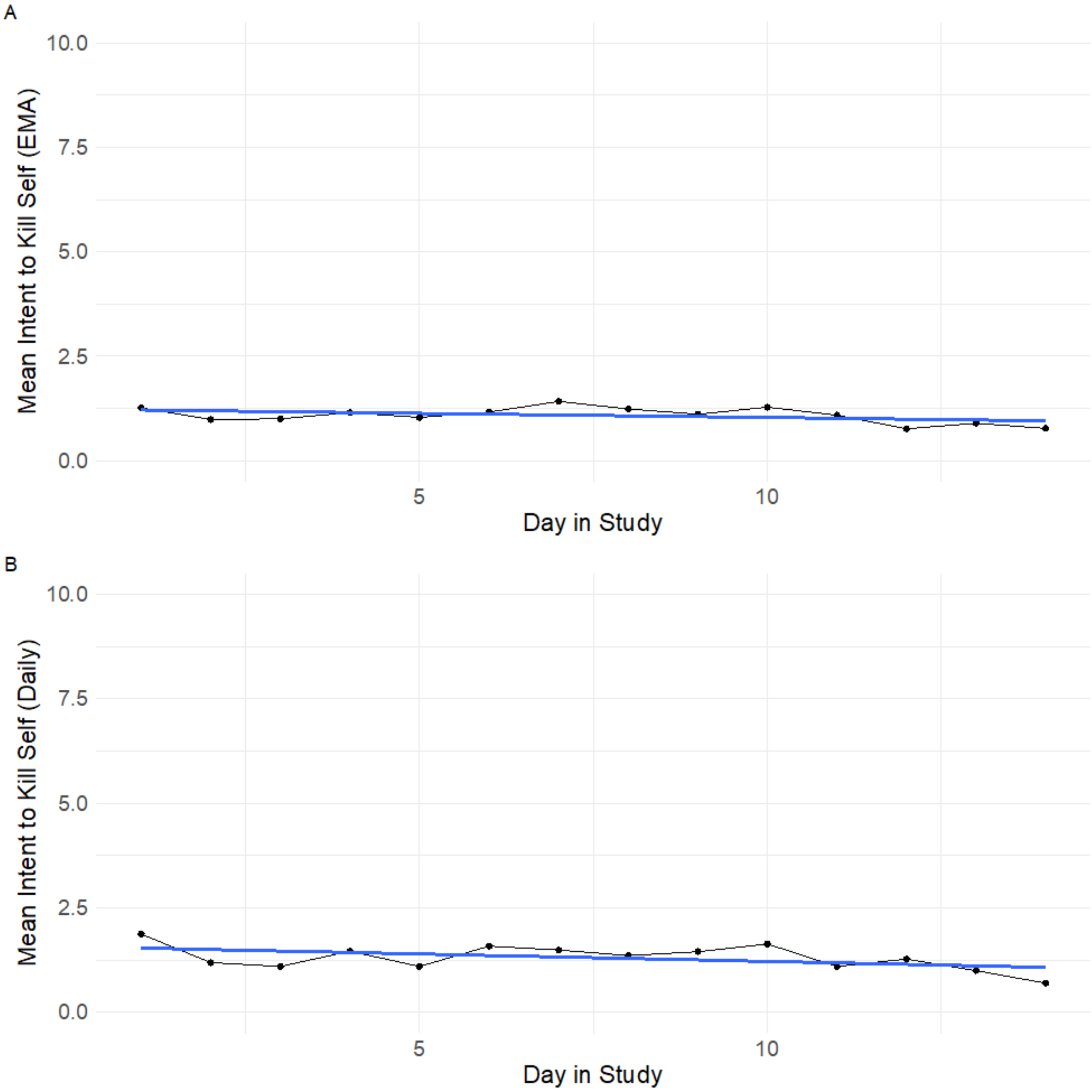
[c]Answered by all participants (n=74).

[d]Answered by all participants (n=74).

[e]Participants could select multiple reasons.

Participants completed 2909 momentary surveys and 679 daily surveys. A total of 74 participants completed at least 1 momentary survey and 72 participants completed at least 1 daily survey. The average number of momentary surveys submitted was 39.3 (range 2 to 70) and for daily surveys, it was 9.4 (range 1 to 14). The average compliance rate for the momentary surveys was 56% (range 3% to 100%) and for the daily surveys was 67% (range 7% to 100%). The daily averages of EMA and Daily survey scores of the intent to kill oneself item are shown in Figure 4.

Figure 4. Mean suicidal intent scores over time in the study. (A) Mean EMA intent scores by day in study. (B) Mean daily intent scores by day in study. Blue line is the linear trend of the mean intent score by day in the study. EMA: ecological momentary assessment.



The ResourceBot was deployed 221 times to 36 participants. A total of 35 participants engaged with the ResourceBot at least once. The ResourceBot was deployed and engaged multiple times by 28 (80%) of the 35 participants who engaged it at least once. The other 7 (20%) of the 35 participants engaged with it only once.

The exit survey was completed by 44 participants. We compared participants who completed the exit survey (n=44) to those who did not (n=30) on EMA compliance percentage, daily survey compliance percentage, mean EMA suicidal intent severity, and mean daily suicidal intent severity. We conducted this descriptive retention analysis to understand if the type of participant who completed the exit survey may be biased in some way. We found that participants who completed the exit survey had higher EMA (72% vs 34%) and daily survey (81% vs 46%) compliance rates than those who did not complete the exit survey. We also found similar mean EMA (1.31 completers vs 1.15 noncompleters) and daily suicidal intent ratings (1.48 completers vs 1.43 noncompleters) by exit survey status.

Among those who completed the exit survey, 19 reported receiving the ResourceBot. We cross-checked participants' self-reports of receiving the ResourceBot with the ResourceBot deployment data. Eighteen of the 19 who reported receiving the ResourceBot in the exit survey matched with the ResourceBot deployment data. The one participant who reported receiving the ResourceBot, but did not actually receive it was excluded from the ResourceBot exit survey analysis. One of the 18 participants did not complete all items for the CSQ and therefore, we report on 17 participants for the CSQ.

Do People Engage With the ResourceBot?

There was 87% (192 out of 221) overall engagement (defined as opening and submitting) with the ResourceBot tool and 86% (165 out of 192) compliance with the ResourceBot follow-up survey. The different components of the ResourceBot and the frequency of responses are provided in Table 3. In the majority of deployments (122 out of 192) participants confirmed that they were in distress. For participants who reported that they were not in distress, the most commonly reported reason was being used for these thoughts or feelings. Text resources (n=81) were more frequently selected than call resources (n=19). For the likelihood of using resources, the most common response was not likely (n=48). For barriers to using resources, the most common responses included that "it won't help" and "I can handle it on my own." As shown in Table 3, participants often skipped out of the ResourceBot at various stages of the tool.

Table 3. ResourceBot responses for all ResourceBot engagements.

Question	Response (n)
It sounds like you're feeling pretty down or having a hard time. Is that right? (n=192)	• Yes (122) • No (14) • Skipped (56)
Why are you having high intent but not feeling down/having a hard time? (n=8)	• I am used to these thoughts/feelings (7) • I don't need help for these thoughts/feelings (1)
Prefer places could call or places you could text? (n=130)	• Call (19) • Text (81) • Skipped (30)
How likely to use the resources shared? (n=100)	• Not Likely (48) • Somewhat Likely (34) • Very Likely (6) • Skipped (12)
Reasons people may not use these resources (n=82; shown if not likely or somewhat likely to use resources)	• It won't help (37) • I can handle it on my own (26) • Too much time/effort (5) • No police (5) • I may not use these resources for a reason not otherwise listed (4) • No professionals (3) • Skipped (2)
Reasons people may not use these resources (n=6; shown if very likely to use resources)	• It won't help (1) • I can handle it on my own (2) • Too much time/effort (0) • No police (0) • Not really - another reason (3) • No professionals (0) • Skipped (0)

Responses to the ResourceBot by level of momentary intent are presented in Table 4. The patterns of responding were similar across low and high levels of intent. For example, across both low and high levels of intent, participants most commonly reported being not likely to use the Resources presented and most frequently endorsed the barrier of the belief that the resources would not help. Results from the first encounter with ResourceBot only are presented in Table 5. Results from the first encounters showed similar trends to data from all encounters.

Table 4. ResourceBot responses by ecological momentary assessment suicidal intent severity (n=182).

Question	Low Intent Responses (n)	High Intent Responses (n)
It sounds like you're feeling pretty down or having a hard time. Is that right?	• Yes (53) • No (8) • Skipped (32)	• Yes (64) • No (6) • Skipped (19)
Why are you having high intent but not feeling down/having a hard time?	• I am used to these thoughts/feelings (5) • I don't need help for these thoughts/feelings (0)	• I am used to these thoughts/feelings (2) • I don't need help for these thoughts/feelings (1)
Prefer places could call or places you could text?	• Call (7) • Text (41) • Skipped (10)	• Call (12) • Text (36) • Skipped (17)
How likely to use the resources shared?	• Not Likely (21) • Somewhat Likely (18) • Very Likely (2) • Skipped (7)	• Not Likely (27) • Somewhat Likely (12) • Very Likely (4) • Skipped (5)
Reasons people may not use these resources	• It won't help (15) • I can handle it on my own (15) • Too much time/effort (3) • No police (3) • I may not use these resources for a reason not otherwise listed (2) • No professionals (0) • Skipped (1)	• It won't help (21) • I can handle it on my own (9) • Too much time/effort (2) • No police (2) • I may not use these resources for a reason not otherwise listed (2) • No professionals (3) • Skipped (0)
Reasons people may not use these resources	• It won't help (0) • I can handle it on my own (2) • Too much time/effort (0) • No police (0) • Not really - another reason (0) • No professionals (0) • Skipped (0)	• It won't help (1) • I can handle it on my own (0) • Too much time/effort (0) • No police (0) • Not really - another reason (3) • No professionals (0) • Skipped (0)

Table 5. ResourceBot responses (first engagement only).

Question	Response (n)
It sounds like you're feeling pretty down or having a hard time. Is that right? (n=35)	• Yes (16) • No (0) • Skipped (19)
Prefer places could call or places you could text? (n=16)	• Call (1) • Text (12) • Skipped (3)
How likely to use the resources shared? (n=13)	• Not Likely (4) • Somewhat Likely (5) • Very Likely (0) • Skipped (4)
Reasons people may not use these resources (n=9)	• It won't help (5) • I can handle it on my own (2) • Too much time/effort (0) • No police (1) • I may not use these resources for a reason not otherwise listed (1) • No professionals (0) • Skipped (0)

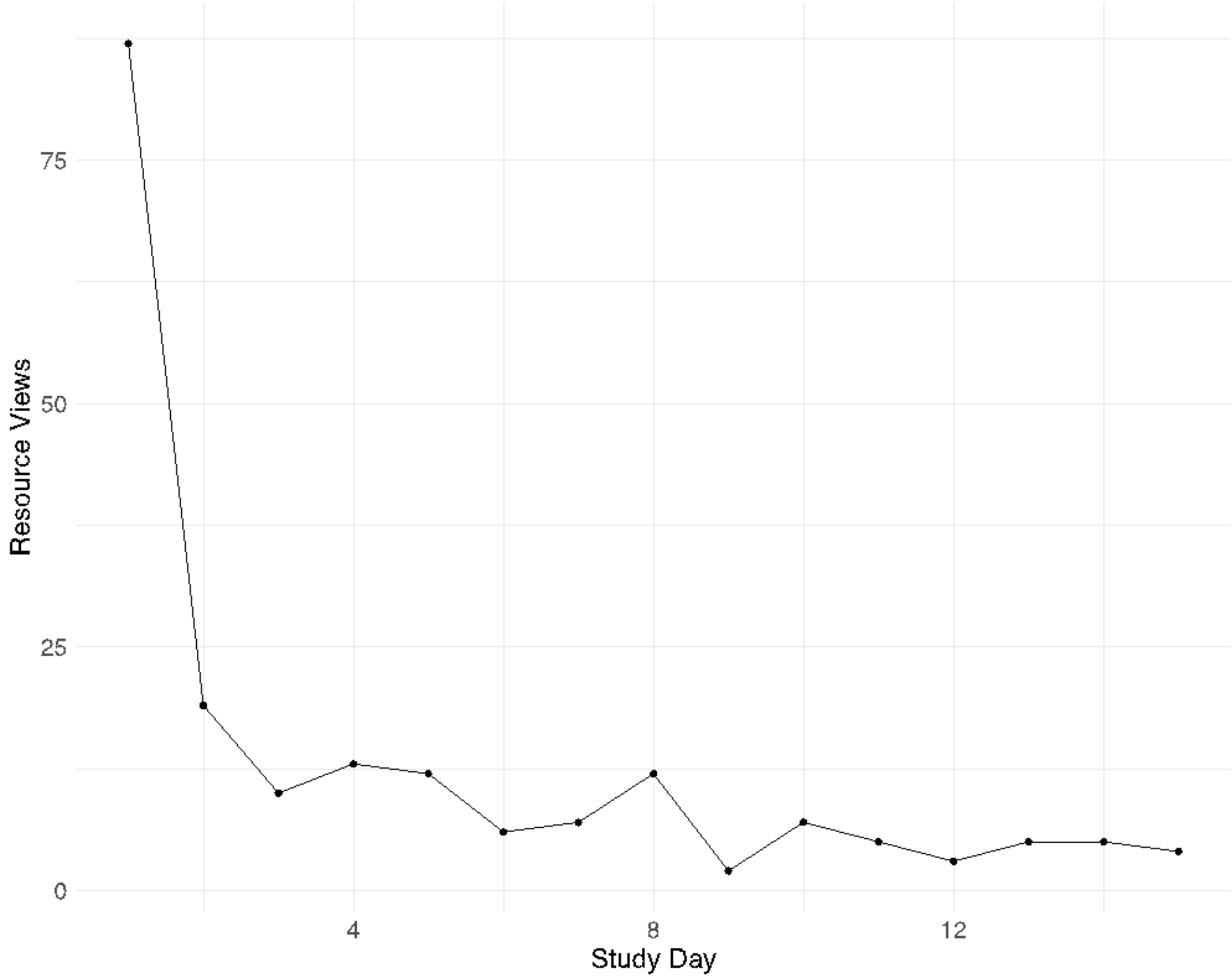
Do People Contact Crisis Resources After Engaging With ResourceBot?

In the ResourceBot follow-up survey, 0 participants reported using crisis resources. Participants could endorse multiple reasons for not using resources in the follow-up survey. The frequency of the reasons for not using the resources was as follows: did not think it would help (76/165, 46%), I handled it on my own (70/165, 42%), too much time and effort (33/165, 20%), felt better without them (24/165, 15%), and other (5/165, 3%). Two times participants did not answer this question.

In the daily survey, there were 3 total times participants reported using crisis resources that day from 3 separate participants. One participant reported calling a crisis line, one reported attending group therapy, and one did not remember the exact resource they used. The helpfulness ratings for 3 instances were 0, 4, and 5 (out of 5).

The crisis resources survey (ie, the constantly available list of resources that participants could open) was opened and submitted a total of 312 times in the survey across 59 participants. A total of 113 times participants opened the survey and skipped out without selecting a resource to view, we therefore report on the remaining 199 times participants selected a resource to view. The frequency of viewing by time in the study is shown in Figure 4. In Figure 4, day 1 refers to the day a participant first logged into the Metricwire smartphone app and day 2 refers to the first day of smartphone surveys. The frequency of viewing for each type of resource was the following (Figure 5): Crisis Text Line (69/197, 35%), National Suicide Prevention Lifeline (58/197, 29%), Trevor Project Chat (29/197, 15%), Trans Lifeline (17/197, 9%), Veteran's Crisis Line Chat (13/197, 6%), and Veteran's Crisis Line (13/197, 6%).

Figure 5. Number of views of resources by day in study. Study day 1 refers to the day a participant first logged into the Metricwire smartphone app and day 2 refers to the first day of smartphone surveys.



What Do People Report About the Acceptability and Usability of ResourceBot?

The mean for the CSQ was 21 (out of 32) and the SD was 5.96. The mean for the SUS was 76.7 (out of 100) and the SD was 17.06. The mean and SD for all items on both scales are provided in Multimedia Appendix 1. For the comfort with answering questions honestly item, the average rating was 4.30 (out of 5).

Discussion

The aims of the current project were to adapt an evidence-based BRI into a new tool for smartphone-based delivery

(ResourceBot) that aims to increase the use of crisis resources and to test the feasibility, acceptability, and utility of this tool in real time monitoring. There were 3 key findings from this study. First, it is feasible to develop and deploy a real-time resource tool. Second, there was low use of crisis resources overall (including those specifically associated with ResourceBot) in the study. Third, participants rated the ResourceBot with moderate satisfaction and good usability. Each of these findings warrants further comment.

We found that it is possible to build, deploy, and receive high engagement with a real-time crisis resource tool. Much has been written about the promise of smartphone technologies for suicide research and intervention [11,27,28]. Although a plethora of mobile apps exist for suicide prevention [29], little systematic research to date has examined mobile interventions for suicide prevention [30]. This study found that in a severe sample (ie, recent suicidal thoughts, elevated symptoms of depression and anxiety), most of the time participants engaged with the ResourceBot and provided helpful data on their use of the tool. In this paper, we intentionally used a simple decision rule (eg, if suicidal intent is greater than 1 then send ResourceBot) to increase the feasibility and interpretability of findings. To promote greater engagement, future research could increase the complexity in 2 ways. First, the decision rule for the tool could be more adaptive and tailored to the individual, for example, deploying the tool based on participant's deviations from their own within person average level of suicidal intent or a rule that incorporates additional variables beyond just suicidal intent. Second, a greater number of messages with distinct content could be used in the tool. The barriers and psychoeducation messages were static in this study, which could have resulted in habituation to the ResourceBot and low use of resources. Therefore, a broader more dynamic message base to draw on may promote greater engagement over time in future work. In short, this study found it is feasible to deliver tools for participant safety immediately after participants complete real-time monitoring surveys. Future work can expand upon what type of tools are offered in that immediate moment after a participant has reported suicidal thoughts.

The second finding of this work is that no participants reported using crisis hotlines in the follow-up survey after the ResourceBot. This finding was counter to our expectations given prior work finding that a BRI can increase the self-reported use of crisis resources [9]. Our work highlights how in digital health research one cannot assume findings from one intervention context necessarily translate to another context. There are numerous reasons why there may be differences between past work and this study. This includes differences in the trigger for the BRI and the participants in the studies. For example, prior work was done with naturalistic users of social media platforms and this work was done with EMA study participants. Most participants in this study had a lifetime history of mental health care and many had previously used crisis hotlines. At baseline, participants reported on average feeling unlikely to use crisis hotlines in the future and therefore may have been more resistant to the BRI. Past work has also been with larger samples (ie, hundreds of participants), and we may have seen more participants use hotlines with a larger sample size.

More broadly, given the increasing role of crisis hotlines in national suicide prevention efforts [31] and suicide research safety protocols [5], this study highlights the need to continue to understand participants' concerns about and experiences with these hotlines. A recent nationally representative survey found that about 5% (23/388) of participants with serious distress had used the 988 Lifeline and only 29% (7/23) of those participants with serious distress reported being very likely to use it in the future [32]. This work suggests the skepticism of crisis hotlines (eg, beliefs that they won't help) in this study are not unique and perhaps a need to offer a broader range of resources in future work. For example, providing suggestions of coping skills (eg, distraction and relaxation) from interventions such as the safety planning intervention may be incrementally useful [33,34].

Finally, participants rated the ResourceBot with good satisfaction and usability. According to one normative rating scale of the SUS [35,36], the average score in our study for ResourceBot would get a grade of a "B." According to another rating system [37], it would be considered "good." These findings provide further support for the feasibility of the real-time deployment of suicide prevention tools. To our knowledge, publicly available norms for the CSQ are not available. Using a transformation suggested by the scale developers to put the score on a 25 to 100 scale where one multiplies the original total score by 3.125, would produce an average score of 65.6 for the ResourceBot. This suggests significant room for improvement with the ResourceBot tool. It is possible that participants' skepticism of crisis lines influenced their satisfaction with ResourceBot given that the tool promoted the use of these crisis lines. In the future, offering a broader type of message, resources, and skills may increase satisfaction with the tool. Another important finding from the exit survey is that participants reported on average being very comfortable answering questions honestly in the study. Examination of temporal trends in the intent to kill oneself scores also showed no changes in severity by day in the study. If participants were trying to avoid triggering the ResourceBot tool, one may expect to see lower intent scores toward the end of the study period and we do not see this. Both the exit survey honesty ratings and the lack of temporal trends in intent scores, suggest a lack of reactivity to the ResourceBot tool, which is a concern with real-time interventions for suicide prevention [6]. The lack of reactivity could also be due, at least in part, to the anonymous nature of the study [38] and the clear language in the consent form regarding active risk monitoring. The structure of the study and the ResourceBot could have contributed to participants feeling more comfortable disclosing suicidal thoughts [39].

This study provides new information on real-time risk management and crisis resource use but still has important limitations that warrant discussion. First, the current sample was a convenience sample recruited through a web-based research platform. It is unclear how the results would generalize to a clinical sample. Second, the threshold used to deploy the ResourceBot tool was relatively low compared with thresholds used in past research [6]. It is possible that participants did not use the resources because they did not consider their own current suicidal thoughts severe enough to warrant reaching out to a crisis hotline. To try to limit overwhelming participants with the ResourceBot, we implemented a stop rule so that the ResourceBot was only deployed up to once a day. Without this stop rule, the deployment rate would have been 939 rather than 221. Nevertheless, most participants who were sent the ResourceBot in the current study were sent it multiple times. As shown in Figure 4, the average levels of suicidal intent were relatively low, which is one reason why the lower threshold was used. These issues of severity and frequency highlight the challenge of selecting an appropriate threshold of suicidal intent. The engagement with the ResourceBot tool and use of the crisis resources may have been different if a higher threshold was used. Third, the ResourceBot deployment was contingent on compliance with the real-time survey and it is possible that participants may be less likely to fill out a survey when they are experiencing higher levels of distress. More work is needed to better understand compliance in real-time monitoring studies and the best way to incentivize compliance [40]. Fourth, the compliance rate for the exit survey (59.4%) was relatively low, which could have potentially biased the exit survey results. Finally, this study can provide information on the feasibility and usability of automated tools, but cannot fully speak to the ethics of automated interventions versus clinician outreach interventions [4,41]. In this study, the Prolific platform requires that participants maintain anonymity and therefore clinician contact in this setting would not be possible. Decisions related to the type and timing of real-time interventions depend upon the context of the study as well as discussions with ethics boards and regulatory bodies [4,42].

Future studies could build upon this study in multiple ways. First, this study only offered crisis lines as resources. Future work could offer more types of resources in this automated tool format, such as reminders or skills for coping with suicidal thoughts. Second, this study is focused on feasibility and acceptability and no randomization was used. Future work could consider a between-participants randomized control trial where different groups were provided with an automated interactive intervention or pop-up reminders. Future work could also attempt to use a within-person micro-randomized trial design [43] where participants are randomized to different types of automated tools at different levels of suicidal thinking [11]. This type of trial design could allow future studies to empirically test the effect of presenting different kinds of resources on future resource use or self-reported momentary suicidal thinking. Finally, this work highlights the immense complexity and challenge of building ethical and effective real-time interventions for suicide prevention. Future work could use focus groups and other qualitative methods from individuals with lived experience to better understand and develop tools that would be the most helpful to people during moments of elevated suicide risk [44].

Mobile technologies have the potential to advance the understanding of suicide and contribute to new suicide prevention approaches. These technologies, however, present immense ethical challenges in which researchers grapple with both collecting helpful data and preserving participant safety. This study highlights the nuance of this issue and the need for the rigorous development of real-time safety tools.

Acknowledgments

This study was funded by the APF/The Trust Grant in Honor of Eric A. Harris, EdD, JD. DDLC is funded by the National Science Foundation Graduate Research Fellowship (grant DGE-1745303) and the National Institute of Mental Health (grant F31MH130055). KHB is supported by the National Institute of Mental Health (grant K23MH120436). MKN is funded by grants from the National Institute of Mental Health (P50MH129699), the Chet and Will Griswold Fund, and a gift from the Fuss Family Research Fund. This content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation or the National Institutes of Health. The authors wish to thank all participants for their participation and valuable feedback. The authors also wish to thank Kayla Edwards for her help with the study and Robert Morris for feedback on the ResourceBot.

Conflicts of Interest

MKN receives publication royalties from Macmillan, Pearson, and UpToDate. He has been a paid consultant in the past 3 years for Microsoft Corporation, the Veterans Health Administration, and COMPASS Pathways, and for legal cases regarding a death by suicide. He has stock options in Cerebral Inc. He is an unpaid scientific advisor for Empatica, Koko, and TalkLife. EMK has been a paid consultant in the past 3 years for Boehringer Ingelheim Pharmaceuticals.

Multimedia Appendix 1
Images of the ResourceBot and the decision logic and mean (SD) for all items on both scales.
[DOCX File , 3398 KB - mental_v11i1e58409_app1.docx]

References

1. Kleiman EM, Nock MK. Real-time assessment of suicidal thoughts and behaviors. Curr Opin Psychol 2018;22:33-37. [doi: 10.1016/j.copsyc.2017.07.026] [Medline: 30122275]
2. Millner AJ, Robinaugh DJ, Nock MK. Advancing the understanding of suicide: the need for formal theory and rigorous descriptive research. Trends Cogn Sci 2020;24(9):704-716 [FREE Full text] [doi: 10.1016/j.tics.2020.06.007] [Medline: 32680678]

3. Kivelä L, van der Does WAJ, Riese H, Antypa N. Don't miss the moment: a systematic review of ecological momentary assessment in suicide research. Front Digit Health 2022:4 [FREE Full text] [doi: 10.3389/fdgth.2022.876595]
4. Nock MK, Kleiman EM, Abraham M, Bentley KH, Brent DA, Buonopane RJ, et al. PRCP 2021;3(2):57-66 [FREE Full text] [doi: 10.1176/appi.prcp.20200029] [Medline: 34414359]
5. Bentley KH, Maimone JS, Kilbury EN, Tate MS, Wisniewski H, Levine MT, et al. Practices for monitoring and responding to incoming data on self-injurious thoughts and behaviors in intensive longitudinal studies: a systematic review. Clin Psychol Rev 2021;90:102098 [FREE Full text] [doi: 10.1016/j.cpr.2021.102098] [Medline: 34763126]
6. Bentley KH, Millner AJ, Bear A, Follet L, Fortgang RG, Zuromski KL, et al. Intervening on high-risk responses during ecological momentary assessment of suicidal thoughts: is there an effect on study data? Psychol Assess 2024;36(1):66-80. [doi: 10.1037/pas0001288] [Medline: 37917497]
7. Kleiman EM, Glenn CR, Liu RT. The use of advanced technology and statistical methods to predict and prevent suicide. Nat Rev Psychol 2023;2(6):347-359 [FREE Full text] [doi: 10.1038/s44159-023-00175-y] [Medline: 37588775]
8. Blease C, Torous J. ChatGPT and mental healthcare: balancing benefits with risks of harms. BMJ Ment Health 2023;26(1):e300884 [FREE Full text] [doi: 10.1136/bmjment-2023-300884] [Medline: 37949485]
9. Jaroszewski AC, Morris RR, Nock MK. Randomized controlled trial of an online machine learning-driven risk assessment and intervention platform for increasing the use of crisis services. J Consult Clin Psychol 2019;87(4):370-379. [doi: 10.1037/ccp0000389] [Medline: 30883164]
10. Cohen K, Dobias M, Morris R, Schleider J. Improving uptake of mental health crisis resources: randomized test of a single-session intervention embedded in social media. J Behav Cogn Ther 2023;33(1):24-34. [doi: 10.1016/j.jbct.2022.12.001]
11. Coppersmith DDL, Dempsey W, Kleiman EM, Bentley KH, Murphy SA, Nock MK. Just-in-time adaptive interventions for suicide prevention: promise, challenges, and future directions. Psychiatry 2022;85(4):317-333 [FREE Full text] [doi: 10.1080/00332747.2022.2092828] [Medline: 35848800]
12. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018;52(6):446-462 [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
13. Nahum-Shani I, Shaw SD, Carpenter SM, Murphy SA, Yoon C. Engagement in digital interventions. Am Psychol 2022;77(7):836-852 [FREE Full text] [doi: 10.1037/amp0000983] [Medline: 35298199]
14. Douglas BD, Ewell PJ, Brauer M. Data quality in online human-subjects research: comparisons between MTurk, prolific, CloudResearch, qualtrics, and SONA. PLoS One 2023;18(3):e0279720 [FREE Full text] [doi: 10.1371/journal.pone.0279720] [Medline: 36917576]
15. Peer E, Rothschild D, Gordon A, Evernden Z, Damer E. Data quality of platforms and panels for online behavioral research. Behav Res Methods 2022;54(4):1643-1662 [FREE Full text] [doi: 10.3758/s13428-021-01694-3] [Medline: 34590289]
16. Smith DMY, Lipson SM, Wang SB, Fox KR. Online methods in adolescent self-injury research: challenges and recommendations. J Clin Child Adolesc Psychol 2022;51(2):143-154. [doi: 10.1080/15374416.2021.1875325] [Medline: 33656941]
17. Fox KR, Harris JA, Wang SB, Millner AJ, Deming CA, Nock MK. Self-injurious thoughts and behaviors interview-revised: development, reliability, and validity. Psychol Assess 2020;32(7):677-689. [doi: 10.1037/pas0000819] [Medline: 32324021]
18. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
19. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
20. Bentley KH, Coppersmith DL, Kleiman EM, Nook EC, Mair P, Millner AJ, et al. Do patterns and types of negative affect during hospitalization predict short-term post-discharge suicidal thoughts and behaviors? Affect Sci 2021;2(4):484-494 [FREE Full text] [doi: 10.1007/s42761-021-00058-6] [Medline: 35465415]
21. Coppersmith DDL, Ryan O, Fortgang RG, Millner AJ, Kleiman EM, Nock MK. Mapping the timescale of suicidal thinking. Proc Natl Acad Sci U S A 2023;120(17):e2215434120 [FREE Full text] [doi: 10.1073/pnas.2215434120] [Medline: 37071683]
22. Wang SB, Coppersmith DDL, Kleiman EM, Bentley KH, Millner AJ, Fortgang R, et al. A pilot study using frequent inpatient assessments of suicidal thinking to predict short-term postdischarge suicidal behavior. JAMA Netw Open 2021;4(3):e210591 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.0591] [Medline: 33687442]
23. Larsen DL, Attkisson CC, Hargreaves WA, Nguyen TD. Assessment of client/patient satisfaction: development of a general scale. Eval Program Plann 1979;2(3):197-207. [doi: 10.1016/0149-7189(79)90094-6] [Medline: 10245370]
24. Lewis JR, Sauro J. The factor structure of the system usability scale. In: Kurosu M, editor. Human Centered Design. Lecture Notes in Computer Science. San Diego, CA, USA: Springer; 2009:94-103.
25. Short CE, DeSmet A, Woods C, Williams SL, Maher C, Middelweerd A, et al. Measuring engagement in eHealth and mHealth behavior change interventions: viewpoint of methodologies. J Med Internet Res 2018;20(11):e292 [FREE Full text] [doi: 10.2196/jmir.9397] [Medline: 30446482]
26. ResourceBot: developing and testing an automated real-time tool for promoting crisis resource use for suicide risk. OSF HOME. URL: https://osf.io/yxmdg/?view_only=adb42a1907eb46638fba1ff9032a0ac4 [accessed 2024-09-26]

27. Allen NB, Nelson BW, Brent D, Auerbach RP. Short-term prediction of suicidal thoughts and behaviors in adolescents: can recent developments in technology and computational science provide a breakthrough? J Affect Disord 2019;250:163-169 [FREE Full text] [doi: 10.1016/j.jad.2019.03.044] [Medline: 30856493]
28. Torous J, Larsen ME, Depp C, Cosco TD, Barnett I, Nock MK, et al. Smartphones, sensors, and machine learning to advance real-time prediction and interventions for suicide prevention: a review of current progress and next steps. Curr Psychiatry Rep 2018;20(7):51. [doi: 10.1007/s11920-018-0914-y] [Medline: 29956120]
29. Wilks CR, Chu C, Sim D, Lovell J, Gutierrez P, Joiner T, et al. User engagement and usability of suicide prevention apps: systematic search in app stores and content analysis. JMIR Form Res 2021;5(7):e27018 [FREE Full text] [doi: 10.2196/27018] [Medline: 34259163]
30. Jiménez-Muñoz L, Peñuelas-Calvo I, Díaz-Oliván I, Gutiérrez-Rojas L, Baca-García E, Porras-Segovia A. Suicide prevention in your pocket: a systematic review of ecological momentary interventions for the management of suicidal thoughts and behaviors. Harv Rev Psychiatry 2022;30(2):85-99. [doi: 10.1097/HRP.0000000000000331] [Medline: 35267251]
31. Miller AB, Oppenheimer CW, Glenn CR, Yaros AC. Preliminary research priorities for factors influencing individual outcomes for users of the US national suicide prevention Lifeline. JAMA Psychiatry 2022;79(12):1225-1231. [doi: 10.1001/jamapsychiatry.2022.3270] [Medline: 36223084]
32. Purtle J, McSorley AMM, Adera AL, Lindsey MA. Use, potential use, and awareness of the 988 suicide and crisis lifeline by level of psychological distress. JAMA Netw Open 2023;6(10):e2341383 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.41383] [Medline: 37906197]
33. Stanley B, Martínez-Alés G, Gratch I, Rizk M, Galfalvy H, Choo TH, et al. Coping strategies that reduce suicidal ideation: an ecological momentary assessment study. J Psychiatr Res 2021;133:32-37 [FREE Full text] [doi: 10.1016/j.jpsychires.2020.12.012] [Medline: 33307352]
34. Stanley B, Brown GK. Safety planning intervention: a brief intervention to mitigate suicide risk. Cogn Behav Ther 2012;19(2):256-264. [doi: 10.1016/j.cbpra.2011.01.001]
35. Lewis JR. The system usability scale: past, present, and future. Int J Hum Comput Interact 2018;34(7):577-590. [doi: 10.1080/10447318.2018.1455307]
36. Sauro J, Lewis JR. Quantifying the User Experience: Practical Statistics for User Research. United States: Morgan Kaufmann; 2016.
37. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud 2009;4(3):114-123.
38. Deming CA, Harris JA, Castro-Ramirez F, Glenn JJ, Cha CB, Millner AJ, et al. Inconsistencies in self-reports of suicidal ideation and attempts across assessment methods. Psychol Assess 2021;33(3):218-229. [doi: 10.1037/pas0000976] [Medline: 33705163]
39. Hallford DJ, Rusanov D, Winestone B, Kaplan R, Fuller-Tyszkiewicz M, Melvin G. Disclosure of suicidal ideation and behaviours: a systematic review and meta-analysis of prevalence. Clin Psychol Rev 2023;101:102272. [doi: 10.1016/j.cpr.2023.102272] [Medline: 37001469]
40. Wrzus C, Neubauer AB. Ecological momentary assessment: a meta-analysis on designs, samples, and compliance across research fields. Assessment 2023;30(3):825-846 [FREE Full text] [doi: 10.1177/10731911211067538] [Medline: 35016567]
41. Pearson JL, Stanley B, King CA, Fisher CB. Intervention research with persons at high risk for suicidality: safety and ethical considerations. J Clin Psychiatry 2001;62 Suppl 25:17-26. [Medline: 11765091]
42. Hom MA, Podlogar MC, Stanley IH, Joiner TE. Ethical issues and practical challenges in suicide research. Crisis 2017;38(2):107-114. [doi: 10.1027/0227-5910/a000415] [Medline: 27561221]
43. Klasnja P, Hekler EB, Shiffman S, Boruvka A, Almirall D, Tewari A, et al. Microrandomized trials: an experimental design for developing just-in-time adaptive interventions. Health Psychol 2015;34S:1220-1228 [FREE Full text] [doi: 10.1037/hea0000305] [Medline: 26651463]
44. Hom MA, Bauer BW, Stanley IH, Boffa JW, Stage DL, Capron DW, et al. Suicide attempt survivors' recommendations for improving mental health treatment for attempt survivors. Psychol Serv 2021;18(3):365-376. [doi: 10.1037/ser0000415] [Medline: 32068415]

Abbreviations

BRI: barrier reduction intervention
CSQ: client satisfaction questionnaire
EMA: ecological momentary assessment
GAD-7: Generalized Anxiety Disorder 7-item
PHQ-9: Patient Health Questionnaire-9
SITBI-R: Self-Injurious Thoughts and Behaviors Interview—Revised
STB: suicidal thoughts and behavior
SUS: System Usability Scale

Edited by J Torous; submitted 14.03.24; peer-reviewed by A Horwitz, U Leong; comments to author 21.03.24; revised version received 05.08.24; accepted 16.08.24; published 31.10.24.

Please cite as:
Coppersmith DDL, Bentley KH, Kleiman EM, Jaroszewski AC, Daniel M, Nock MK
Automated Real-Time Tool for Promoting Crisis Resource Use for Suicide Risk (ResourceBot): Development and Usability Study
JMIR Ment Health 2024;11:e58409
URL: https://mental.jmir.org/2024/1/e58409
doi:10.2196/58409
PMID:

Technologies for Supporting Individuals and Caregivers Living With Fetal Alcohol Spectrum Disorder: Scoping Review

Joanna Ting Wai Chu[1,2,3,4], BA, BSc (Hon), MSc, PhD; Holly Wilson[1,4], BSoSci (Hon), MSc; Cynthia Zhiyin Cai[1,4], BSc, MEnvMgt; Jessica C McCormack[1,5], BA/LLB (Hons), MA, PhD; David Newcombe[3,4], BA (Hon), PhD; Chris Bullen[1], MBChB, MPH, PhD

1
2
3
4
5

Corresponding Author:
Joanna Ting Wai Chu, BA, BSc (Hon), MSc, PhD

Abstract

Background: Fetal alcohol spectrum disorder (FASD) is a common developmental disability that requires lifelong and ongoing support but is often difficult to find due to the lack of trained professionals, funding, and support available. Technology could provide cost-effective, accessible, and effective support to those living with FASD and their caregivers.

Objective: In this review, we aimed to explore the use of technology available for supporting people living with FASD and their caregivers.

Methods: We conducted a scoping review to identify studies that included technology for people with FASD or their caregivers; focused on FASD; used an empirical study design; were published since 2005; and used technology for assessment, diagnosis, monitoring, or support for people with FASD. We searched MEDLINE, Web of Science, Scopus, Embase, APA PsycINFO, ACM Digital Library, JMIR Publications journals, the Cochrane Library, EBSCOhost, IEEE, study references, and gray literature to find studies. Searches were conducted in November 2022 and updated in January 2024. Two reviewers (CZC and HW) independently completed study selection and data extraction.

Results: In total, 17 studies exploring technology available for people with FASD showed that technology could be effective at teaching skills, supporting caregivers, and helping people with FASD develop skills.

Conclusions: Technology could provide support for people affected by FASD; however, currently there is limited technology available, and the potential benefits are largely unexplored.

Trial Registration: PROSPERO CRD42022364885; https://tinyurl.com/3zaatu9u

(JMIR Ment Health 2024;11:e51074) doi:10.2196/51074

KEYWORDS

fetal alcohol; scoping review; technology; caregivers; diagnosis; support; intervention; fetal alcohol spectrum disorder; FASD; developmental disability; lifelong support; caregiver; accessibility; alcohol; alcohol intake; pregnant substance; pregnant; fetal; PRISMA; Preferred Reporting Items for Systematic Reviews and Meta-Analyses; mobile phone

Introduction

Background

Fetal alcohol spectrum disorder (FASD) is a diagnostic term that describes the effects of prenatal alcohol exposure (PAE) on an individual [1]. Globally, it is estimated that approximately 7.7 of every 1000 births are affected by FASD [2]. People with FASD may experience restricted growth, diminished neurological and cognitive functioning, and behavioral problems [1]. These symptoms result in common disabilities, such as difficulties with memory, language, cognition, executive function, social skills, and attention [1]. People living with FASD face challenges with peer relationships, education, employment, independence [3,4], and mental health, including higher rates of suicide [5] and lower life expectancy [6].

Early interventions [3], including assessment, diagnosis, and effective support from a young age, are key to favorable outcomes [7]. Without such interventions, people with FASD are more likely to experience difficulties throughout their lifetime, such as school failure, poor mental health, and substance dependency [5,8]. However, accessing early interventions is difficult [9], and people can experience lengthy delays waiting for a diagnosis and support. Several factors contribute to these difficulties including a lack of FASD-informed and trained health professionals [10], lack of FASD-informed support, and lack of effective interventions available for those with FASD or their caregivers [11]. Technology could offer a low-cost and effective support to people with FASD.

The term *technology* encompasses a range of technologies; however, in this review, we limited technology to the delivery of support through virtual reality (VR), computer-based interventions, smartphone apps, artificial intelligence (AI), telehealth, computer games, wearable technology, or technologies that facilitate therapy or support. We use the National Institute for Health and Care Research (NIHR) [12] definition of technologies that include VR assessment or therapy, digital technologies, telehealth, computer-based assessment or therapies, real-time monitoring and wearable devices, smartphone apps, and sensors that can improve patient outcomes and health service efficiencies.

Innovations in technology hold promise to drive efficiencies, improve outcomes, and widen access to health care delivery [13]. Technology can support people's health including aiding diagnosis and delivering effective support. Technology has improved diagnosis by facilitating the detection of neurocognitive impairments experienced by those with neurodevelopmental disorders. Technology has aided the screening of brain anatomy and activity and the screening of social and motor skill deficits [14,15]. Likewise, telehealth can provide specialist expertise to individuals living in remote communities [15,16], such as delivering FASD assessments to people living in remote Canada [17]. Digital interventions can support peoples' mental health [18,19], while for those with intellectual disabilities, digital interventions can support well-being [20] and the development of key skills [21]. VR-delivered interventions can provide favorable outcomes for mental illness [18,19,22,23] and pain management [24] and improve emotional recognition for people with autism spectrum disorder (ASD) [25] and cognitive skills for those with attention-deficit/hyperactivity disorder (ADHD) [26]. These technologies have been found to be capable of delivering automated and self-directed support, at low cost without burdening health professionals, to patients and families with a range of conditions and disabilities [14,19]. Technology has been well received by people with intellectual disabilities [16,20,21,27], but the content needs to be adapted to the skills and needs of the intended population [27].

Despite the increasing adoption of technological advances in health care delivery, research on the application of such technologies used to support or assist people with FASD is sparse. A systematic review conducted on the implementation of technologies to assess, monitor, and treat neurodevelopmental disorders did not report on any FASD-related studies [28]. To date, no study has been conducted to review available technologies to support people living with FASD. Much of this research has focused on digital interventions that can reduce PAE rather than support for FASD [29]. A recent systematic review of screening tools for FASD identified several screening tools to support clinicians to diagnose FASD or fetal alcohol syndrome (FAS) [17], and telehealth has been explored for people with FASD [30]. However, the focus of the review was not on technology-based tools alone [17].

Objectives

We conducted a scoping review to explore what technologies have been applied or implemented to support caregivers and people with FASD with assessment, monitoring, and support and to identify current gaps and limitations within the existing evidence base. Scoping reviews are particularly relevant to disciplines with emerging evidence where the existence of few randomized controlled trials makes it difficult for researchers to conduct systematic reviews. Another strength is that scoping reviews can include a range of study designs, in both published and gray literature, addressing questions beyond those related to intervention effectiveness [31].

Methods

Protocol and Registration

We conducted a scoping review following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) [32]. The review protocol was registered on PROSPERO (CRD42022364885).

Eligibility Criteria

Studies were eligible if they met the following criteria: technology (modified and adapted from NIHR [28]) was intended for people living with FASD or their caregivers; focused on FASD; used an empirical study design (qualitative or quantitative); were published after 2005 (the year when the first FASD diagnostic criteria were made available); included digital technology designed to aid the assessment, diagnosis, monitoring, or support for people with FASD and involved direct interaction from the person with FASD or their caregiver; and were published in the English language. We excluded technology that did not require direct interaction by the person with FASD or their caregiver, such as tools designed to assist professionals; that were noninteractive and provided only written material, such as noninteractive websites; and if the focus was not FASD.

Information Sources and Search Strategy

Between November 14 and November 30, 2022, we searched the literature to identify studies that examined technology to support people with FASD. We searched the following databases: MEDLINE, Web of Science, Scopus, Embase, APA PsycINFO, ACM Digital Library, JMIR Publications journals, the Cochrane Library, EBSCOhost, and IEEE. To find gray literature and publicly available technology, we searched the ProQuest Dissertations and Theses database, PROSPERO, ClinicalTrials.gov, World Health Organization, International Clinical Trials Registry Platform, Google Scholar (first 100 results), Apple App Store, and Google Play Store. Searches were rerun in April 2024. Of the studies that were included in the final analysis, we searched the reference list and identified no new references. The search terms combined FASD and technology-related terms (see Multimedia Appendix 1 for an example of a search strategy). For the gray literature, the search terms were modified when it was not possible to combine terms.

Study Selection

Search results were extracted, and then duplicates were removed. Then, the title and abstract were independently screened based on the inclusion criteria by 2 reviewers (CZC and HW). Following this, we obtained the full text of all possible papers, and they were independently screened by 2 reviewers (CZC and HW).

Data Charting

Data were extracted to a bespoke Microsoft Excel (Microsoft Corp) spreadsheet; one author (CZC) extracted data for all studies, and a random sample of 10% of papers was reviewed by another author (HW) to check reliability. The data extracted included author, date, full citation, country, number and characteristics of participants, description of technology, research approach, methods, data analysis technique, and main findings. Any disagreement between reviewers was resolved by consensus.

Critical Appraisal of Individual Sources of Evidence

The methodological quality of each study was accessed using the Mixed Methods Appraisal Tool (MMAT) [33,34]. We chose to determine the quality of each study to explore the quality of literature, and all studies were included in the final analysis regardless of their quality. We used the MMAT, as the quality of different study designs can be assessed with the MMAT. The MMAT determines the quality of each study with two steps. First, each paper is screened to ensure the MMAT can be applied. Then, based on the study type, several questions determine the quality of the studies' research approach, methodology, and results based on each criterion for different study types. Each study has 5 quality questions applied to determine the quality of each study, and no overall score is given, but rather each response is considered. Two authors (CZC and HW) independently applied the MMAT to each study, and any disagreements were resolved through discussion with all authors.

Synthesis of Results

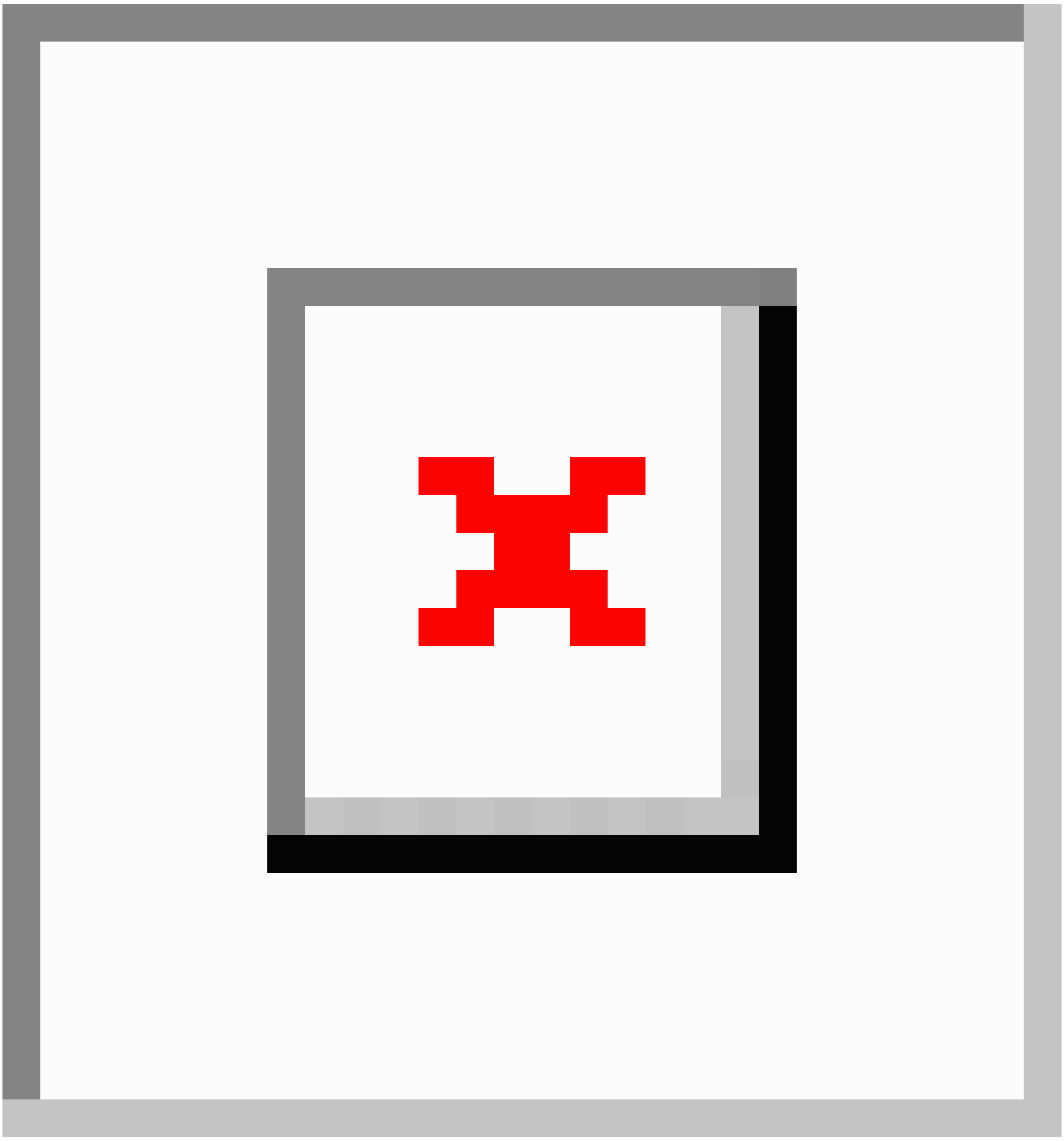
To synthesize technology available to support those with FASD, a narrative synthesis was conducted [35]. The main findings from each study were exported into NVivo (QSR International), where they were grouped by common technology for assessment, diagnosis, and support for FASD. A narrative synthesis was conducted, as it enables the inclusion of qualitative and quantitative findings and provides an overview of the existing technology available to support people with FASD [35].

Results

Selection of Sources of Evidence

Our search produced 676 results. After duplicates were removed (n=75) and criteria were applied, 17 studies met the inclusion criteria and were included in this review (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of included studies. FASD: fetal alcohol spectrum disorder.



Characteristics of Sources of Evidence

In total, 17 studies [36-52] were included in this review. All studies explored the experiences of stakeholders with technology, but no studies included publicly available apps or technology. Of the 17 included studies, 15 studies [36-41,43-47,49-52] included participants; the remainder were study protocols. In total, 231 caregivers of people with FASD and 168 children participated. Of the 168 children, 157 (94%) had FASD, and 11 (6%) were typically developing. All children who participated were 18 years of age or younger, and of these, 97 (58%) were younger than 10 years of age. Most studies were conducted in the United States (n=10, 59%) [36,37,40,41,43-47,52], Canada (n=3, 18%) [38,39,48], and Australia (n=2, 12%) [49,50], with 1 conducted in South Africa (n=1, 6%) [42] and the United Kingdom (n=1, 6%) [51]. The included studies can be classified as those that evaluated the efficacy of a technology on an outcome (n=8, 47%) [36,37,40,41,43-45,52], focused on the design or implementation (n=7, 41%) [38,39,46,47,49-51], were study protocols (n=2, 12%) [42,48], and had no published follow-up studies. Of these studies that evaluated the efficacy (n=8, 47%), the study designs used were randomized controlled trials (n=5, 63%) [37,40,41,43,52], case studies (n=1, 12%) [45], and pre-post study designs (n=2, 25%) [36,44]. The type of technologies in these studies included web-based interactive platforms (n=3, 18%), VR-based technology (n=4, 23%), computer games (n=4, 23%), apps (n=2, 12%), and interventions delivered via technology (n=4, 23%).

Critical Appraisal Within Sources of Evidence

The MMAT was applied to 15 of the 17 studies. The tool could not be applied to 2 because they were study protocols [42,48]. The results of the quality assessment are shown in Multimedia Appendix 2 [36-52]. Most (n=10, 71%) [36,38-41,44-47,49] of the studies meet all criteria. While 3 (18%) [37,43,52] studies only meet 2 or 3 of the criteria as the methods lacked sufficient information to determine the quality of the studies.

Results of Individual Sources of Evidence

In total, 17 studies were included in this review, and 2 of these studies were study protocols (see Multimedia Appendix 3 [36-52] for characteristics of individual studies).

Synthesis of Results

Overview

Our narrative analysis identified 2 main themes (assessment and diagnosis of FASD and support) and 2 subthemes within support (caregiver support and skill development). The majority of technology was designed to support people with FASD or their caregivers. Most studies focused on support, with only 1 study exploring both assessment and support.

Assessment and Diagnosis of FASD

Only 1 study explored the experiences of caregivers with technology to support the assessment and diagnosis of FASD. This study explored the perceptions of families of the use of telehealth-assisted assessment and diagnosis of FASD in remote communities in Manitoba, Canada [38]. The birth and adoptive caregivers who participated in the interviews were satisfied with the use of telehealth to aid the assessment processes and reduced barriers to access.

Support

Overview

All other studies focused on supporting people with FASD or their caregiver's following diagnosis of FASD. Two main subthemes were identified: (1) caregiver support, where technology was designed to support the caregivers of people with FASD, and (2) skill development, where technology aimed to develop skills for people living with FASD.

Caregiver Support

Technology designed to support caregivers of people with FASD was used in 8 studies [39,40,46-51]. These studies explored 3 distinct technologies. First, Families Moving Forward Connect [46,47], an app designed to educate and support caregivers, provided resources and opportunities to connect with others with experience with FASD. These studies explored the design and feasibility of the app and found the app was received positively by caregivers, including the content of the app and the connection to others, but some participants reported difficulty with using the app.

Second, Strongest Families FASD [39,48], a web-based interactive program, was designed to promote parenting and reduce distress in caregivers through video or audio clips and interactive activities. One study explored the usability of the website through the experiences of biological and adoptive caregivers, who found the program to be easy to use and the content relevant [39]. The second study was a protocol study with no currently available outcomes [48]. Finally, a study explored the efficacy of web-based, in-person, and information-only formats in supporting the knowledge and behavioral regulation in caregivers of people diagnosed with FAS and partial FAS [40]. All 3 formats (online, in-person, and information) showed improvements in knowledge of behavioral regulation, but the online format showed no improvements in actual behaviors when the in-person and information formats did.

Finally, educational training programs delivered online were designed to support caregivers. The Salford Parents and Carers Education Course for Improvements in FASD Outcomes in Children (SPECIFiC) [51] is a 7-session educational program that was delivered online through videoconferencing. SPECIFiC was designed to be delivered in person; however, due to the COVID-19 pandemic, the program was delivered remotely. Getting on With It (GOWI) is a 6-session program delivered online via Zoom (Zoom Video Communications) that aimed to support caregivers of those with FASD [50]. GOWI was adapted from an in-person program [53] and delivered via Zoom. The original program consists of 1 session delivered over several weeks, each with a different topic. Some of the topics include an introduction to FASD and dealing with systems and getting on with professionals [53]. These programs then developed into Families Linking with Families, which is a 7-session program delivered online via Zoom [49]. Caregivers who completed GOWI and Families Moving Forward Connect were satisfied with the program and reported improvements in their knowledge of FASD [49,50].

Skill Development

In total, 7 studies explored the use of technology to promote skill development for people with FASD. Within this subtheme, technology focused on the development of motor skills, life skills, disruptive behavior, and executive functioning.

Motor Skills

In total, 3 studies [43,44,52] explored the efficacy of Sensorimotor Training to Affect Balance, Engagement and Learning (STABEL) technology, a VR program that required children and adolescents with FASD to complete virtual tasks while the surface they were standing on changed. STABEL was designed to improve balance and motor skills for those living with FASD. STABEL was found to be engaging and showed improvement in motor skills for those with FASD who were assigned to the STABEL group relative to a control group.

Life Skills

In total, 2 studies [36,45] explored technology that focused on developing life skills, specifically fire and road safety, for children younger than 10 years of age with FASD. Both technologies required children to complete tasks in either a VR or computer-based game. The tasks focused on developing the relevant fire and road safety steps that would be translated into real-world settings. Across both studies, people with FASD effectively learned road and fire safety skills. However, when the road and fire skills were tested in real-world settings 1 week after receiving the technology, participants did not maintain the skills they had developed.

Disruptive Behavior

In total, 2 studies [37,41] explored the GoFAR program, a 3-component program aimed to reduce disruptive behaviors. In this program, children younger than 10 years of age with FASD use a computer game, where they navigate characters through a range of challenges designed to improve self-regulation. Therapy sessions are offered for the primary parents or caregivers of the children to educate them to promote self-regulation. Then, behavior analogy therapy is offered for the child and caregiver to support the implementation of behavior in real-life settings. The program was well received by caregivers and people with FASD. For individuals with FASD, the program showed improvements in behavioral skills and a reduction in disruptive behavior.

Executive Functioning

One protocol planned to test the efficacy of a game developed by the Foundation of Alcohol Related Research (FARR) known as the FARR game [42], a computerized cognitive training game in which people with FASD must navigate through more demanding cognitive tasks, designed to improve attention, memory, inhibition, and working memory. At the time of this review, no publication with outcomes for this protocol was available.

Stakeholder Engagement

Of the studies that were included in the review, none involved caregivers or people living with FASD in the design or development process. Two studies sought caregiver input into the design and usability of an app and web-based program during the development of the technology [39,47]. However, the content and design of the app had been largely established prior to stakeholder engagement.

Discussion

Principal Results and Summary of Evidence

Although research is limited, there is growing evidence to show that technology can support skill development for people with FASD and support the well-being of caregivers of those with FASD. Of the few studies, most focused on developing skills for those with FASD and showed improvements in motor skills, life skills, disruptive behavior, and executive functioning. Technology has only been investigated in few of the neurocognitive domains PAE can affect, such as affect regulation and behavioral regulation [1]. A small number of studies have explored the use of technology to support caregivers. These studies showed that the technology (such as app, telehealth, and web-based programs) is well received and useful to caregivers. This review suggests that technology could be a viable option to support people with FASD and their caregivers, but currently technology is underused and remains largely unexplored.

Assessment and Diagnosis of FASD

Evidence suggests that professionals feel unprepared to support and diagnose people with FASD [10,54] and can hold stigmatizing beliefs [55]. This stigma and lack of knowledge of health professionals can result in hesitancy to diagnose people with FASD and long wait times for a diagnosis. Significant delay between seeking help and a confirmed diagnoses of FASD can lead to poor outcomes for both the families and individuals. Technology potentially offers a way of improving the diagnostic pathway for FASD. For example, there is growing evidence supporting the potential for using telehealth and computer-based assessment methods to improve access to assessment and diagnosis of ASD [15]. Yet, the use of telehealth for those with FASD has only been explored in 3 studies [38,56,57], of which only 1 looked at the experiences of caregivers [38], which showed that telehealth can bring FASD diagnosis and support to those living in remote locations. Similarly, technology has been used to aid health professionals in diagnosing people with FASD [17]. Future research could explore the use of technology to facilitate self-screening of suspected FASD, such as detecting common difficulties experienced by those with FASD. If technology could screen for FASD, this could encourage the seeking of formal diagnosis, therefore facilitating a formal diagnosis.

There is potential for emerging technologies, like AI and machine learning, to be of use in assessment and diagnosis, although we did not find any examples in our scoping review. For example, combining facial recognition software and AI could be used to detect facial dysmorphia in FAS. Additionally, recent studies have made use of machine learning models to analyze brain imaging data to improve diagnosis and treatment outcomes for ADHD [58]. Telehealth solutions incorporating virtual assessments and consultations could provide FASD assessment, diagnosis, and support to those living in remote locations or underserved areas.

Support and Interventions

People with FASD and their caregivers often experience difficulties accessing support due to a lack of support available, trained professionals [8,37], and funding to access services. Despite a rapid growth in the use of telehealth, computer-based therapies, and digital technologies to support physical and mental health [18,19], relatively few technologies have been designed specifically for FASD. As technology has showed promise in supporting those with intellectual disability [20,21,27], ASD [25], and neurodevelopmental disabilities [14], technology could provide effective support to those with FASD. For instance, telehealth has been used extensively in delivering health care remotely to those with intellectual disabilities [16]. Future research could explore the use of telehealth services to provide support for FASD to families living in remote areas or to enable the delivery of group-based parent training.

International reviews indicate that VR experiences can support people with anxiety [22] and depression [59] and support mental health for those with intellectual disabilities [20]. Specifically, VR programs have been used to improve attention and memory for people with ADHD [26] and emotional recognition for those with ASD [25]. Our review shows that VR can improve motor skills [43,44] and safety skills [45] for people with FASD, demonstrating the potential for VR technology to target specific skill development for those with FASD.

In addition, VR enables people to be immersed in a virtual world and situations, to experience situations to develop skills, and to experience the viewpoint of other people. VR can promote perspective taking and empathy and allow people to experience a condition, which in turn has the potential to reduce stigma toward those with mental illness [40], depression [41], and schizophrenia [13]. People with FASD can experience poorer mental health [5,38], in part driven by stigma [39], social isolation, and everyday difficulties people with FASD experience [7]. Professionals believe that FASD can be a highly stigmatized disability [54,55]. Therefore, VR experiences for professionals might promote understanding and reduce FASD stigma.

There is currently no published research on the use of real-time monitoring and wearable devices to support people with FASD. Wearable devices, such as those that monitor real-time body signals of heartbeat and respirations for stress level, have been used to capture stress response in daily life in children with ASD [60]. The findings can aid the development of real-time interventions and tailored treatment. A systematic review on wearables and mobile technologies interventions in ASD alone found 83 publications [61]. Compared with other neurodevelopmental disorders, we found the use of technological advances in FASD to be limited. Similar technology, wearable technology, VR experiences, and telehealth could support those living with FASD and their caregivers. Therefore, future research could adapt and co-design technology for those with FASD. The potential of technology for supporting individuals and caregivers living with FASD is largely unexplored. However, there may be challenges to adopting these technologies, such as cost of equipment, training requirements, and issues with data quality [62,63].

Support Across the Life Span

The challenges people with FASD experience persist throughout their lifetime [9]. Those who are not provided with adequate and early support are likely to experience challenges with employment, independence, living situations, and forensic engagement [3,8]. In this review, no technology or studies explored the use of technology to support people older than 18 years of age. FASD research rarely focuses on those older than 18 years of age living with FASD [11]. This gap is problematic, given that people with FASD require lifelong support, and with very little support and trained professionals available, technology could provide a cost-effective and potentially effective avenue to support people with FASD.

Given the potential of technology to support those with FASD and their caregivers, there is a need to advocate for policy changes to investigate, develop, and sustain technology for those with FASD. Digital platforms could be used to facilitate collaboration between policy makers, researchers, health care providers, and community stakeholders to develop evidence-based policies and allocate resources. By focusing on policies and solutions, researchers, practitioners, and policy makers can harness the potential of digital solutions to improve the prevention, diagnosis, intervention, and support for individuals affected by FASD, ultimately enhancing their quality of life and well-being.

End-User Involvement

Technology designed for caregivers and people with FASD could be best created through consultation and feedback from those whom the technology is intended for; however, only 2 studies in our review sought feedback from end users in the development and only after the material had largely been finalized. Evidence consistently suggests the importance of stakeholder engagement during the development process [64,65]. People with FASD can face unique challenges with executive functioning, memory, and attention [1], which could affect their ability to engage with and gain benefits through technology. The diversity of individuals with FASD must also be considered when developing technological tools. If people with FASD and their caregivers are not considered in the development of technologies, it could drive health inequities, as the support and interventions are less likely to be accessible to people with FASD. Therefore, it is important to involve and respond to stakeholder voices at all stages of developments to increase the uptake and usability of technology. Future research needs to explore the accessibility of technology, feasibility of types of technologies, and the type of technology solutions that would support those with FASD. Identifying the views of those with FASD on technology can provide insight into the main ethical issues of concern and could lead to effective solutions and supports that could be scaled up.

Limitations

Our review may be limited through the databases and search terms that were used. We used a broad definition of technology to capture the wide-ranging and rapidly advancing field. However, our definition of technology may have limited our searches to identify different types of technology. A scoping review enables us to include unpublished literature and publicly available technology such as through the app store. As with any scoping review, we are unable to examine the effectiveness of technology for those with FASD but rather just explore the scope of the existing technology.

It should also be noted that technology might be used in clinical practices, but in this review, we have found limited published evidence about efficacy, costs, or user impact. Technology is an area of rapid development, and there are likely barriers between the use of technology in research and its use clinically. Future research, such as pragmatic trials and service evaluation to evaluate the real-world impact of technology in clinical practices, is needed.

Conclusions

The potential for technology to provide vital support to people living with FASD and their caregivers is largely unexplored. Present examples are limited to specific skills and caregiver support. Learnings can be drawn from the ASD field where transformative technological advances have been examined and applied to those with ASD in clinical and real-world practices. There are areas for development of tools and technology to promote skill development, connection to others affected by

FASD, and well-being and provide cost-effective diagnosis, assessment, and support for the those affected by FASD. The value of technologies is yet to be applied in any area of FASD, despite a growing demand for more effective and efficient services. There is an urgent need to explore the opportunities offered by technology to support those living with FASD.

Acknowledgments

This study was funded by the Health Research Council (21/751) and Ember Korowai Takitini.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Cochrane Library search strategy.
[DOCX File, 14 KB - mental_v11i1e51074_app1.docx]

Multimedia Appendix 2
Quality assessment of each study.
[DOCX File, 16 KB - mental_v11i1e51074_app2.docx]

Multimedia Appendix 3
Characteristics of included studies.
[DOCX File, 25 KB - mental_v11i1e51074_app3.docx]

Checklist 1
Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) checklist.
[PDF File, 520 KB - mental_v11i1e51074_app4.pdf]

References

1. Cook JL, Green CR, Lilley CM, et al. Fetal alcohol spectrum disorder: a guideline for diagnosis across the lifespan. CMAJ 2016 Feb 16;188(3):191-197. [doi: 10.1503/cmaj.141593] [Medline: 26668194]
2. Lange S, Probst C, Gmel G, Rehm J, Burd L, Popova S. Global prevalence of fetal alcohol spectrum disorder among children and youth: a systematic review and meta-analysis. JAMA Pediatr 2017 Oct 1;171(10):948-956. [doi: 10.1001/jamapediatrics.2017.1919] [Medline: 28828483]
3. Streissguth AP, Bookstein FL, Barr HM, Sampson PD, O'Malley K, Young JK. Risk factors for adverse life outcomes in fetal alcohol syndrome and fetal alcohol effects. J Dev Behav Pediatr 2004 Aug;25(4):228-238. [doi: 10.1097/00004703-200408000-00002] [Medline: 15308923]
4. Rangmar J, Hjern A, Vinnerljung B, Strömland K, Aronson M, Fahlke C. Psychosocial outcomes of fetal alcohol syndrome in adulthood. Pediatrics 2015 Jan;135(1):e52-e58. [doi: 10.1542/peds.2014-1915] [Medline: 25535260]
5. Flannigan K, McMorris C, Ewasiuk A, et al. Suicidality and associated factors among individuals assessed for fetal alcohol spectrum disorder across the lifespan in Canada. Can J Psychiatry 2022 May;67(5):361-370. [doi: 10.1177/07067437211053288] [Medline: 34738837]
6. Thanh NX, Jonsson E. Life expectancy of people with fetal alcohol syndrome. J Popul Ther Clin Pharmacol 2016;23(1):e53-e59. [Medline: 26962962]
7. Popova S, Lange S, Burd L, Chudley AE, Clarren SK, Rehm J. Cost of fetal alcohol spectrum disorder diagnosis in Canada. PLoS One 2013;8(4):e60434. [doi: 10.1371/journal.pone.0060434] [Medline: 23593216]
8. McLachlan K, Flannigan K, Temple V, Unsworth K, Cook JL. Difficulties in daily living experienced by adolescents, transition-aged youth, and adults with fetal alcohol spectrum disorder. Alcohol Clin Exp Res 2020 Aug;44(8):1609-1624. [doi: 10.1111/acer.14385] [Medline: 32472600]
9. Domeij H, Fahlström G, Bertilsson G, et al. Experiences of living with fetal alcohol spectrum disorders: a systematic review and synthesis of qualitative data. Dev Med Child Neurol 2018 Aug;60(8):741-752. [doi: 10.1111/dmcn.13696] [Medline: 29479676]
10. McCormack JC, Chu JTW, Marsh S, Bullen C. Knowledge, attitudes, and practices of fetal alcohol spectrum disorder in health, justice, and education professionals: a systematic review. Res Dev Disabil 2022 Dec;131:104354. [doi: 10.1016/j.ridd.2022.104354] [Medline: 36375286]
11. Reid N, Dawe S, Shelton D, et al. Systematic review of fetal alcohol spectrum disorder interventions across the life span. Alcohol Clin Exp Res 2015 Dec;39(12):2283-2295. [doi: 10.1111/acer.12903] [Medline: 26578111]

12. National Institute for Health and Care Research. 2024. URL: https://www.nihr.ac.uk/ [accessed 2024-06-08]
13. Abernethy A, Adams L, Barrett M, et al. The promise of digital health: then, now, and the future. NAM Perspect 2022 Jun 27;2022. [doi: 10.31478/202206e] [Medline: 36177208]
14. Ribas MO, Micai M, Caruso A, Fulceri F, Fazio M, Scattoni ML. Technologies to support the diagnosis and/or treatment of neurodevelopmental disorders: a systematic review. Neurosci Biobehav Rev 2023 Feb;145:105021. [doi: 10.1016/j.neubiorev.2022.105021]
15. Alfuraydan M, Croxall J, Hurt L, Kerr M, Brophy S. Use of telehealth for facilitating the diagnostic assessment of autism spectrum disorder (ASD): a scoping review. PLoS One 2020;15(7):e0236415. [doi: 10.1371/journal.pone.0236415] [Medline: 32702017]
16. Selick A, Bobbette N, Lunsky Y, Hamdani Y, Rayner J, Durbin J. Virtual health care for adult patients with intellectual and developmental disabilities: a scoping review. Disabil Health J 2021 Oct;14(4):101132. [doi: 10.1016/j.dhjo.2021.101132] [Medline: 34127406]
17. Lim YH, Watkins RE, Jones H, Kippin NR, Finlay-Jones A. Fetal alcohol spectrum disorders screening tools: a systematic review. Res Dev Disabil 2022 Mar;122:104168. [doi: 10.1016/j.ridd.2021.104168] [Medline: 34996007]
18. Garrido S, Millington C, Cheers D, et al. What works and what doesn't work? A systematic review of digital mental health interventions for depression and anxiety in young people. Front Psychiatry 2019 Nov;10:759. [doi: 10.3389/fpsyt.2019.00759] [Medline: 31798468]
19. De Witte NAJ, Joris S, Van Assche E, Van Daele T. Technological and digital interventions for mental health and wellbeing: an overview of systematic reviews. Front Digit Health 2021 Dec;3:754337. [doi: 10.3389/fdgth.2021.754337] [Medline: 35005695]
20. Oudshoorn CEM, Frielink N, Nijs SLP, Embregts P. Psychological eHealth interventions for people with intellectual disabilities: a scoping review. J Appl Res Intellect Disabil 2021 Jul;34(4):950-972. [doi: 10.1111/jar.12877] [Medline: 33704872]
21. Oudshoorn CEM, Frielink N, Nijs SLP, Embregts P. eHealth in the support of people with mild intellectual disability in daily life: a systematic review. J Appl Res Intellect Disabil 2020 Nov;33(6):1166-1187. [doi: 10.1111/jar.12758] [Medline: 32578361]
22. van Loenen I, Scholten W, Muntingh A, Smit J, Batelaan N. The effectiveness of virtual reality exposure-based cognitive behavioral therapy for severe anxiety disorders, obsessive-compulsive disorder, and posttraumatic stress disorder: meta-analysis. J Med Internet Res 2022 Feb 10;24(2):e26736. [doi: 10.2196/26736] [Medline: 35142632]
23. Kalyanaraman SS, Penn DL, Ivory JD, Judge A. The virtual doppelganger: effects of a virtual reality simulator on perceptions of schizophrenia. J Nerv Ment Dis 2010 Jun;198(6):437-443. [doi: 10.1097/NMD.0b013e3181e07d66] [Medline: 20531123]
24. Dreesmann NJ, Su H, Thompson HJ. A systematic review of virtual reality therapeutics for acute pain management. Pain Manag Nurs 2022 Oct;23(5):672-681. [doi: 10.1016/j.pmn.2022.05.004] [Medline: 35868974]
25. Farashi S, Bashirian S, Jenabi E, Razjouyan K. Effectiveness of virtual reality and computerized training programs for enhancing emotion recognition in people with autism spectrum disorder: a systematic review and meta-analysis. Int J Dev Disabil 2024 Apr;70(1):110-126. [doi: 10.1080/20473869.2022.2063656] [Medline: 38456137]
26. Corrigan N, Păsărelu CR, Voinescu A. Immersive virtual reality for improving cognitive deficits in children with ADHD: a systematic review and meta-analysis. Virtual Real 2023 Feb 18:1-20. [doi: 10.1007/s10055-023-00768-1] [Medline: 36845650]
27. MacHale R, Ffrench C, McGuire B. The experiences and views of adults with intellectual disabilities accessing digital mental health interventions: a qualitative systematic review and thematic synthesis. J Appl Res Intellect Disabil 2023 May;36(3):448-457. [doi: 10.1111/jar.13082] [Medline: 36781378]
28. Valentine AZ, Brown BJ, Groom MJ, Young E, Hollis C, Hall CL. A systematic review evaluating the implementation of technologies to assess, monitor and treat neurodevelopmental disorders: a map of the current evidence. Clin Psychol Rev 2020 Aug;80:101870. [doi: 10.1016/j.cpr.2020.101870] [Medline: 32712216]
29. Oh SS, Moon JY, Chon D, et al. Effectiveness of digital interventions for preventing alcohol consumption in pregnancy: systematic review and meta-analysis. J Med Internet Res 2022 Apr 11;24(4):e35554. [doi: 10.2196/35554] [Medline: 35404257]
30. Whittingham LM, Coons-Harding KD. Connecting people with people: diagnosing persons with fetal alcohol spectrum disorder using telehealth. J Autism Dev Disord 2021 Apr;51(4):1067-1080. [doi: 10.1007/s10803-020-04607-z] [Medline: 32648146]
31. Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med Res Methodol 2018 Nov 19;18(1):143. [doi: 10.1186/s12874-018-0611-x] [Medline: 30453902]
32. Tricco AC, Lillie E, Zarin W, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 2;169(7):467-473. [doi: 10.7326/M18-0850] [Medline: 30178033]
33. Hong QN, Pluye P, Fabregues S, et al. Mixed Methods Apprisal Tool (MMAT) version 2018. : Canadian Intellectual Property Office, Industry Canada; 2018 URL: http://mixedmethodsappraisaltoolpublic.pbworks.com/w/file/fetch/127425851/MMAT_2018_criteria-manual_2018-04-04.pdf [accessed 2024-06-22]

34. Hong QN, Fàbregues S, Bartlett G, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Education for Information 2018 Dec;34(4):285-291. [doi: 10.3233/EFI-180221]
35. Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis: JBI; 2020. URL: https://synthesismanual.jbi.global [accessed 2024-07-04]
36. Coles CD, Strickland DC, Padgett L, Bellmoff L. Games that "work": using computer games to teach alcohol-affected children about fire and street safety. Res Dev Disabil 2007;28(5):518-530. [doi: 10.1016/j.ridd.2006.07.001] [Medline: 16965894]
37. Coles CD, Kable JA, Taddeo E, Strickland DC. A metacognitive strategy for reducing disruptive behavior in children with fetal alcohol spectrum disorders: GoFAR pilot. Alcohol Clin Exp Res 2015 Nov;39(11):2224-2233. [doi: 10.1111/acer.12885] [Medline: 26503069]
38. Hanlon-Dearman A, Edwards C, Schwab D, Millar MC, Longstaffe S. “Giving voice”: evaluation of an integrated telehealth community care model by parents/guardians of children diagnosed with fetal alcohol spectrum disorder in Manitoba. Telemed J E Health 2014 May;20(5):478-484. [doi: 10.1089/tmj.2013.0161] [Medline: 24568178]
39. Hundert AS, Huguet A, Green CR, et al. Usability testing of guided internet-based parent training for challenging behavior in children with fetal alcohol spectrum disorder (Strongest Families FASD). J Popul Ther Clin Pharmacol 2016;23(1):e60-e76. [Medline: 27115205]
40. Kable JA, Coles CD, Strickland D, Taddeo E. Comparing the effectiveness of on-line versus in-person caregiver education and training for behavioral regulation in families of children with FASD. Int J Ment Health Addict 2012 Dec;10(6):791-803. [doi: 10.1007/s11469-012-9376-3] [Medline: 31024223]
41. Kable JA, Taddeo E, Strickland D, Coles CD. Improving FASD children's self-regulation: piloting phase 1 of the GoFAR intervention. Child Fam Behav Ther 2016 May;38(2):124-141. [doi: 10.1080/07317107.2016.1172880] [Medline: 29104359]
42. Louw JG, Olivier L, Skeen S, van Heerden A, Tomlinson M. Evaluation of a custom-developed computer game to improve executive functioning in 4- to 6-year-old children exposed to alcohol in utero: protocol for a feasibility randomized controlled trial. JMIR Res Protoc 2019 Oct 8;8(10):e14489. [doi: 10.2196/14489] [Medline: 31596247]
43. McCoy SW, Hsu LY, Jirikowic T, Price R, Ciol M, Kartin D. Sensorimotor training to affect balance, engagement, and learning for children with fetal alcohol spectrum disorders. Physiotherapy 2015 May;101:e971-e972. [doi: 10.1016/j.physio.2015.03.1828]
44. McCoy SW, Jirikowic T, Price R, et al. Virtual sensorimotor balance training for children with fetal alcohol spectrum disorders: feasibility study. Phys Ther 2015 Nov;95(11):1569-1581. [doi: 10.2522/ptj.20150124] [Medline: 26112255]
45. Padgett LS, Strickland D, Coles CD. Case study: using a virtual reality computer game to teach fire safety skills to children diagnosed with fetal alcohol syndrome. J Pediatr Psychol 2006 Jan;31(1):65-70. [doi: 10.1093/jpepsy/jsj030] [Medline: 15829610]
46. Petrenko CLM, Kautz-Turnbull CC, Roth AR, et al. Initial feasibility of the “Families Moving Forward Connect” mobile health intervention for caregivers of children with fetal alcohol spectrum disorders: mixed method evaluation within a systematic user-centered design approach. JMIR Form Res 2021 Dec 2;5(12):e29687. [doi: 10.2196/29687] [Medline: 34860661]
47. Petrenko CL, Parr J, Kautz C, Tapparello C, Olson HC. A mobile health intervention for fetal alcohol spectrum disorders (Families Moving Forward Connect): development and qualitative evaluation of design and functionalities. JMIR Mhealth Uhealth 2020 Apr 6;8(4):e14721. [doi: 10.2196/14721] [Medline: 32250274]
48. Turner K, Reynolds JN, McGrath P, et al. Guided internet-based parent training for challenging behavior in children with fetal alcohol spectrum disorder (Strongest Families FASD): study protocol for a randomized controlled trial. JMIR Res Protoc 2015 Oct 13;4(4):e112. [doi: 10.2196/resprot.4723] [Medline: 26462968]
49. Gibbs A, Flanagan J, Gray L. An Australian online training and support program for caregivers of children and youth with fetal alcohol spectrum disorder: Families Linking with Families. J Intellect Dev Disabil 2024 Apr 2;49(2):175-185. [doi: 10.3109/13668250.2023.2271757]
50. Gibbs A, Harrington S, Robinson C, Brooks C, Dedman C. “Getting on With It”: a course by caregivers, for caregivers—pilot overview and evaluation report. : NOFASD Australia; 2020 URL: https://www.nofasd.org.au/parents-carers-and-families/support-for-caregivers-families/ [accessed 2024-06-22]
51. Price AD, Mukherjee RAS, Webster A, et al. Development and pre-feasibility testing of specific: a psychoeducation programme for caregivers of children with fetal alcohol spectrum disorder (FASD). J Child Fam Stud 2023 Oct;32(10):3026-3041. [doi: 10.1007/s10826-023-02637-6]
52. Jirikowic T, Westcott McCoy S, Price R, Ciol MA, Hsu LY, Kartin D. Virtual sensorimotor training for balance: pilot study results for children with fetal alcohol spectrum disorders. Pediatr Phys Ther 2016;28(4):460-468. [doi: 10.1097/PEP.0000000000000300] [Medline: 27661243]
53. Gibbs A. An evidence-based training and support course for caregivers of children with foetal alcohol spectrum disorder (FASD) in New Zealand. Adv Dual Diagn 2019 Feb 18;12(1/2):73-84. [doi: 10.1108/ADD-10-2018-0013]
54. Chu JTW, McCormack JC, Marsh S, Bullen C. Knowledge, attitudes, and practices towards fetal alcohol spectrum disorder in New Zealand educators: an online survey. J Intellect Disabil 2023 Sep;27(3):762-776. [doi: 10.1177/17446295221104618] [Medline: 35634949]

55. McCormack JC, Chu JTW, Wilson H, Rahman J, Marsh S, Bullen C. Knowledge, attitudes, and practices towards fetal alcohol spectrum disorder in the New Zealand social and community sector: an online survey. J Intellect Disabil 2023 Apr 24:17446295231172234. [doi: 10.1177/17446295231172234] [Medline: 37092706]
56. Ens CDL, Hanlon-Dearman A, Millar MC, Longstaffe S. Using telehealth for assessment of fetal alcohol spectrum disorder: the experience of two Canadian rural and remote communities. Telemed J E Health 2010 Oct;16(8):872-877. [doi: 10.1089/tmj.2010.0070] [Medline: 20925569]
57. Benoit T, Bowes C, Bowman N, et al. Telemedicine diagnosis for fetal alcohol syndrome—the Manitoba experience. Paediatr Child Health 2002 Mar;7(3):147-151. [doi: 10.1093/pch/7.3.147] [Medline: 20046287]
58. Sethu N, Vyas R. Overview of Machine Learning Methods in ADHD Prediction: Springer Singapore; 2020:51-71.
59. Fodor LA, Cote CD, Cuijpers P, Szamoskozi , David D, Cristea IA. The effectiveness of virtual reality based interventions for symptoms of anxiety and depression: a meta-analysis. Sci Rep 2018 Jul 9;8(1):10323. [doi: 10.1038/s41598-018-28113-6] [Medline: 29985400]
60. Fioriello F, Maugeri A, D'Alvia L, et al. A wearable heart rate measurement device for children with autism spectrum disorder. Sci Rep 2020 Oct 29;10(1):18659. [doi: 10.1038/s41598-020-75768-1] [Medline: 33122808]
61. Koumpouros Y, Kafazis T. Wearables and mobile technologies in autism spectrum disorder interventions: a systematic literature review. Res Autism Spectr Disord 2019 Oct;66:101405. [doi: 10.1016/j.rasd.2019.05.005]
62. Zakerabasali S, Ayyoubzadeh SM, Baniasadi T, Yazdani A, Abhari S. Mobile health technology and healthcare providers: systemic barriers to adoption. Healthc Inform Res 2021 Oct;27(4):267-278. [doi: 10.4258/hir.2021.27.4.267] [Medline: 34788907]
63. Cho S, Ensari I, Weng C, Kahn MG, Natarajan K. Factors affecting the quality of person-generated wearable device data and associated challenges: rapid systematic review. JMIR Mhealth Uhealth 2021 Mar 19;9(3):e20738. [doi: 10.2196/20738] [Medline: 33739294]
64. Cole AC, Adapa K, Khasawneh A, Richardson DR, Mazur L. Codesign approaches involving older adults in the development of electronic healthcare tools: a systematic review. BMJ Open 2022 Jul 6;12(7):e058390. [doi: 10.1136/bmjopen-2021-058390] [Medline: 35793923]
65. Fischer B, Peine A, Östlund B. The importance of user involvement: a systematic review of involving older users in technology design. Gerontologist 2020 Sep 15;60(7):e513-e523. [doi: 10.1093/geront/gnz163] [Medline: 31773145]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
AI: artificial intelligence
ASD: autism spectrum disorder
FARR: Foundation of Alcohol Related Research
FAS: fetal alcohol syndrome
FASD: fetal alcohol spectrum disorder
GOWI: Getting on With It
MMAT : Mixed Methods Appraisal Tool
NIHR: National Institute for Health and Care Research
PAE : prenatal alcohol exposure
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews
SPECIFiC: Salford Parents and Carers Education Course for Improvements in FASD Outcomes in Children
STABEL: Sensorimotor Training to Affect Balance, Engagement and Learning
VR: virtual reality

Edited by J Torous; submitted 20.07.23; peer-reviewed by J Chan, L D'Adamo, SQ Yoong; revised version received 14.04.24; accepted 15.04.24; published 11.07.24.

Please cite as:
Chu JTW, Wilson H, Cai CZ, McCormack JC, Newcombe D, Bullen C
Technologies for Supporting Individuals and Caregivers Living With Fetal Alcohol Spectrum Disorder: Scoping Review
JMIR Ment Health 2024;11:e51074
URL: https://mental.jmir.org/2024/1/e51074
doi:10.2196/51074

Review

Health Care Professionals' Views on the Use of Passive Sensing, AI, and Machine Learning in Mental Health Care: Systematic Review With Meta-Synthesis

Jessica Rogan[1,2], BSc, MSc, ClinPsy; Sandra Bucci[1,2], BSc, ClinPsy; Joseph Firth[1], BSc, PhD

[1]Division of Psychology and Mental Health, School of Health Sciences, Faculty of Biology, Medicine and Health, Manchester Academic Health Sciences, The University of Manchester, Manchester, United Kingdom

[2]Greater Manchester Mental Health NHS Foundation Trust, Manchester, United Kingdom

Corresponding Author:
Sandra Bucci, BSc, ClinPsy
Division of Psychology and Mental Health, School of Health Sciences, Faculty of Biology, Medicine and Health, Manchester Academic Health Sciences
The University of Manchester
2nd Floor, Zochonis Building, Brunswick Street
Manchester, M13 9PL
United Kingdom
Phone: 44 161 306 0422
Email: sandra.bucci@manchester.ac.uk

Abstract

Background: Mental health difficulties are highly prevalent worldwide. Passive sensing technologies and applied artificial intelligence (AI) methods can provide an innovative means of supporting the management of mental health problems and enhancing the quality of care. However, the views of stakeholders are important in understanding the potential barriers to and facilitators of their implementation.

Objective: This study aims to review, critically appraise, and synthesize qualitative findings relating to the views of mental health care professionals on the use of passive sensing and AI in mental health care.

Methods: A systematic search of qualitative studies was performed using 4 databases. A meta-synthesis approach was used, whereby studies were analyzed using an inductive thematic analysis approach within a critical realist epistemological framework.

Results: Overall, 10 studies met the eligibility criteria. The 3 main themes were uses of passive sensing and AI in clinical practice, barriers to and facilitators of use in practice, and consequences for service users. A total of 5 subthemes were identified: barriers, facilitators, empowerment, risk to well-being, and data privacy and protection issues.

Conclusions: Although clinicians are open-minded about the use of passive sensing and AI in mental health care, important factors to consider are service user well-being, clinician workloads, and therapeutic relationships. Service users and clinicians must be involved in the development of digital technologies and systems to ensure ease of use. The development of, and training in, clear policies and guidelines on the use of passive sensing and AI in mental health care, including risk management and data security procedures, will also be key to facilitating clinician engagement. The means for clinicians and service users to provide feedback on how the use of passive sensing and AI in practice is being received should also be considered.

Trial Registration: PROSPERO International Prospective Register of Systematic Reviews CRD42022331698; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=331698

(JMIR Ment Health 2024;11:e49577) doi:10.2196/49577

KEYWORDS

artificial intelligence; machine learning; passive sensing; mental health care; clinicians; views; meta-synthesis; review; mental health; health care; health care professionals; psychology; psychiatry; mental health professionals; mobile phone

Introduction

Background

Mental health problems are highly prevalent globally, with approximately 1 in 8 people experiencing mental health difficulties, which can have significant personal and economic consequences [1]. Rapid growth in digital technology innovation has led to an increased interest in digital mental health interventions [2]. Digital tools with built-in sensors, such as smartphones, smartwatches, and other wearable devices, allow for the unobtrusive and continuous collection of objective data, providing insight into user behavior and physiology [3]. Machine learning, which is a branch of artificial intelligence (AI), can be applied to these data to *learn* from it and generate clinically actionable insights and predictions [4]. It has therefore been suggested that passive sensing data and applied machine learning methods could overcome what some describe as trial-and-error–driven approaches used in mental health care by supporting precise diagnoses and prognoses [5]. Indeed, mental health remains one of the only domains in health care that relies only on service users' self-report of cognitive and emotional states and symptoms and on clinicians to accurately recognize and map these states to make diagnostic, prognostic, and therapeutic decisions [6]. Passive sensing data and AI may offer a means to overcome the pitfalls of current clinical measures by presenting a more complete picture of a person's difficulties [7]. For example, raw sensor data captured regarding speech characteristics, location, and activity can be transformed to derive high-level behavioral markers, such as fatigue, sleep disruption, and mood, which can be used to identify clinical states, such as depression [8]. In addition, digital tools that allow for passive sensing can support service users' self-management of symptoms and access to digitally delivered therapies [4]. Through self-management, service users may feel empowered [9], and service user and clinician access to digital remote data capture has the potential to identify early warning signs of deterioration, providing the opportunity to reduce the risk of relapse of mental health difficulties via early identification and intervention [10]. This may be particularly useful, as current health care systems generally rely on the delivery of treatment by scheduled appointments, which can result in warning signs of mental health relapse being missed or treated too late [11]. Using sensors from digital tools, such as smartphones and wearable devices, to identify clinical and behavioral features of worsening mental health and applying machine learning methods to identify patterns in the data could augment mental health care by delivering more precise treatment at the time it is needed [12].

Despite the potential benefits, there remains a persistent gap between the rapid developments in digitally supported mental health care and the successful adoption of these tools in clinical practice [13]. A key driver to the potential success of digitally supported health care uptake is the willingness, confidence, and capacity of clinicians to make changes to their practice [9]. Resistance to incorporating digital approaches in clinical practice can occur for various reasons, including the lack of technological literacy, fear that AI models could replace professionals, and concerns about ethical and legal issues [6]. There is trepidation that core aspects of clinician roles, such as diagnosis, assessment, formulation and treatment, may be delegated to AI models without human input [14]. This has been viewed as dehumanizing and could have negative implications for therapeutic relationship [15]. Ethical issues have also been raised, such as implications for service user privacy and data security [16]. As clinicians' perceptions and attitudes pose a potential barrier to implementation [17], it is important that they are invited into the dialog around digitally supported AI in mental health care, to embrace any benefits there might be, as well as share their concerns and explore the limitations and risks [2]. However, it has been noted that stakeholder's views are rarely considered in model design or evaluation in relation to machine learning approaches [18]. Indeed, professionals have felt that their knowledge and views have been disregarded in the design of digital health solutions or are only considered as an afterthought [19]. As the extent to which these methods can be successfully implemented in health care depends on their acceptability [3], research is needed to understand stakeholders' perspectives on digital health systems [11].

Objectives

Although there have been some qualitative studies exploring mental health care professionals' views and experiences of passive sensing and AI in mental health care, there are no published reviews that systematically aggregate these findings, specifically through examining participants' experiences and perspectives, both deeply (because of the qualitative approach) and broadly (because of the integration of studies from different health care contexts and participants) [20]. This meta-synthesis aims to synthesize and evaluate the relative strengths of the qualitative literature regarding mental health care professionals' views on the use of passive sensing and AI in mental health care to provide a new, comprehensive interpretation of the findings that goes beyond the depth and breadth of the original studies [21]. Although research continues to grow in this area, it is now an appropriate time to review the literature, as the COVID-19 pandemic has increased the urgency for creating digital interventions that can fulfill the full potential of digital health [22], and it is necessary to engage multiple stakeholder groups early in the design and development process [23].

Methods

Overview

Meta-synthesis is a systematic review and integration of findings from qualitative studies to facilitate the transfer of knowledge and bring together a broad range of participants and descriptions [20]. A systematic approach for identifying and assessing the quality of potential papers, followed by analysis of the data and synthesis, was used with the aim of understanding what mental health care professionals think about the use of passive sensing and AI in mental health care. The review protocol was developed and registered with the International Prospective Register of Systematic Reviews (PROSPERO CRD42022331698).

Eligibility Criteria

Eligible papers for this review (1) were peer-reviewed studies published in English that used a qualitative method—mixed methods studies were also included, but only the qualitative findings were considered; and (2) examined health care professionals' views on hypothetical or actual use of service user-facing digital tools that use passive sensing and AI in mental health care. Studies with participants that included other stakeholders, as well as health care professionals, were not discounted; however, findings were only included if they were explicitly associated with mental health professionals. There were no limits on the publication year.

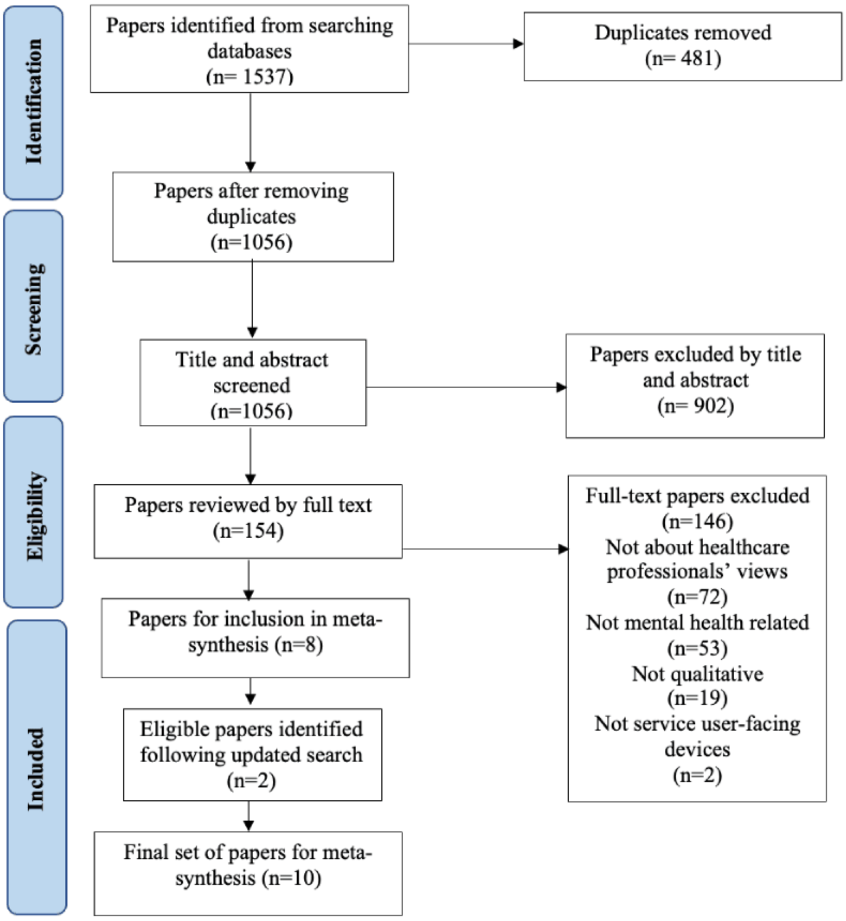
Search Strategy

A discussion within the research team and a review of the literature allowed for the identification of common terminology used in this research area and the selection of search terms. The search tool "SPIDER" (sample, phenomenon of interest, design, evaluation, research type) was used to ensure that all relevant areas were covered when developing the search terms. Relevant studies were identified through systematic searches of the following electronic databases: AMED, PsycINFO, Embase, and Medline. The search terms were (clinician*) OR (health care professional*) OR (staff) OR (physician*) OR (provider*) OR (practitioner*) OR (psychologist*) OR (doctor*) OR (therapist*) OR (care coordinator*) OR (mental health nurse*) OR (psychiatric nurse*) OR (support worker*) OR (counsellor *) OR (case manager*) OR (GP*) AND (view*) OR (opinion*) OR (perception*) OR (qualitative) OR (interview*) AND (remote monitoring) OR (digital phenotyping) OR (machine learning) OR (passive sens*) OR (passive monitor*) OR (passive data) OR (artificial intelligence) OR (wearables). A manual search of references and citations from eligible articles was also performed by JR to identify additional studies. Papers were initially screened according to title and abstract, followed by a full article.

Study Selection

The study selection and exclusion processes were conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [24] in October 2022 and are outlined in Figure 1. Article titles and abstracts were screened for eligibility by JR. If the inclusion criteria were unclear, full-text articles were obtained and reviewed. Any uncertainty regarding study eligibility was resolved through discussion with a wider research team. A second independent rater screened 10% (106/1056) of titles and reviewed 10% (16/154) of full-text articles to assess the reliability of the study selection. There was an "almost perfect" level of agreement between the raters at the screening stage (k=0.918) and at the full-text stage (k=1) [25]. As all studies were published in recent years, the search was conducted again in February 2023. Overall, 10 studies met the eligibility criteria and were included in the review.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of systematic search.



Quality Appraisal

Study quality was assessed using the Critical Appraisal Skills Programme (CASP) tool for quality appraisal in qualitative evidence synthesis (CASP, 2018, [26]), which assesses the strengths, limitations, relevance, and credibility of qualitative research. The CASP comprises 10 items that focus on different methodological aspects of qualitative studies, such as method, design, recruitment, data collection, and reflexivity. It is considered a good measure of transparency of research practice and reporting standards and is recommended for use in health-related research [27]; therefore, it was deemed appropriate for use in this review. A 3-point scale was used, with a score applied to each criterion (0=criterion not met, 1=criterion partially met, and 2=criterion totally met) [28]. Therefore, papers were given a total quality score of 20.

The scoring was completed by JR. A second independent rater assessed the quality of 50% (5/10) of the studies, and the scores were compared at the item level. Interrater reliability estimates showed good agreement between raters (k=0.832) [25]. Disagreements in ratings were resolved through discussion among raters until agreement was reached.

Data Extraction and Synthesis

The included studies were read and reread to ensure that they met the inclusion criteria. Key study information was then recorded, including the number and characteristics of participants, aims of the research, analysis method used, and the settings (Table 1). In addition, the original authors' analysis of primary qualitative data was extracted (second-order constructs), and individual participants' quotes were also noted (first-order constructs), in line with meta-synthesis principles [29]. An inductive thematic analysis approach within a critical realist epistemological framework was then taken with the aim of developing a cohesive, synthesized understanding of the data [30] and new interpretations [31]. JR completed the coding of the text and quotes using NVivo qualitative data analysis software (NVivo version 12, Lumivero). The constructs were then grouped into core themes. These themes were discussed by a broader research team, considering how each paper contributed to each core theme. The themes were then grouped into final higher-order themes, which were again reviewed and agreed upon by the research team. These themes are considered third-order constructs and allow for reflection on how each paper's findings fit within the wider literature and for findings to extend beyond the original papers [21]. JR returned to the papers to ensure that the themes identified reflected the data and that other themes were not overlooked.

Table 1. Summary of included studies.

Study	Participants	Aim	Data analysis	Setting	Results and themes	Critical Appraisal Skills Programme quality appraisal score (out of 20)
Greer et al [32] (the United Kingdom)	Five focus groups, made up of mental health nurses (N=25); age range 22-64 y; mean age 42.7 y; male: n=9; female: n=16	To explore staff views, specifically benefits and barriers to using remote monitoring to predict risk of inpatient aggression	Thematic analysis	Inpatient forensic mental health service, the United Kingdom	Utility in clinical practice, risk to user safety and well-being, data security and privacy, impact on staff workload, engagement, and adherence	17
Ng et al [33] (the United States)	Interviews with mental health professionals (N=17); age and gender not stated	Explore opportunities and barriers mental health staff perceive in applying sensor-captured patient-generated data among populations with PTSD[a] in routine care settings and how providers perspectives inform the design of tracking systems and strategies to implement technologies	Thematic analysis	Centre for Veterans With PTSD, the United States	Patient-driven uses of Fitbit and its data; integrating Fitbit data into treatment protocols; challenges to the use in treatment	18
Blease et al [34] (global)	Web-based survey of psychiatrists (N=791); age range 25-≥65 y; mean age group 35-44 y; male: n=550; female: n=230; other: n=11	To explore psychiatrists' opinions about the potential impact of innovations in AI[b] and machine learning on psychiatric practice	Qualitative descriptive analysis of written responses	Web-based survey across 22 countries	Patient-psychiatrist interactions, quality of patient medical care, profession of psychiatry, health systems	18
Thenral and Annamalai [35] (India)	Interviews with psychiatrists (n=14), patients (n=14), technology experts (n=13), and chief executive officers (n=5); overall (N=46); psychiatrist characteristics: mean age 35.5 y; male: n=7; female: n=7	To understand the perceived challenges in building, deploying, and using AI-enabled telepsychiatry for clinical practice	Grounded theory	Practices in urban areas of India: Chennai, Mumbai, Bangalore, and Delhi	Knowledge and gaps deficit; attitudes and perception; data challenges; ethical, legal accountability; AI related; health system infrastructure; human resources and skills; technology; clinical practice	15
Dawoodbhoy et al [36] (the United Kingdom)	Interviews with health care professionals (n=9) and AI experts (n=11); overall (N=20); age and gender of health care professionals not stated	To identify issues in patient flow on mental health units and align them with potential AI solutions, ultimately devising a model for their integration at service level	Thematic analysis	Acute mental health inpatient units, the United Kingdom	Current mental health inpatient service and patient flow model: patient factors; problems with social care; problems with clinical management; problems with inpatient service and system; solutions	16

Study	Participants	Aim	Data analysis	Setting	Results and themes	Critical Appraisal Skills Programme quality appraisal score (out of 20)
Rodriguez-Villa et al [23] (the United States and India)	Focus groups and interviews with mental health clinicians (n=53) and service users and their families (n=75); overall (N=128); clinician characteristics: age range 23-72 y; mean age 36 y; gender not stated	To engage clinicians and people living with schizophrenia spectrum disorders and their family members from 3 study sites distinct in culture and setting in developing new features and co-designing the mindLAMP app	Thematic analysis	Mental health services in the United States and India	COVID-19 led to an uptake of virtual therapy, presenting new challenges and opportunities for providers and people with schizophrenia, using technology; access to data may offer providers and people with schizophrenia new insight into illness and treatment, but too much data elicit discomfort; relevance and integrated experience increase engagement	18
Byrne et al [10] (Australia)	Focus groups with service users (n=12) and interviews with mental health clinicians (n=10); overall (N=22); clinician characteristics: age range 25-60 y; male: n=2; female: n=8	Explore patient and clinician-related acceptability of an mHealth[c] device to monitor stress for severe mental illness	Thematic analysis	Community youth mental health service, Australia	Self-monitoring improves insight; clinician monitoring as a benefit to treatment; privacy and data misuse concerns; ease of use; engaging design; procedural guidelines	20
de Angel et al [3] (the United Kingdom)	Three focus groups: 2 focus groups with patients (n=16) and 1 focus group with mental health clinicians (n=6); overall (N=22); characteristics of clinicians: mean age 36.7 y; male: n=1; female: n=5	To identify clinically meaningful targets for digital health research and to explore patient and clinician attitudes toward the use of remote monitoring technologies and identify any perceived barriers to and facilitators of using these methods in psychological treatments for depression	Thematic analysis	Improving Access to Psychological Therapies services, the United Kingdom	Promoters of change (internal and external); markers of change (internal and external)	14
Gotzl et al [37] (Germany)	Two focus groups with young people (n=8) and interviews with experts, including psychologists (n=2); overall (N=13); age and gender of psychologists not stated	To investigate the subjective needs, attitudes, and preferences of key stakeholders toward an AI-informed mHealth app	Mixed methods-only qualitative component considered: content analysis	This study formed part of the living laboratory “AI4U-Artifical Intelligence for personalized digital mental health promotion and prevention in youth”	Young peoples’ understanding of mental health; experts understanding of mental health in youth; opportunities and risks seen by experts; experts’ recommendations	17

Study	Participants	Aim	Data analysis	Setting	Results and themes	Critical Appraisal Skills Programme quality appraisal score (out of 20)
Reis and Maier [38] (Germany)	Interviews with mental health professional (N=15); mean age 35 y; male: n=9; female: n=6	To explore the application scenarios for artificial intelligence in mental health from the mental health professionals' perspective and to evaluate the implementation readiness of scenarios	Content analysis	Mental health facilities in Germany	Application scenarios of AI in mental health: preselection and scheduling; patient monitoring during waiting times; documentation of treatment sessions; diagnosis support; relapse prophylactics; emergency care	16

[a]PTSD: posttraumatic stress disorder.

[b]AI: artificial intelligence.

[c]mHealth: mobile health.

Results

Summary of Papers

A total of 10 papers were deemed eligible for inclusion in this review. A total of 3 studies were conducted in the United Kingdom, 1 in India, 1 in the United States, 1 across both the United States and India, 1 in Australia, 2 in Germany, and 1 in a global study. In total, 6 studies used thematic analysis, 1 used a grounded theory approach, 1 used qualitative descriptive analysis, and 2 used content analysis. Participants in 4 of the papers were health care professionals only, with the remaining 6 papers including health care professionals as well as other stakeholders, such as service users and their families, technology experts, and technology company owners. The findings were only included if they were explicitly associated with health care professionals. The number of health care professionals ranged from 2 to 53 (mean 17). The age of the mental health professionals where this was reported (6 papers) ranged from 22 to 72 years. Among the 5 papers that reported gender, 28 participants were male, and 42 were female. Owing to the high number of participants in the global web-based survey [34], these data are described separately, with 791 participants taking part, ranging in age from 25 to ≥65 years. Of the participants, 550 identified as male, 230 identified as female, and 11 identified as others.

Study Quality

The overall CASP quality appraisal scores are included in Table 1, and a breakdown of these scores is provided in Table 2. There was variation in the scores across the papers. Those given stronger scores tended to provide more detail as to why certain qualitative approaches were selected over others, provided details regarding the sample including why participants may have opted not to take part, and ethical considerations were reported. It should be noted that although some studies did make some reference to the relationship between researcher and participants, this was the area that scored lowest, with few studies referencing bias and considering the influence their own roles may have had on results and reporting. For papers that included other stakeholders alongside mental health professionals, higher scores were given if the results were written to clearly distinguish which themes were associated with which participant group.

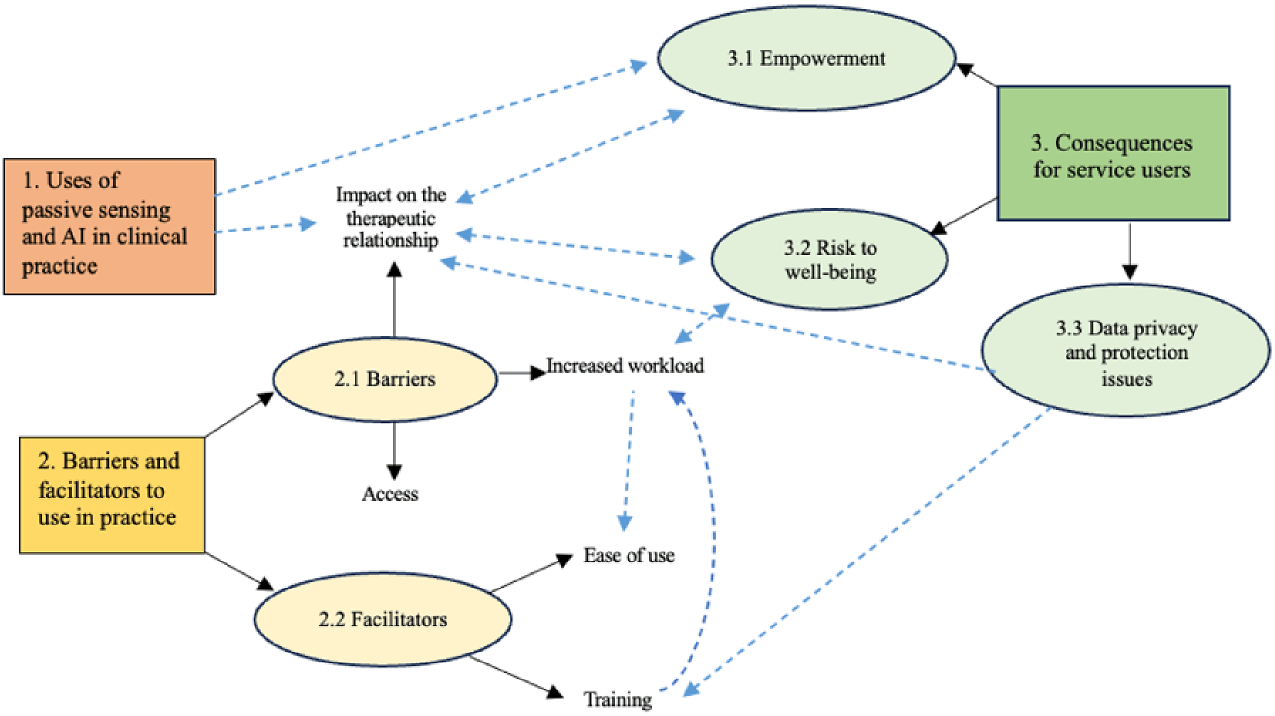
Table 2. Quality ratings on each of the Critical Appraisal Skills Programme domains.

Study	Aims	Method	Design	Recruitment	Data collection	Bias and reflexivity	Ethical issues	Data analysis	Clear findings	Value
Greer et al [32]	2	2	1	2	2	0	2	2	2	2
Ng et al [33]	2	2	2	2	2	1	1	2	2	2
Blease et al [34]	2	2	1	2	2	1	2	2	2	2
Thenral and Annamalai [35]	2	2	2	2	2	0	0	2	1	2
Dawoodbhoy et al [36]	2	2	2	2	1	1	1	2	1	2
Rodriguez-Villa et al [23]	2	2	2	2	2	0	2	2	2	2
Byrne et al [10]	2	2	2	2	2	2	2	2	2	2
de Angel et al [3]	2	2	2	1	1	0	1	2	1	2
Gotzl et al [37]	2	2	2	1	2	1	2	2	1	2
Reis and Maier [38]	2	2	2	2	2	0	0	2	2	2

Findings

Analysis of the data revealed three distinct but interrelated themes: (1) the use of passive sensing and AI in clinical practice, (2) barriers to and facilitators of use in practice, and (3) consequences for service users. A total of 5 subthemes were identified from the data. The themes, subthemes, and relationships between them are summarized in Figure 2.

Figure 2. Themes, subthemes, and the relationships between them. AI: artificial intelligence.



Theme 1: Uses of Passive Sensing and AI in Clinical Practice

Findings across the reviewed papers included how clinicians felt they could use passive sensing and AI in their practice. Passive sensing technology has been shown to be particularly useful, as the unobtrusive collection of objective service user data could offer a new information source [23]. This could provide insight into factors such as speech, social media use, and activity levels, which could be considered alongside self-report questionnaires and assessment tools to facilitate more accurate assessment of service users' mental health difficulties [35]. In assessing service user needs, it was also felt that these data could clarify discrepancies between self-report, observation, and psychometrics and validate service users' concerns [33]. It has been suggested that passive sensing data and AI could entirely replace some methods of assessment, such as questionnaires, to improve the clinical experience for both service users and clinicians [3,38]. Passive sensing data can reduce errors and biases in clinical decisions regarding diagnosis, treatment, and medication [34], as data-driven technologies may uncover correlations that humans cannot [36]:

> *The benefits would be greater reliability in diagnosis and prognosis, being able to choose specific customized treatment plans after analysis.* [Participant in Blease et al [34]]

Further suggestions were made as to how passive sensing and AI could be useful in therapeutic work, such as guiding productive discussions [33], setting treatment goals, delivering low-intensity support [3], tracking the efficacy of brief interventions [23], and encouraging ongoing engagement and regular self-reflection [38]. Furthermore, discussions were conducted about how AI's ability to process, connect, and make conclusions from large amounts of data could be used to risk-stratify service users according to their personal factors and needs [36] and support identification and awareness of early warning signs, thus reducing the risk of relapse of mental health difficulties [32-34,37]. As clinicians have access to these data, it was also felt that they could identify when to intervene [38], which may further reduce a service user's risk of deterioration in mental health [10]. Indeed, clinicians have reported that seeing a change in the data regarding a service user's speech and self-care habits would promote awareness of a decline in their well-being [3]. This was considered useful in community and ward environments, where staff may not always have eyes on service users [32], particularly for those who may lack insight into their difficulties or do not volunteer information themselves [23,32]:

> *...not all our patients will be able to say, "oh well I feel agitated" or be able to come out and say it, but within themselves all the physical, you know, changes are taking place so I think it's good, it will help us to see the covert, you know, things that are not outward that the patients cannot express.* [Participant in Greer et al [32]]

The idea that aspects of psychiatric work could gradually be replaced by AI was viewed as positive progression by some, but others were concerned that overreliance on AI and technology in practice may result in staff becoming deskilled [34]:

> *May lead to less skilled mental health staff.* [Participant in Blease et al [34]]

Theme 2: Barriers and Facilitators to Use in Practice

Overview

Throughout the papers, participants discussed the perceived barriers to and facilitators of using passive sensing and AI in mental health care. The barriers discussed included access, concerns about clinicians' workloads, and the potential negative impact on the therapeutic relationship. Facilitators included ease of use and training.

Subtheme 2.1: Barriers

Access

It was highlighted that technology is now readily available, and this was reported as a benefit to using passive sensing and AI in mental health practice as it may improve service user access to mental health care [34]. However, not all mental health services have sufficient access to technology [37] because of factors such as cost and the lack of the necessary infrastructure to support digital tools [3]. For example, participants reported that in India, most hospitals do not have access to the internet [35], and service users do not always have access to smartphones [23]. This would likely present significant barriers to health care professionals using such technology in mental health care. Therefore, it is important to consider the digital divide, that is, the gap between those who benefit from the digital age and those who do not [3]:

> *All these devices, technology, AI, etc., require high-speed internet...Majority of the hospitals do not have internet...patients who have basic livelihood issues cannot afford a device or internet.* [Participant in Thenral and Annamalai [35]]

Increased Workload

A further barrier discussed was the impact of passive sensing and AI could have on clinicians' workloads. Clinicians wondered about the amount of time and effort required to incorporate data flows into their practice and whether they would be required to review data before sessions [33], which could result in clinicians trawling through a significant amount of data to generate actionable insights [3]. Indeed, participants reported feeling "overwhelmed" when presented with passively collected data [23]:

> *I feel overwhelmed with the data to begin with.* [Participant in Rodriguez-Villa et al [23]]

Queries were also raised around documentation and whether the use of these tools would increase administrative work for clinicians [33,34]. Furthermore, it was suggested that clinicians may have to spend time with service users reviewing the use of devices and verifying data, which could take time away from evidence-based practices. If clinicians are alerted to changes in behavior that are the result of inaccurate readings, this could cause unnecessary alarm and waste clinicians' time [33]:

> *There's always so much to do...You're already kind of preparing for sessions, and at a certain point it's like, "How much am I treating or assessing what I'm seeing on a screen or on paper compared to just talking to someone and figuring stuff out together?"* [Participant in Ng et al [33]]

Managing risk was another concern raised, with participants wondering about their clinical responsibility for monitoring the data for risk issues [3,10,35] because responding to constant data streams would not be possible [23]. This is important because risk aversion is cited as a potential barrier to engagement [36].

In contrast, it has been suggested that passive sensing and AI have the potential to multiply resources, in that it offers a means of support to service users when health care professionals are unavailable [38]:

> *Having a system that treats people would be awesome. We cannot be everywhere, and the number of mental health professionals is too low.* [Participant in Reis and Maier [38]]

Impact on the Therapeutic Relationship

An issue that arose across studies was the impact that use of passive sensing and AI in practice could have on the therapeutic relationships. Some service users may prefer in-person consultations [35]; therefore, using digital tools and AI methods as a replacement of human contact could be determinantal to the therapeutic alliance [3] because of a loss of empathy and inaccurate interpretations of service users' presentations [34]:

> *Psychiatry is incompetent and incomplete without empathy. I doubt a machine could ever empathise with a live human being...I don't think affect of patient and mood, feelings, emotions can be analysed accurately.* [Participant in Blease et al [34]]

A potential lack of meaningful interactions between service users and clinicians [33] has led some to believe that service users may become resistant to or refuse treatment [34]. It was also suggested that, as service users are less accountable to clinicians, this could negatively impact motivation [3]. Furthermore, service users may become reliant on a device during treatment, and having this subsequently removed could have negative repercussions, including service users becoming mistrusting of services [3]. It was also highlighted that clinicians may not be able to fully trust the data they receive, as participants suggested that service users may influence these data deliberately [34]:

> *If a patient simulates a disease AI might not be able to determine it.* [Participant in Blease et al [34]]

Having said this, it was proposed that allowing service users to submit data to clinicians, who could then respond with recommendations, would enable remote support and continuity of care, which could strengthen the therapeutic relationship [23]. It appeared that the general consensus was that although digital tools may enhance practice, they should not replace service user or staff interactions, something which was viewed as integral to the therapeutic relationship [32].

Subtheme 2.2: Facilitators

Ease of Use

Throughout the studies, it was highlighted that to improve engagement with digital tools that allow for passive sensing and applied AI methods, the technology and systems would have to be relatively straightforward to use in terms of accessibility and convenience [3]. Health care professionals discussed that clinicians are very busy and would not have time to navigate complicated systems [10]:

> *I think it needs to be relatively simple...not over complicated and very easy to navigate.* [Participant in Byrne et al [10]]

To ensure ease of use, suggestions were made, such as including relevant stakeholders in the development of such technologies and related systems to ensure that they use accessible language [37] and presenting the data in a simple way that is easy to understand [3]:

> *Being able to have a really simple, easy way to compare the progress throughout the weeks of treatment. So, you would obviously be collecting a huge amount of data but if there was a way that we could somehow get, "Okay, you did an average of X amount of steps in week one, and your average sleep was X amount with waking up X amount of times."* [Participant in de Angel et al [3]]

Training

The importance of training to support clinicians in using digital tools, passively collected data, and applied AI methods in practice has been emphasized in most studies. This was discussed in the context of being given time to access the training as well as time to consider implementation [3]:

> *As long as we've had adequate training...And it's not just having the training, it's then having the time to think about that afterwards and incorporate it into your practice which would require a corresponding decrease in clinical work.* [Participant in de Angel et al [3]]

Clinicians may differ in technological literacy, and some may generally find technology challenging [35]. It was discussed that this technology will only be useful if clinicians understand it and feel comfortable using it [23]. For training to be adequate, it was suggested that clinicians would value clear procedures and guidance on when and how these digital technologies should be used [10], how to connect the data to their established clinical practice [33], and how to interpret data. Depending on the condition, certain markers of behavior could be interpreted positively or negatively [3]. Clinicians would also require clear guidelines regarding responsibility, interoperability, information governance, and potential risks [36]:

> *You're going to need [to] train them on how to balance all of these different pieces of data they have access to and how to prioritize the data. I think it would be especially important for new therapists coming on. It would probably be pretty overwhelming for some to have access to that much data and I think we would need to do like a standard operating procedure of how to [deal with] the information.* [Participant in Ng et al [33]]

Clinicians would further benefit from being informed of the evidence base around passive sensing and AI in mental health care [33], especially as the belief that there is a lack of studies to support the use of AI technology in health care could be a barrier to clinician engagement [35]:

> *There is also a lack of well-established trials and studies to understand the applicability of AI-related technology. They build solutions with no real-time applications.* [Participant in Thenral and Annamalai [35]]

Theme 3: Consequences for Service Users

Overview

Throughout the studies, findings on the consequences that passive sensing and AI in mental health care could have for service users were discussed. There appeared to be a positive notion that this could empower service users, although it was also acknowledged that there could be risks to service users' well-being. Concerns have also been raised regarding the protection and safety of service users' data.

Subtheme 3.1: Empowerment

It was suggested that passive sensing and AI could facilitate "knowledge transfer" and empower service users to understand how their actions, feelings, and thoughts are intertwined [37]. By increasing insight into mental health, self-monitoring allows service users to respond to symptoms and take action themselves [10]. Service users' monitoring and managing their mental health may involve connecting with other users and supporting one another, setting reminders to take medication, and responding to prompts to engage in helpful strategies [3]. Thus, service users' ability to monitor their mental health–related data can be empowering [33]:

> *A lot of the thoughts you have are that you are incapable, inadequate, cannot accomplish things. So, that [data] kind of directly speaks against that, right? "I am able to accomplish something, like reaching 10,000 steps a day.* [Participant in Ng et al [33]]

This self-management could increase awareness of early warning signs, reduce the risk of relapse, and therefore decrease demand for services, for example, by reducing hospital admissions [36].

Subtheme 3.2: Risk to Well-Being

In contrast, concerns have been raised about the accuracy of sensing technology, as making health care decisions based on unreliable sensors could potentially be harmful [3]. In addition, if service users have access to their health data, this could cause some to become hyperfocused on their data, catastrophize, or become disheartened by lack of progress or negative trends [33]. This was thought to be particularly pertinent to those who experience health anxiety [3] or paranoia, as "tracking" behavior could exacerbate symptoms [32]:

> *When you give this to a paranoid patient, they will think you are monitoring them. It will be so difficult*

> *to explain it to them to understand it that this is what you're monitoring...This paranoia could also lead to them not even wearing this.* [Participant in Greer et al [32]]

Ng et al [33] highlighted the importance of clinicians, suggesting alternate ways for service users to frame or interact with their data. This may be important in ensuring that recommendations delivered to service users do not arouse false expectations of users [37]:

> *I would see a risk if the app claimed: "if you go through these ten steps...then you are a different person" [laughs]...A good app would be characterized by the fact that the user does not internalize a problem centred perspective, but that he...gets the feeling: "I am okay."* [Participant in Gotzl et al [37]]

Another concern highlighted by participants working in mental health wards was that service users could harm themselves using a digital device. For example, if an armband that stretches it could be used as a ligature, the design of devices is an important consideration [32]:

> *...how far does it stretch, can you put it round your neck? Well, that might be an issue, you know, ligatures.* [Participant in Greer et al [32]]

> *...should be something that they cannot use as a weapon, like, there shouldn't be any metal or something that they can use to self-harm. [Participant in Greer et al]* [Participant in Greer et al [32]]

Subtheme 3.3: Data Privacy and Protection Issues

Participants reported that in practice, they will often recommend apps to service users without reviewing privacy policies, citing a lack of time as the reason for not investigating this further [23]. However, in most studies, concerns have been raised regarding privacy in relation to passive sensing data. It has been suggested that the collection of personal data through digital devices that allow passive sensing could increase the risk of loss of confidentiality and misuse of data [34,36], which could negatively impact therapeutic relationships [10]. In line with this, it was felt that service users would have less control over what they chose to share, which may feel uncomfortable for service users and lead clinicians to feel as though they are invading their privacy [10,23]:

> *...it does feel like it is personal information and to share all the details about their sleep and their activity levels—that could be quite tricky for some of them to share openly and knowing that we kind of can access it without them knowing or without them being there.* [Participant in Byrne et al [10]]

Data management was therefore seen as an important consideration, and it was highlighted that service users should be given choice over what they share and be made aware of who could access the data, what will happen if their data are leaked [3], how their data will be kept private and secure, and what their data will be used for [10]:

> *I think it's about having a conversation with the client at the beginning about boundaries really, and once it's clearer, how the information will be shared and can be shared, then you can...Normally put in those boundaries and then you can understand those concerns.* [Participant in de Angel et al [3]]

It is important to ensure that service users have capacity when making these decisions, as mental states can change and influence decision-making [32].

> *um, making sure they understand completely, cos some people are more paranoid on days...than other days so it could be they're fine for 5 days then the sixth day they're really paranoid.* [Participant in Greer et al [32]]

Discussion

Principal Findings

Across the papers reviewed, multiple ways in which passive sensing technology and applied AI methods could augment mental health care were identified, such as supporting service users in managing their mental health, improving diagnostic accuracy, monitoring treatment trajectories, and increasing access to timely support, thereby reducing the risk of relapse of mental health difficulties. Indeed, research has shown that passive data and AI methods have the potential to provide insight into service user behavior outside the clinic environment and provide real-time detection of behavioral anomalies, which could allow early identification and intervention before a deterioration in mental health [39]. However, despite the potential benefits, concerns have been raised that clinicians could become overreliant on digital technology in practice [40]. This could have negative consequences, as participants discussed that they may not be able to fully trust the data they receive because of service user influence and inaccurate sensors. Therefore, overreliance on inaccurate data can lead to misdiagnoses or missed diagnoses. Thus, decision-making should not be delegated to technology alone [41], and it is important for clinicians to acknowledge the limitations of objective data collection and applied AI methods to avoid tension between service users and clinicians [42]. This is particularly important, as research has shown that discrepancies between experience and tracking data can lead to upset, confusion, and disengagement [43], which may negatively impact the therapeutic relationship.

The influence that the use of passive sensing and applied AI methods could have on the therapeutic relationship was further discussed across papers. Although service users should feel empowered to make choices and manage their own mental health, access to human in-person support is deemed necessary. This reflects concerns that the use of AI in health care could lead to neglect of the therapeutic aspects of in-person consultation, such as consideration of motivation and self-advocacy, attendance to nonverbal cues, and social connection that can be provided by in-person clinical contact [44]. Fears were further raised that the absence of a therapeutic relationship may lead service users to disengage or refuse mental health care altogether. Research suggests that a therapeutic alliance can exist between a person seeking change and a change agent, which does not necessarily have to be a human health care professional, with digital tools and apps themselves having the potential to act as change agents [45].

Concerns have been raised across studies that service users may notice a decline in their mental health if they were to monitor aspects of their behavior and interpret subsequent passively collected data in such a way that increases anxiety or results in demotivation. Research has shown that tracking behavior can reduce enjoyment in walking-based activities [46] and increase eating disorder symptomology [47]. However, research has also found that the use of digital devices that allow passive sensing, such as wearables, can be a positive experience, with multiple psychological benefits identified by users, including increased sense of motivation and accountability [48]. Individual differences are therefore important for clinicians to consider, as certain characteristics may impact a service user's ability to interpret their health data in a helpful way. For example, research has shown that high health literacy supports the understanding of passively collected health data and how to use it to work toward goals [49].

Across studies, clinicians discussed the impact that use of passive sensing and AI could have on their workload. Concerns appeared to be around reviewing significant amounts of data to identify clinically relevant information and risk monitoring. However, previous research has suggested that AI may in fact reduce clinicians' workloads, as less time will be required to read through notes to understand a service user's history, particularly because certain AI methods, such as natural language processing, could be applied to patient notes to summarize important information [50]. Furthermore, machine learning methods can facilitate work by highlighting previously inaccessible or less understood symptoms and patterns [6]. It has also been suggested that data received by clinicians regarding a service user's behavior may allow them to identify those most in need of support and prioritize their workload, thus using their time more effectively [51]. To reduce concerns about increased workload, it would be useful for clinicians to receive data in a user-friendly format, allowing seamless access to relevant information. If devices and associated systems are not considered user-friendly and there are multiple technical issues, this will likely result in frustration and reluctance to engage [52]. Along with ease of use, training was discussed as a means to encourage clinicians to engage with devices that allow passive sensing and applied AI methods in their practice. Ways to make training useful for clinicians included ensuring that clinicians have access to clear guidance around incorporating data flows into their practice, managing risk issues, and data privacy and protection procedures. The latter is especially pertinent, as concerns about data security were a reoccurring theme throughout studies. Transparent guidelines will need to be developed, and codes of practice enforced around storage, ownership, and sharing of data [52]. However, it has been suggested that concerns about confidentiality of data may always remain; therefore, to facilitate engagement, the perceived value to clinicians and service users will need to outweigh these concerns [53]. As discussed in the reviewed studies, training should involve increasing awareness of the evidence base so that clinicians can understand the cost-benefits of engaging in passive sensing and AI in practice.

The final key issue is access. As highlighted in this review, access to technology could pose a barrier to engagement at both the service user and clinician level. For example, studies conducted in India have highlighted that not all hospitals offering mental health care have access to the internet. Therefore, considering the digital context within low- and middle-income countries, it is important to create digital-based mental health interventions intended for a global rollout. Indeed, Lee et al [54] highlighted that methods such as machine learning have the potential to advance health equity by supporting opportunities for equality in patient outcomes, performance, and resource allocation.

Strengths and Limitations

Meta-synthesis allows for greater scope and generalizability than individual primary studies [55]. However, as data are transposed into third-order constructs, there is potential for the findings to move away from the empirical, conceptual, and theoretical contexts of primary qualitative studies [56]. Of the papers included, 2 used content analysis, which is a more descriptive approach to coding and data interpretation (Vaismoradi et al [57]). Thus, the findings may have been more heavily influenced by studies that used more robust qualitative methods, such as thematic analysis, which can provide a more detailed and nuanced account of the data (Braun and Clarke [58]). Furthermor e, the process and methods of meta-synthesis are heavily influenced by the focus and expertise of the authors, meaning that some concepts and theories may not have been considered. This limitation was managed through discussion with the research team on coding and themes as well as remaining attuned to personal perspectives that could introduce bias [59].

Efforts were made to include all eligible studies in this review and to avoid neglecting potentially important findings, such as checking the reference lists of all papers and searching the databases again at a later date to identify further studies that might have been published. However, it is possible that some studies were overlooked, particularly as the terminology in this research area can be diverse and studies were only included if they were published in the English language in peer-reviewed journals, meaning that important contributions to the literature may have been missed because of language and publication bias. The included studies were conducted across different mental health settings, such as primary care and inpatient settings, and across different countries. It is important to note that the health care systems and services within them differ globally, so the generalizability of the results may be limited. However, meta-synthesis of qualitative studies can transform findings into highly abstracted and generalizable theoretical frameworks [21].

Future Directions

Considering the findings from this review and wider research in this area, a key barrier to implementing digital technology innovations is end user perceptions rather than technology innovation itself [3]. Therefore, it will be important for future research to gain a deeper understanding of service user views as well as other stakeholders, such as policy makers. Further research into the efficacy of passive sensing and AI in mental health care is necessary to build an evidence base that would support the scaling up of these approaches to routine service delivery. Real-world studies implementing passive sensing and AI in practice are needed to understand the contextual factors that impact uptake, which will be useful to gain knowledge that can support the development of implementation frameworks [60].

Clinical Implications

These findings suggest that although clinicians are open-minded about the use of passive sensing and applied AI methods in mental health care, factors such as service user well-being, clinicians' workloads, and the therapeutic relationship need to be considered. It is important to involve both service users and clinicians in the development of digital technologies and systems to ensure their ease of use. The development of policies, training, and clear guidelines on the use of passive sensing and AI in mental health care, including risk management and data security procedures, will also be key to facilitating clinician engagement and wide-scale adoption. Means for clinicians and service users to provide feedback on how the use of passive sensing and AI in practice is being received should also be considered, allowing reflection on any impact there might be on the therapeutic relationship, service user well-being, and clinicians' workloads.

Acknowledgments

JF is supported by a UK Research and Innovation Future Leaders Fellowship (MR/T021780/1). SB is supported by the National Institute for Health and Care Research Professorship (NIHR300794) and acknowledges the support from the Manchester Biomedical Research Centre (NIHR203308).

Conflicts of Interest

SB is a Director and shareholder of CareLoop Health Ltd, which develops and markets digital therapeutics for schizophrenia and a digital screening app for postnatal depression. SB also reports research funding from the National Institute for Health and Care Research and the Wellcome Trust.

References

1. World mental health report: transforming mental health for all. World Health Organisation. 2022 Jun 16. URL: https://www.who.int/publications/i/item/9789240049338 [accessed 2024-01-05]
2. Wilson RL, Higgins O, Atem J, Donaldson AE, Gildberg FA, Hooper M, et al. Artificial intelligence: an eye cast towards the mental health nursing horizon. Int J Ment Health Nurs 2023 Jun;32(3):938-944. [doi: 10.1111/inm.13121] [Medline: 36715172]
3. de Angel V, Adeleye F, Zhang Y, Cummins N, Munir S, Lewis S, et al. The feasibility of implementing remote measurement technologies in psychological treatment for depression: mixed methods study on engagement. JMIR Ment Health 2023 Jan 24;10:e42866 [FREE Full text] [doi: 10.2196/42866] [Medline: 36692937]
4. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021 Oct;20(3):318-335 [FREE Full text] [doi: 10.1002/wps.20883] [Medline: 34505369]
5. Chekroud AM, Bondar J, Delgadillo J, Doherty G, Wasil A, Fokkema M, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry 2021 Jun;20(2):154-170 [FREE Full text] [doi: 10.1002/wps.20882] [Medline: 34002503]
6. Koutsouleris N, Hauser TU, Skvortsova V, De Choudhury M. From promise to practice: towards the realisation of AI-informed mental health care. Lancet Digit Health 2022 Nov;4(11):e829-e840 [FREE Full text] [doi: 10.1016/s2589-7500(22)00153-4]
7. Fisher CE, Appelbaum PS. Beyond googling: the ethics of using patients' electronic footprints in psychiatric practice. Harv Rev Psychiatry 2017;25(4):170-179 [FREE Full text] [doi: 10.1097/hrp.0000000000000145]
8. Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. Annu Rev Clin Psychol 2017 May 08;13:23-47 [FREE Full text] [doi: 10.1146/annurev-clinpsy-032816-044949] [Medline: 28375728]
9. Spadaro B, Martin-Key NA, Bahn S. Building the digital mental health ecosystem: opportunities and challenges for mobile health innovators. J Med Internet Res 2021 Oct 13;23(10):e27507 [FREE Full text] [doi: 10.2196/27507] [Medline: 34643537]
10. Byrne S, Tohamy A, Kotze B, Ramos F, Starling J, Karageorge A, et al. Using a mobile health device to monitor physiological stress for serious mental illness: a qualitative analysis of patient and clinician-related acceptability. Psychiatr Rehabil J 2022 Sep;45(3):219-225 [FREE Full text] [doi: 10.1037/prj0000514] [Medline: 35298226]
11. Bucci S, Lewis S, Ainsworth J, Haddock G, Machin M, Berry K, et al. Digital interventions in severe mental health problems: lessons from the Actissist development and trial. World Psychiatry 2018 Jun;17(2):230-231 [FREE Full text] [doi: 10.1002/wps.20535] [Medline: 29856569]

12. Zhang X, Lewis S, Chen X, Berry N, Bucci S. Mental health professionals views and the impact of COVID-19 pandemic on implementing digital mental health in China: a nationwide survey study. Internet Interv 2022 Dec;30:100576 [FREE Full text] [doi: 10.1016/j.invent.2022.100576] [Medline: 36185346]
13. Fiske A, Henningsen P, Buyx A. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J Med Internet Res 2019 May 09;21(5):e13216 [FREE Full text] [doi: 10.2196/13216] [Medline: 31094356]
14. Rodriguez F, Scheinker D, Harrington RA. Promise and perils of big data and artificial intelligence in clinical medicine and biomedical research. Circ Res 2018 Dec 07;123(12):1282-1284 [FREE Full text] [doi: 10.1161/circresaha.118.314119]
15. Chancellor S, De Choudhury M. Methods in predictive techniques for mental health status on social media: a critical review. NPJ Digit Med 2020;3:43 [FREE Full text] [doi: 10.1038/s41746-020-0233-7] [Medline: 32219184]
16. Rubeis G. iHealth: the ethics of artificial intelligence and big data in mental healthcare. Internet Interv 2022 Apr;28:100518 [FREE Full text] [doi: 10.1016/j.invent.2022.100518] [Medline: 35257003]
17. Berry N, Lobban F, Bucci S. A qualitative exploration of service user views about using digital health interventions for self-management in severe mental health problems. BMC Psychiatry 2019 Jan 21;19(1):35 [FREE Full text] [doi: 10.1186/s12888-018-1979-1] [Medline: 30665384]
18. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med 2019 Sep;25(9):1337-1340 [FREE Full text] [doi: 10.1038/s41591-019-0548-6] [Medline: 31427808]
19. Wilson R. Why nurses need to be at the centre of new developments in digital technology. Nurse Res 2020;28(2):6-8. [doi: 10.7748/nr.28.2.6.s2]
20. Lachal J, Revah-Levy A, Orri M, Moro MR. Metasynthesis: an original method to synthesize qualitative literature in psychiatry. Front Psychiatry 2017;8:269 [FREE Full text] [doi: 10.3389/fpsyt.2017.00269] [Medline: 29249996]
21. Mohammed MA, Moles RJ, Chen TF. Meta-synthesis of qualitative research: the challenges and opportunities. Int J Clin Pharm 2016 Jun;38(3):695-704. [doi: 10.1007/s11096-016-0289-2] [Medline: 27052213]
22. Hasson-Ohayon I, Lysaker PH. Special challenges in psychotherapy continuation and adaption for persons with schizophrenia in the age of coronavirus (COVID-19). Counsel Psychol Q 2020 Jun 16;34(3-4):577-585 [FREE Full text] [doi: 10.1080/09515070.2020.1781595]
23. Rodriguez-Villa E, Rozatkar AR, Kumar M, Patel V, Bondre A, Naik SS, et al. Cross cultural and global uses of a digital mental health app: results of focus groups with clinicians, patients and family members in India and the United States. Glob Ment Health (Camb) 2021;8:e30 [FREE Full text] [doi: 10.1017/gmh.2021.28] [Medline: 34512999]
24. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009 Jul 21;339:b2535 [FREE Full text] [doi: 10.1136/bmj.b2535] [Medline: 19622551]
25. Richard Landis J, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174 [FREE Full text] [doi: 10.2307/2529310]
26. CASP Checklists. URL: https://casp-uk.net/casp-tools-checklists/ [accessed 2024-01-19]
27. Long HA, French DP, Brooks JM. Optimising the value of the critical appraisal skills programme (CASP) tool for quality appraisal in qualitative evidence synthesis. Res Method Med Health Sci 2020 Aug 06;1(1):31-42 [FREE Full text] [doi: 10.1177/2632084320947559]
28. Boeije HR, van Wesel F, Alisic E. Making a difference: towards a method for weighing the evidence in a qualitative synthesis. J Eval Clin Pract 2011 Aug;17(4):657-663 [FREE Full text] [doi: 10.1111/j.1365-2753.2011.01674.x] [Medline: 21504515]
29. Chrastina J. Meta-synthesis of qualitative studies: background, methodology and applications. In: Proceedings of the NORDSCI International Conference. 2018 Presented at: NORDSCI International Conference; July 17, 2018; Helsinki, Finland.
30. Braun V, Clarke V. Successful Qualitative Research: A Practical Guide for Beginners. Thousand Oaks, CA: SAGE Publications; 2013.
31. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008 Jul 10;8:45 [FREE Full text] [doi: 10.1186/1471-2288-8-45] [Medline: 18616818]
32. Greer B, Newbery K, Cella M, Wykes T. Predicting inpatient aggression in forensic services using remote monitoring technology: qualitative study of staff perspectives. J Med Internet Res 2019 Oct 19;21(9):e15620 [FREE Full text] [doi: 10.2196/15620] [Medline: 31538943]
33. Ng A, Kornfield R, Schueller SM, Zalta AK, Brennan M, Reddy M. Provider perspectives on integrating sensor-captured patient-generated data in mental health care. Proc ACM Hum Comput Interact 2019 Nov;3(CSCW):115 [FREE Full text] [doi: 10.1145/3359217] [Medline: 33585802]
34. Blease C, Locher C, Leon-Carlyle M, Doraiswamy M. Artificial intelligence and the future of psychiatry: qualitative findings from a global physician survey. Digit Health 2020 Oct 27;6:2055207620968355 [FREE Full text] [doi: 10.1177/2055207620968355] [Medline: 33194219]

35. Thenral M, Annamalai A. Challenges of building, deploying, and using AI-enabled telepsychiatry platforms for clinical practice among urban Indians: a qualitative study. Indian J Psychol Med 2021 Jul;43(4):336-342 [FREE Full text] [doi: 10.1177/0253717620973414] [Medline: 34385728]
36. Dawoodbhoy FM, Delaney J, Cecula P, Yu J, Peacock I, Tan J, et al. AI in patient flow: applications of artificial intelligence to improve patient flow in NHS acute mental health inpatient units. Heliyon 2021 May 12;7(5):e06993 [FREE Full text] [doi: 10.1016/j.heliyon.2021.e06993] [Medline: 34036191]
37. Götzl C, Hiller S, Rauschenberg C, Schick A, Fechtelpeter J, Fischer Abaigar U, et al. Artificial intelligence-informed mobile mental health apps for young people: a mixed-methods approach on users' and stakeholders' perspectives. Child Adolesc Psychiatry Ment Health 2022 Nov 17;16(1):86 [FREE Full text] [doi: 10.1186/s13034-022-00522-6] [Medline: 36397097]
38. Reis L, Maier C. Artificial intelligence in mental health: a qualitative expert study on realistic application scenarios and future directions. 2022 Presented at: SIGMIS-CPR '22: 2022 Computers and People Research Conference; June 2-4, 2022; Atlanta, GA. [doi: 10.1145/3510606.3550209]
39. Barnett I, Torous J, Staples P, Sandoval L, Keshavan M, Onnela JP. Relapse prediction in schizophrenia through digital phenotyping: a pilot study. Neuropsychopharmacology 2018 Jul;43(8):1660-1666 [FREE Full text] [doi: 10.1038/s41386-018-0030-z] [Medline: 29511333]
40. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 2012;19(1):121-127 [FREE Full text] [doi: 10.1136/amiajnl-2011-000089] [Medline: 21685142]
41. Fry H. Hello World: How to be Human in the Age of the Machine. London, UK: Transworld Digital; 2018.
42. Matthews M, Abdullah S, Gay G, Choudhury T. Tracking mental well-being: balancing rich sensing and patient needs. Computer 2014 Apr;47(4):36-43 [FREE Full text] [doi: 10.1109/MC.2014.107]
43. Yli-Kauhaluoma S, Pantzar M. Seeking connectivity to everyday health and wellness experiences: specificities and consequences of connective gaps in self-tracking data. Digit Health 2018 May 30;4:2055207618779714 [FREE Full text] [doi: 10.1177/2055207618779714] [Medline: 29942640]
44. Brown JE, Halpern J. AI chatbots cannot replace human interactions in the pursuit of more inclusive mental healthcare. SSM Ment Health 2021 Dec;1:100017 [FREE Full text] [doi: 10.1016/j.ssmmh.2021.100017]
45. Tong F, Lederman R, D'Alfonso S, Berry K, Bucci S. Digital therapeutic alliance with fully automated mental health smartphone apps: a narrative review. Front Psychiatry 2022 Jun 22;13:819623 [FREE Full text] [doi: 10.3389/fpsyt.2022.819623] [Medline: 35815030]
46. Etkin J. The hidden cost of personal quantification. J Consum Res 2016 Apr;42(6):967-984 [FREE Full text] [doi: 10.1093/jcr/ucv095]
47. Simpson CC, Mazzeo SE. Calorie counting and fitness tracking technology: associations with eating disorder symptomatology. Eat Behav 2017 Aug;26:89-92 [FREE Full text] [doi: 10.1016/j.eatbeh.2017.02.002] [Medline: 28214452]
48. Ryan J, Edney S, Maher C. Anxious or empowered? a cross-sectional study exploring how wearable activity trackers make their owners feel. BMC Psychol 2019 Jul 03;7(1):42 [FREE Full text] [doi: 10.1186/s40359-019-0315-y] [Medline: 31269972]
49. McKinney P, Cox AM, Sbaffi L. Information literacy in food and activity tracking among parkrunners, people with type 2 diabetes, and people with irritable bowel syndrome: exploratory study. J Med Internet Res 2019 Aug 01;21(8):e13652 [FREE Full text] [doi: 10.2196/13652] [Medline: 31373277]
50. Lovejoy CA, Buch V, Maruthappu M. Technology and mental health: the role of artificial intelligence. Eur Psychiatry 2019 Jan;55:1-3. [doi: 10.1016/j.eurpsy.2018.08.004] [Medline: 30384105]
51. Lim HM, Teo CH, Ng CJ, Chiew TK, Ng WL, Abdullah A, et al. An automated patient self-monitoring system to reduce health care system burden during the COVID-19 pandemic in Malaysia: development and implementation study. JMIR Med Inform 2021 Feb 26;9(2):e23427 [FREE Full text] [doi: 10.2196/23427] [Medline: 33600345]
52. Kelly JT, Campbell KL, Gong E, Scuffham P. The internet of things: impact and implications for health care delivery. J Med Internet Res 2020 Nov 10;22(11):e20135 [FREE Full text] [doi: 10.2196/20135] [Medline: 33170132]
53. Jalali MS, Kaiser JP, Siegel M, Madnick S. The internet of things promises new benefits and risks: a systematic analysis of adoption dynamics of IoT products. IEEE Secur Priv 2019;17(2):39-48. [doi: 10.1109/msec.2018.2888780]
54. Lee EE, Torous J, De Choudhury M, Depp CA, Graham SA, Kim HC, et al. Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. Biol Psychiatry Cogn Neurosci Neuroimaging 2021 Sep;6(9):856-864 [FREE Full text] [doi: 10.1016/j.bpsc.2021.02.001] [Medline: 33571718]
55. Malterud K. Qualitative Metasynthesis: A Research Method for Medicine and Health Sciences. London, UK: Routledge; 2019.
56. Sim J, Mengshoel AM. Metasynthesis: issues of empirical and theoretical context. Qual Quant 2022 Sep 06;57(4):3339-3361 [FREE Full text] [doi: 10.1007/s11135-022-01502-w]
57. Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: implications for conducting a qualitative descriptive study. Nurs Health Sci 2013 Sep;15(3):398-405. [doi: 10.1111/nhs.12048] [Medline: 23480423]
58. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]

59. Finfgeld-Connett D. Generalizability and transferability of meta-synthesis research findings. J Adv Nurs 2010 Feb;66(2):246-254 [FREE Full text] [doi: 10.1111/j.1365-2648.2009.05250.x] [Medline: 20423407]
60. Gama F, Tyskbo D, Nygren J, Barlow J, Reed J, Svedberg P. Implementation frameworks for artificial intelligence translation into health care practice: scoping review. J Med Internet Res 2022 Jan 27;24(1):e32215 [FREE Full text] [doi: 10.2196/32215] [Medline: 35084349]

Abbreviations

AI: artificial intelligence
CASP: Critical Appraisal Skills Programme
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by J Torous; submitted 02.06.23; peer-reviewed by JD Carrier, K Cho; comments to author 04.08.23; revised version received 30.10.23; accepted 01.11.23; published 23.01.24.

Please cite as:
Rogan J, Bucci S, Firth J
Health Care Professionals' Views on the Use of Passive Sensing, AI, and Machine Learning in Mental Health Care: Systematic Review With Meta-Synthesis
JMIR Ment Health 2024;11:e49577
URL: https://mental.jmir.org/2024/1/e49577
doi:10.2196/49577
PMID:38261403

Review

Effectiveness of Online and Remote Interventions for Mental Health in Children, Adolescents, and Young Adults After the Onset of the COVID-19 Pandemic: Systematic Review and Meta-Analysis

Linda Fischer-Grote[1,2], BSc, MSc; Vera Fössing[3], BSc, MSc; Martin Aigner[1,4], Dr med, PhD; Elisabeth Fehrmann[1], MSc, PhD; Markus Boeckle[3], MSc, PhD

[1]Department of Psychology and Psychodynamics, Karl Landsteiner University of Health Sciences, Krems, Austria
[2]Department of Clinical Psychology and Psychotherapy, University Hospital Krems, Krems, Austria
[3]Research Centre Transitional Psychiatry, Karl Landsteiner University of Health Sciences, Krems, Austria
[4]Department of Psychiatry for Adults, University Hospital Tulln, Tulln, Austria

Corresponding Author:
Elisabeth Fehrmann, MSc, PhD
Department of Psychology and Psychodynamics
Karl Landsteiner University of Health Sciences
Dr.-Karl-Dorrek-Straße 30
Krems, 3500
Austria
Phone: 43 6503032923
Email: elisabeth.fehrmann@kl.ac.at

Abstract

Background: The prevalence of mental illness increased in children, adolescents, and young adults during the COVID-19 pandemic, while at the same time, access to treatment facilities has been restricted, resulting in a need for the quick implementation of remote or online interventions.

Objective: This study aimed to give an overview of randomized controlled studies examining remote or online interventions for mental health in children, adolescents, and young adults and to explore the overall effectiveness of these interventions regarding different symptoms.

Methods: A systematic literature search was conducted according to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) guidelines using PubMed, PsycInfo, Psyndex, Embase, and Google Scholar. A meta-analysis was conducted using a random effects model to calculate overall effect sizes for interventions using standardized mean differences (SMDs) for postintervention scores.

Results: We identified 17 articles with 8732 participants in the final sample, and 13 were included in the quantitative analysis. The studies examined different digital interventions for several outcomes, showing better outcomes than the control in some studies. Meta-analyses revealed significant medium overall effects for anxiety (SMD=0.44, 95% CI 0.20 to 0.67) and social functioning (SMD=0.42, 95% CI –0.68 to –0.17) and a large significant effect for depression (SMD=1.31, 95% CI 0.34 to 2.95). In contrast, no significant overall treatment effects for well-being, psychological distress, disordered eating, and COVID-19–related symptoms were found.

Conclusions: The qualitative and quantitative analyses of the included studies show promising results regarding the effectiveness of online interventions, especially for symptoms of anxiety and depression and for training of social functioning. However, the effectiveness needs to be further investigated for other groups of symptoms in the future. All in all, more research with high-quality studies is required.

(JMIR Ment Health 2024;11:e46637) doi:10.2196/46637

KEYWORDS

COVID-19 pandemic; online/digital mental health intervention; e-mental health; anxiety; social functioning; depression; well-being; psychological distress; eating disorder; COVID-19 symptoms

Introduction

The high prevalence of psychological disorders in children and adolescents is well known, has been reported for a long time [1-4], and was estimated in 2015 to be 13.4% worldwide [4]. Psychological disorders in these age groups often show long-term impacts on adult life as well [2,5]. Childhood and adolescence are relevant periods for learning and brain maturing, possibly resulting in either a positive or negative impact [6]. Due to these developmental aspects, adolescents have, for example, been found to be especially vulnerable to addiction and addictive behavior [7].

Because of the COVID-19 pandemic and all the accompanying characteristics, prevalence rates of mental health issues have increased in the general population [8], adolescents [9], and young adults, who are among the groups most at risk of suffering from a COVID-19–related decrease in mental health [10-15]. A systematic review reported a lockdown-associated increase in anxiety and depressive symptoms in children and adolescents and an increase in sleep disorders; as risk factors, a priori mental illness and high media exposure were identified [16]. Increased stress levels are associated with respective containment measures [15].

Earlier research spanning from 1946 until 2020 showed an increased risk of depression and anxiety in children and adolescents due to loneliness and isolation [17]. This is an important aspect the current pandemic brought about in many countries [9,17] due to lockdowns and homeschooling, possibly impacting adolescents especially, as emotional support by peers is highly relevant at this age [9]. School closures resulted in a change in daily routines, which is particularly important for young people with mental health problems. Additionally, social isolation poses a risk factor for domestic violence, and an increase in worries related to the future, like school success, university access, and employment chances, has been noted [18].

However, the negative impact of the pandemic consists not only of an increase in mental health issues but also a significant impediment to the accessibility of treatment options, among other aspects, due to the need for social distancing [8,19]. Even before the pandemic, some groups of patients, like migrants [20], different groups of minorities [21], and people in remote areas [22], were difficult to reach through mental health programs. Prior to the pandemic, fewer than 50% of adolescents with depression used adequate services [23,24].

The sudden onset and accompanying restrictions of the pandemic made it even more necessary to increase the offers of online therapy to maintain the treatment of patients with mental health issues. These offers led to a sudden increase in therapists using online interventions [8,25-27], thereby seemingly decreasing perceived barriers by psychotherapists to use online or remote treatment options [8,26]. Nevertheless, the sudden switch also resulted in insecurities and the need for further guidance for therapists [27].

New media and online interventions have been developed and studied for years now [8], including in the context of children and adolescents with psychosomatic illnesses [28], with some studies even finding advantages of virtual therapy compared with face-to-face treatments [29] or at least similar outcomes [30]. Generally, reasonable user satisfaction and feasibility of interventions have been found [30], and studies show that therapeutic alliances can also be reached during videoconferencing, with clients rating bond and presence as equal to face-to-face settings [22]. Online help-seeking seems related to increased anonymity, accessibility, and inclusivity [31], and social media shows benefits for offering mental health care [32]. Applications developed to enhance mental health in children and adolescents show good acceptability [33]. Co-designed eHealth for adolescents appears to be associated with a more engaging and satisfying user experience [34-37].

Still, more research on effectiveness is needed [33], especially considering the sudden switch to online therapies due to COVID-19. Some reviews have been conducted regarding the effectiveness of online interventions for mental health related to the COVID-19 pandemic [38-41]; a review by Bonardi et al [38] focused on randomized controlled trials (RCTs) explicitly designed to address mental health issues related to COVID-19 and found some with promising effects but none for children or adolescents that met the inclusion criteria. Regarding web-based exercise interventions for depressive symptoms and anxiety, a review found no clear recommendations [39], while Valentine et al [40] found telehealth services for neurodevelopmental disorders to be primarily equal to control groups and focused on studies conducted before the onset of the COVID-19 pandemic. Yunus et al [41] found efficacy of digitalized interventions for depression in pregnant women and included studies from before the pandemic.

Nevertheless, it seems of high relevance to identify studies of interventions for mental health conducted after the onset of the pandemic with children, adolescents, and young adults, as persons of these age groups are at a higher risk of being negatively impacted by the pandemic in the long term. Whereas children, adolescents, and young adults can be considered “digital natives” [42], which might make them especially receptive to online interventions, younger individuals also seem to be especially vulnerable to negative aspects of digital media usage (eg, problematic smartphone use) [43]. Not only is it necessary to identify RCTs studying these aspects but one should also take into consideration the specific type of control condition that is used since different types of controls lead to different strengths of studies and especially in mobile health interventions, the combination of a variety of features might account for the resulting effects [44].

Thus, this systematic review and meta-analysis aimed to give a concise overview of studies examining the effectiveness of online or remotely delivered interventions or interventions delivered through digital media since the onset of the COVID-19 pandemic for specific mental health issues in children, adolescents, and young adults.

Methods

Search Strategy

To identify papers published since early 2020 (after the initial onset of the COVID-19 pandemic) until June 2023, a literature search based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) framework [45] was conducted in PubMed, PsycINFO, Psyndex, Embase, and Google Scholar. The detailed search parameters are depicted in Textbox 1. The reference search strategy was applied to locate additional relevant studies, and Google Scholar alerts were enabled. Multimedia Appendix 1 shows the PRISMA checklist, while Multimedia Appendix 2 shows the search strategy in more detail.

Textbox 1. Search parameters used in the literature search.

Databases

- PubMed
- PsycINFO
- Psyndex
- Embase

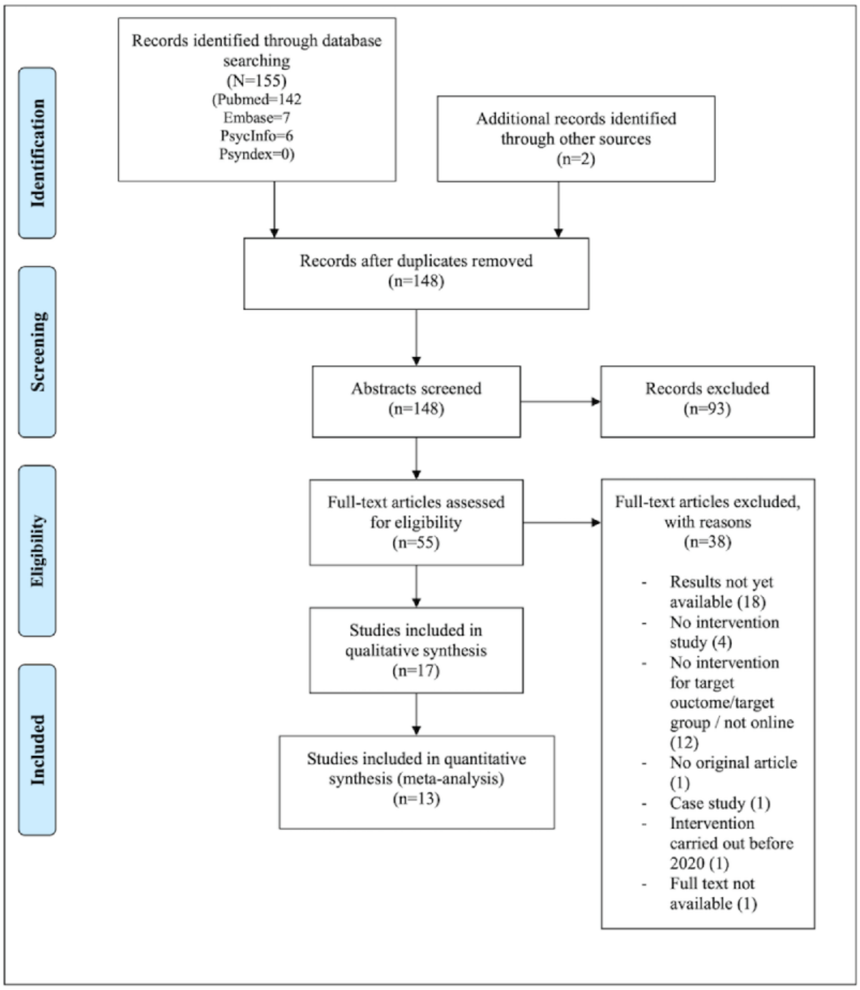
Search parameters

- (((depression) OR (anxiety) OR (mental health) OR (eating disorder) OR (stress) OR (sleeping disorder) OR (quality of life)) AND (((post covid) OR (long covid) OR (Covid) OR (Sars-cov-2)) AND ((adolescent) OR (child) OR (Juvenile) OR (teenager) OR (youth) OR (young adults) OR (emerging adult)) AND ((Psychology) OR (Psychotherapy) OR (psychiatry)) AND ((online) OR (digital) OR (video-based) OR (tele*)) AND ((effectiveness) OR (efficacy)) AND ((RCT) OR (Randomized controlled trial) OR (Case control) OR (observational cohort))

Study Selection Process

Before examining full texts, 2 authors (LFG, VF) independently screened the titles and abstracts for inclusion and exclusion criteria. In case of a mismatch between the 2 authors, all authors conferred and made a joint decision. See Figure 1 for the detailed exclusion process at each stage. Studies were included if they were (1) original, interventional studies; (2) published not earlier than 2020 (after the onset of the COVID-19 pandemic); (3) in peer-reviewed journals; (4) written in English or German; (5) focused on psychological or psychotherapy interventions that were delivered remotely (eg, online, via mobile app, or via telephone); (6) targeted at mental health issues like distress, depression or anxiety, psychological well-being, or quality of life (QoL); (7) conducted with children, adolescents, or young adults (from the age of 6 years to the age of 30 years, as emerging adulthood is defined as ages up to 30 years [46]). As outcome measures, we included standardized, validated, and reliable instruments designed for children, adolescents, and young adults to assess the listed mental health issues.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) flow diagram.



Statistical Analysis

Meta-analyses were conducted to examine the interventions' effectiveness using standardized mean differences (SMDs) as the outcome measure. The SMD compares postintervention scores between treatment and control groups. Only RCTs or pilot RCTs were included in the meta-analyses. A positive SMD indicates lower outcome scores in the treatment group than in the control group. Eligible studies were grouped by outcome type (anxiety, depression, well-being, disordered eating, psychological stress, social functioning, and COVID-19–related outcomes), and separate analyses were carried out for each group. Score polarity had to be reversed in 1 study [47]. Effect sizes were pooled using the "metafor" package [48] in the R environment. A random effects model was fitted to the data to account for variations in sample size, measures, and methodologies between the different studies. Heterogeneity was assessed using Higgins I^2 [49]. Interpretation of the effects sizes is based on Cohen *d* [50,51].

Additionally, a risk of bias assessment for studies included in the meta-analysis was conducted based on the Joanna Briggs Institute Critical Appraisal Checklist for Randomized Controlled Trials and on the Joanna Briggs Institute Critical Appraisal Tool for Quasi-Experimental Studies [52]. All statistical analyses were conducted in the R environment for statistical computing [53].

Results

Sample of Included Studies

A total of 155 articles were found in the initial database search process, and 2 additional studies were identified through the reference search strategy. Of the total number of articles, 9 duplicates had to be removed. We examined 55 articles at a full-text level. Of these, 20 articles were excluded since they were not original, were case studies, were not intervention studies, did not target the right outcome variables or groups, were not available, or were carried out before 2020. Additionally, 18 studies had to be excluded at the end of the search process as results still needed to be published for these trials. See Figure 1 for a detailed description of the inclusion and exclusion process.

The final sample of articles in June 2023 comprised 17 articles for the qualitative analysis, with an overall 8732 participants. Of these studies, 13 articles were included in the quantitative analysis. RCTs were reported in 16 articles, whereas 1 study [54] had only a quasiexperimental design with no control group. Only 1 study [55] explicitly compared the online intervention with an intervention conducted in a face-to-face setting. In addition, 1 study [56] was adapted to an online format during data collection due to the beginning of the COVID-19 pandemic.

Of the 17 studies included in the qualitative analysis, 5 (29%) [47,57-60] were conducted in the United States, 4 (24%) [54,61-63] were conducted in China, 2 (12%) each were conducted in Australia [56,64] and the United Kingdom [65,66], and 1 (6%) each was carried out in Canada [67], Italy [55], Iran [19], and Tunisia [68]. Of the included studies, 4 had an approximately (SD 15%) equal distribution of female and male participants [19,54,66,67]. In contrast, 9 had more female participants [47,57-59,61,62,64,65,68], 2 had more male participants [55,63], and 1 study was conducted with female participants only [56]. In addition, 1 additional article reported on 5 studies, of which 4 had more female participants, and 1 had an approximately equal distribution of female and male participants [60]. Of the included articles, 8 focused on children and adolescents [54,55,57-59,62,65,67], while another 8 included young adults [19,47,56,61,63,64,66,68]. The article reporting on 5 studies had samples with only adolescents and samples including young adults [60].

Characteristics of the included studies can be viewed in Tables 1 and 2, and the risk of bias assessment is depicted in Tables 3 and 4.

Table 1. Sample characteristics of the included studies.

Study	Sample size	Sample recruitment	Gender	Age	Country
Chen et al (2023) [62]	N=76	Research flyers through social media or school referral	Intervention: female=78.9%; male=21.1%; control: female=76.3%; male=23.7%	Intervention: 11-18 years, mean 16.45 (SD 1.52) years; control: 13-18 years, mean 16.37 (SD 1.24) years	China
Duan et al (2022) [54]	N=76	Online broadcasting platform	Female=56.58%; male=43.42%	10-12 years, mean 10.72 (SD 0.48) years	China
He et al (2022) [63]	N=148	Social media, online platforms, university communities, referred by counselors	Female=37.2%	17-21 years, mean 18.78 (SD 0.89) years	China
Krifa et al (2022) [68]	N=366	Health care students: class visits, posters in university, on website	Female=94%	Mean 20.74 (SD 1.64) years	Tunisia
Kutok et al (2021) [57]	N=80	Targeted Instagram advertisement	Female=59%	13-17 years, mean 15.3 (SD 1.35) years	US
Malboeuf-Hurtubise et al (2021) [67]	N=37	In 2 elementary schools	Female=43%; male=57%	Mean 8.18 years	Canada
Nicol et al (2022) [59]	N=18	Adolescents with depressive symptoms treated in practice-based research networks	Female=88%	13-17 years	US
Pavarini et al (2022) [65]	N=100	Advertisement on social media	Female=84%; male=14%; male (transgender)=1%; nonbinary=1%	16-18 years	UK
Prato et al (2022) [55]	N=40	Patients diagnosed with Tourette syndrome at a child and adolescent neuropsychiatry unit	Female=10%; male=90%	9-16 years, mean 13.5 (SD 2.0) years	Italy
Schleider et al (2022) [58]	N=2452	Instagram advertisements	Female=88.09% (biological sex)	13-16 years	US
Shabahang et al (2021) [19]	N=150	Convenient sample from Guilan University, Iran; online advertisement in college student social network	Female=51.33%; male=48.67%	Mean 24.7 (SD 5.4) years	Iran
Simonsson et al (2021) [66]	N=177	Students from the University of Oxford, UK	Female=64.4%	18-24 years (71.8%)	UK
Suffoletto et al (2021) [47]	N=52 (intervention: n=34; usual care group: n=18)	Young adults with a current mental health diagnosis recruited from primary care or a mental health clinic	Female=85%	Intervention: mean 18.7 (SD 0.42) years; usual care group: mean 18.7 (SD 0.48) years	US
Sun et al (2022) [61]	N=114	University students, online via WeChat-based flyers and websites targeting college students	Female=73.7%	Mean 22.21 (SD 2.67) years	China
Torok et al (2022) [64]	N=455	Social media: persons with suicidal thoughts in the past 12 months	Female=84.4%	18-25 years, mean 21.5 (SD 2.18) years	Australia
Yeager et al (2022) [60]	Study 1: n=2534; study 2: n=790; study 3: n=160; study 4: n=200; study 5: n=119; study 6: n=351	Character Lab Research Network, school students, university students (in specific courses)	Study 1: female=49%, male=49%, nonbinary=2%; study 2: female=64%, male=36%; study 3: female=72.3%, male=27.7%; study 4: female=81.5%, male=18.5%; study 6: similar to study 2	Study 1: 13-18 years; study 2: 17-≥21 years; study 3: 18-26 years; study 4: 18-32 years; study 5: 14-16 years; study 6: similar to study 2	US
Zhou and Wade (2021) [56]	N=100 (pre-COVID-19: n=41; during COVID-19: n=59)	University students at risk of developing an eating disorder	Female=100%	17-26 years; mean 19.85 (SD 2.01) years	Australia

Table 2. Study characteristics of the included studies.

Study	Study type	Intervention	Control condition[a]	eHealth technology	Target outcomes	Results
Chen et al (2023) [62]	RCT[b]	Online solution-focused brief therapy (SFBT), active intervention group: 38/76, 50%	No treatment, wait list control, 38/76, 50%	Teleconference	Primary outcome: anxiety; secondary outcomes: depressive symptoms and coping styles	Significant results in intervention group regarding anxiety, depression, and problem-oriented coping styles; depression levels significantly lower in intervention than in control group
Duan et al (2022) [54]	Quasiexperimental, no control group	Online Strength-informed Acceptance and Commitment Therapy (SACT), active intervention group: 76/76, 100%	No control group	Video conferencing system	Quality of life (QoL) and anxiety	Pre to post: significant reduction in anxiety but no significant increase in QoL; pre to 3-month follow-up: reduced anxiety and increased QoL
He et al (2022) [63]	RCT	CBT[c]-based mental health chatbot (XiaoE), active intervention group: 49/148, 33.1%	2 mHealth[d] minimal active controls: e-book, 49/148, 33.1%; general chatbot: 50/148, 33.8%	Main intervention: XiaoE, unguided CBT-based chatbot, 1 module a day for 1 week	Primary outcome: depressive symptoms; secondary outcomes: working alliance, usability, acceptability	Primary outcome: significant reduction in depression in intervention group compared with e-book and general chatbot group; secondary outcome: better working alliance and acceptability in intervention, no significant difference for usability
Krifa et al (2022) [68]	RCT	CARE[e] program: internet-based positive psychology intervention, active intervention group: 183/366, 50%	No treatment, wait list control: 183/366, 50%	8-week online self-program: lecture, videos, psychoeducation, practices	Stress, anxiety, depression, emotional regulation, optimism, hope, study engagement, well-being	Significant positive effects in all variables; significant improvement compared with control group
Kutok et al (2021) [57]	RCT	IMPACT[f], active intervention group: 36/80, 45%	Placebo minimal: enhanced web-based resources: 44/80, 55%	Video intervention plus app-based automated messaging; control: enhanced web-based resources	Cyberbullying: to reduce consequences of cyber victimization, to increase bystander intervention	Feasible, acceptable; improved bystander intervention and well-being in intervention group
Malboeuf-Hurtubise et al (2021) [67]	RCT	MBI[g] (16/37, 43.2%); P4C[h] (21/37, 56.8%); both group-based, delivered online	Comparison of 2 active intervention groups (comparative efficacy)	Teleconferencing platform	Anxiety, inattention symptoms, basic psychological need satisfaction (BPN) in the context of COVID-19	P4C: more impact on anxiety and inattention; MBI: better outcomes for BPN
Nicol et al (2022) [59]	Pilot RCT	CBT, active intervention group: 10/18, 55.6%	No treatment, wait list control, 1:1: 8/18, 44.4%	mHealth app with embedded conversational agent	Primary outcomes: depression severity, anxiety; secondary outcomes: feasibility, acceptability, usability	Reduction in symptom severity from moderate to mild in treatment group, no reduction in control group; usability, acceptability, feasibility high

Study	Study type	Intervention	Control condition[a]	eHealth technology	Target outcomes	Results
Pavarini et al (2022) [65]	RCT	Online peer support training course “Uplift Peer Support Training,” active intervention group: 50/100, 50%	No treatment, wait list control: 50/100, 50%	Zoom, smaller groups in breakout rooms or via WhatsApp	Primary outcomes: motivation to provide support, perceived support-giving skills, frequency of support provided, compassion toward others, connectedness with peers; secondary outcomes: mental well-being, emotional symptoms, self-efficacy, civic engagement	Primary outcomes: no difference regarding motivation, significant effects of training regarding other primary outcomes; secondary outcomes: significant effect of training
Prato et al (2022) [55]	RCT	Behavior therapy for youths with Tourette syndrome during COVID-19	Non-mHealth evidence-based active control: comparison of online vs face-to-face intervention: 20/40, 50% each	Video conference vs face-to-face	Tic symptoms, obsessive compulsive symptoms, ADHD[i] symptoms, anxiety, depressive symptoms	Both forms of delivery equally effective regarding most outcomes; online delivery more effective regarding depressive symptoms
Schleider et al (2022) [58]	RCT	Online single-session intervention (SSI) for depressive symptoms (behavioral activation SSI: 821/2452, 33.5% vs growth mindset SSI: 813/2452, 33.2% vs supportive therapy SSI as the control	Placebo active control: supportive control, (structurally similar [eg, matched in length]): 818/2452, 33.4%	Self-administered online intervention	Depressive symptoms, hopelessness, agency, generalized anxiety, COVID-19–related trauma, restrictive eating	Both active SSIs showed significantly better outcomes regarding depression, hopelessness, agency, and restrictive eating than control group; no difference between behavioral action and control group regarding generalized anxiety and COVID-19– related trauma symptoms but between growth mindset and control
Shabahang et al (2021) [19]	RCT	Video-based CBT intervention for COVID-19 anxiety, active intervention group: 75/150, 50%	No treatment, wait list control: 75/150, 50%	Self-administered video-based strategies, online booklet	COVID-19 anxiety, health anxiety, anxiety sensitivity, somatosensory amplification	Significant differences in outcomes between intervention and control groups; high intervention group participant satisfaction with the intervention
Simonsson et al (2021) [66]	RCT	Online, guided, 8-week mindfulness program, active intervention group: 88/177, 50%	No treatment, wait list control: 89/177, 50%	Online classes via Zoom,	Anxiety, depression	Larger reduction in anxiety in treatment group compared with control group; no difference regarding depression
Suffoletto et al (2021) [47]	Pilot RCT	Mobile Support Tool for Mental Health (MoST-MH), active intervention group: 34/52, 65.4%	mHealth minimally active control: enhanced usual care (eUC; weblink to psychoeducational videos): 18/52, 34.6%	Text messaging, web-based check-ins, video feedback (psychoeducation)	Mental health symptoms, mental health self-efficacy, mental health care use	MoST-MH: reduction in all symptoms except substance abuse; eUC: only reduction in general anxiety, family distress, hostility; no improvements regarding self-efficacy and care use in either group

Study	Study type	Intervention	Control condition[a]	eHealth technology	Target outcomes	Results
Sun et al (2022) [61]	RCT	Mindfulness-based mHealth intervention, active intervention group: 57/114, 50%	mHealth minimally active control (matched social support mHealth control): 57/114, 50%	Videoconferencing via Zoom, WeChat-based mini-program	Primary outcomes: anxiety, depression; secondary outcomes: mindfulness, social support, emotional suppression	Reduction in anxiety and depression and increase in mindfulness and social support in both groups; greater effect on anxiety through mindfulness intervention; greater engagement with and higher acceptability of mindfulness mHealth
Torok et al (2022) [64]	RCT	Self-guided smartphone app based on DBT[j], active intervention group: 228/455, 51.1%	Placebo active control, smartphone app with general information: 227/455, 49.9%	Smartphone app (LifeBuoy)	Primary outcome: suicidal ideation symptom severity; secondary outcomes: depression, generalized anxiety, distress, well-being	Significantly higher effects of intervention regarding suicidal ideation; no superior effects regarding secondary outcomes
Yeager et al (2022) [60]	RCT	Synergistic mindset intervention, active intervention group: study 1: 1208/2534, 47.7%; study 2: 387/790, 49%; study 3: 74/160, 46%; study 4: growth only, 52/200, 26%; stress only, 65/200, 32.5%; synergistic, 39/200, 19.5%; study 5: 61/119, 51.3%; study 6: 179/351, 51%	Placebo active control, study 1: 1326/2534, 52.3%; study 2: 403/790, 51.0%; study 3: 86/160, 54%; study 4: 44/200, 22%; study 5: 58/119, 48.7%; study 6: 172/351, 49%	Self-administered online training module	Studies 1 and 2: stress-related cognition; studies 3 and 4: cardiovascular reactivity; studies 4 and 5: psychological well-being; study 5: daily cortisol levels, academic success; study 6: anxiety levels during COVID-19 lockdowns	Improvements in outcomes greater in treatment group than in control group in all experiments
Zhou and Wade (2021) [56]	RCT	Online intervention to reduce disordered eating, active intervention group: 77/100, 77%	No treatment, assessment only control: 23/100, 23%	Online format introduced in April 2021	Disordered eating, body image flexibility, self-compassion, fear of self-compassion, negative affect	Significantly higher symptomology during COVID-19 than pre-COVID-19, active intervention significantly increased self-compassion compared with control, no other significant time x condition effects

[a]Typology of control groups based on Goldberg et al [44].

[b]RCT: randomized controlled trial.

[c]CBT: cognitive behavioral therapy.

[d]mHealth: mobile health.

[e]CARE: Coherence, Attention, Relationship, and Engagement.

[f]IMPACT: Intervention Media to Prevent Adolescent Cyber-Conflict Through Technology.

[g]MBI: mindfulness-based intervention.

[h]P4C: philosophy for children.

[i]ADHD: attention-deficit/hyperactivity disorder.

[j]DBT: dialectical behavior therapy.

Table 3. Risk of bias assessment for randomized controlled trials.

First author (year)	True randomization	Concealed allocation	Similar groups at baseline	Participants, personnel, or outcome assessors blinded to assignment	Identical treatment of groups	Follow-up: complete or full description	Analysis of participants in their groups	Outcome measurement: equal and reliable	Appropriate statistical analysis	Appropriate trial design and deviations accounted for
Randomized controlled trials										
Chen et al (2023) [62]	Yes	Yes	Yes	Participants and personnel: no; outcome assessors: yes	Yes	Yes	Yes	Yes	Yes	Yes
He et al (2022) [63]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Krifa et al (2022) [68]	Yes	Unclear	Yes	Unclear	Yes	Yes	Yes	Yes	Yes	Yes
Kutok et al (2021) [57]	Randomized but stratified by age and gender	Yes	Yes	Participants and personnel: no; outcome assessors: yes	Yes	Yes	Yes	Yes	Yes	Yes
Malboeuf-Hurtubise et al (2021) [67]	Unclear	Unclear	No	Unclear	Yes	No follow-up	Yes	Yes	Yes	No real control group, 2 interventions
Nicol et al (2022) [59]	Yes	Yes	Yes	Unclear	Yes	Yes	Yes	Yes	Yes	Yes
Pavarini et al (2022) [65]	Yes	Unclear	Yes	Unclear	More assessments in treatment group	Yes	Yes	Yes	Yes	Yes
Prato et al (2022) [55]	Yes	Unclear	Yes	Participants and personnel: no; outcome assessors: unclear	Yes	Yes	Yes	Yes	Yes	Yes
Schleider et al (2022) [58]	Unclear	Unclear	Yes	Unclear	Yes	Yes	Yes	Yes	Yes	Yes
Shabahang et al (2021) [19]	Yes	Unclear	Yes	Unclear	Yes	No follow-up	Yes	Yes	Yes	Yes
Simonsson et al, (2021) [66]	Yes	Unclear	Yes	Unclear	Yes	Yes	Yes	Yes	Yes	Yes
Suffoletto et al (2021) [47]	Unclear	Yes	Partially	Outcome assessors: yes	Yes	Yes	Yes	Yes	Yes	Yes
Sun et al (2022) [61]	Yes	Unclear	Unclear	Participants and research assistant: yes (at first)	Yes	Yes	Yes	Yes	Yes	Yes

First author (year)	True randomization	Concealed allocation	Similar groups at baseline	Participants, personnel, or outcome assessors blinded to assignment	Identical treatment of groups	Follow-up: complete or full description	Analysis of participants in their groups	Outcome measurement: equal and reliable	Appropriate statistical analysis	Appropriate trial design and deviations accounted for
Torok et al (2022) [64]	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Yeager et al (2022) [60]	Yes	Yes	Unclear	Yes	Yes	Unclear	Yes	Yes	Yes	Yes
Zhou and Wade (2021) [56]	Yes	Unclear	Yes	Unclear	Yes	Unclear	Yes	Yes	Yes	Sample size intervention and control unbalanced

Table 4. Risk of bias assessment for quasiexperimental studies.

First author (year)	Clear distinction cause and effect	Similar participant in comparison	Similar treatment of comparison group	Control group	Multiple measurement of outcome pre- and postintervention	Follow-up: complete or full description	Equal measurement of participants in comparisons	Reliable measurement of outcomes	Appropriate statistical analysis
Duan et al (2022) [54]	Yes	Not applicable	Not applicable	No	Unclear	Yes	Not applicable	Yes	Yes

Characteristics of Online Interventions Used in the Included Studies

In most (9/17, 53%) of the included studies [54-57,61,62,65-67], different versions of videoconferencing systems were used to deliver the interventions remotely (see Table 2). In an intervention targeted at cyberbullying (Intervention Media to Prevent Adolescent Cyber-Conflict Through Technology [IMPACT]), Kutok et al [57] added app-based automated messaging to their video-delivered intervention. Pavarini et al [65] added the possibility for smaller group discussions by using breakout rooms and WhatsApp for their online peer support training. The mindfulness-based mobile health intervention by Sun et al [61] was supplemented by a WeChat-based mini-program. Other interventions delivered remotely via videoconferencing were the online Strength-informed Acceptance and Commitment Therapy (SACT) [54], mindfulness-based interventions [66,67], philosophy for children (P4C) [67], behavior therapy for Tourette syndrome [55], an intervention to reduce disordered eating [56], and the online solution-focused brief therapy (SFBT), primarily to reduce symptoms of anxiety [62].

Next to these online interventions with teleconferencing systems, 4 studies used self-administered online interventions. Schleider et al [58] examined online single interventions for depressive symptoms, and Shabahang et al [19] targeted COVID-19–related anxiety with self-administered video-based strategies and online booklets. An 8-week self-program with lectures and videos was delivered as an intervention by Krifa et al [68], and Yeager et al [60] used self-administered online training to reduce stress-related symptoms. Text messaging, web-based check-ins, and video feedback with psychoeducation were applied in a study by Suffoletto et al [47] in their Mobile Support Tool for Mental Health (MoST-MH). A cognitive behavioral therapy (CBT)–based mental health chatbot (XiaoE) was used to reduce depressive symptoms by He et al [63]. One study [64] used a smartphone app (LifeBuoy) based on dialectical behavior therapy (DBT) to target suicidal ideation. In contrast, a second study [59] used an app with an embedded conversational agent based on CBT to primarily reduce depressive symptoms.

Not all studies reported on the feasibility and acceptability of their interventions. Those that did, however, found the intervention to be feasible [57,59] and acceptable [57-59,63], met with high satisfaction [19,62], and more accepted and engaging in the treatment group than in the control group [61].

Effectiveness of Online Interventions Regarding Mental Health Outcomes

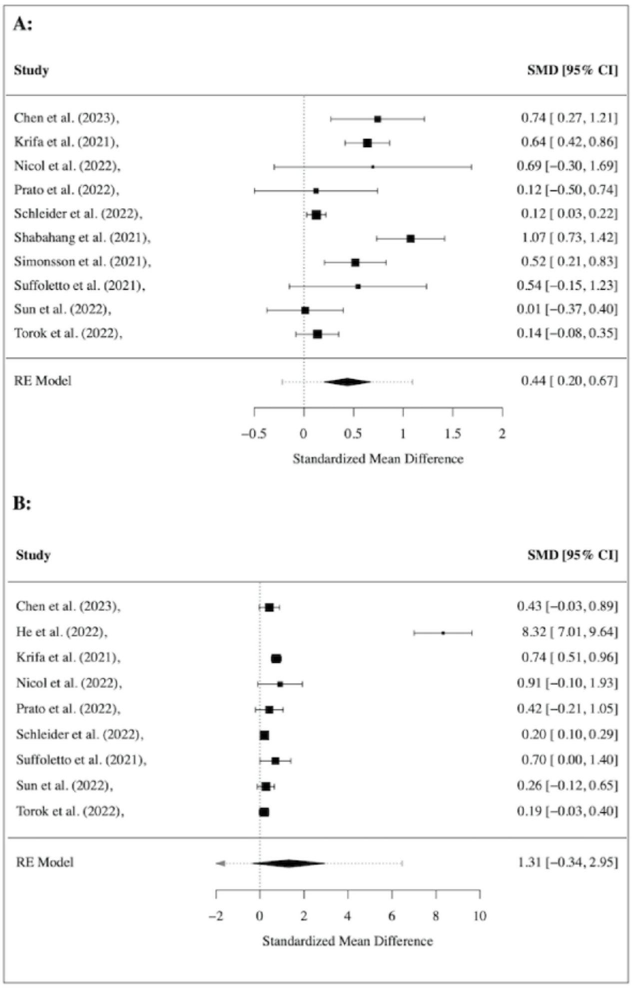
Mental health–related outcomes varied in the included studies (see Table 2). They included anxiety, depression, mental well-being, social functioning, COVID-19–related symptoms, cyberbullying, Tourette syndrome, disordered eating, suicidal ideation, and psychological stress, among others.

Anxiety

Several studies reported reduced anxiety [19,47,54,55,59-62,66-68]. The impact on anxiety was more prominent in some studies for the treatment group than for the control group [19,59-61,66,68]. In contrast, others found only partial differences [58], equal effects, or no differences between groups [55,64]. P4C had a more significant impact on anxiety than a mindfulness-based intervention in one study [67].

A meta-analysis of 10 studies, with 9 targeting generalized anxiety disorder and 1 targeting health anxiety, showed an overall significant positive effect of interventions in the form of decreased symptoms (SMD=0.44, 95% CI 0.20 to 0.67; I^2=82.9%). Figure 2A shows a forest plot of the observed outcomes.

Figure 2. Meta-analysis of treatment effect regarding (A) anxiety and (B) depression, shown using the overall and individual study standardized mean difference (SMD) and 95% CIs (those that include 0 show nonsignificant effects), where a positive effect size indicates a decrease in symptoms.



Depression

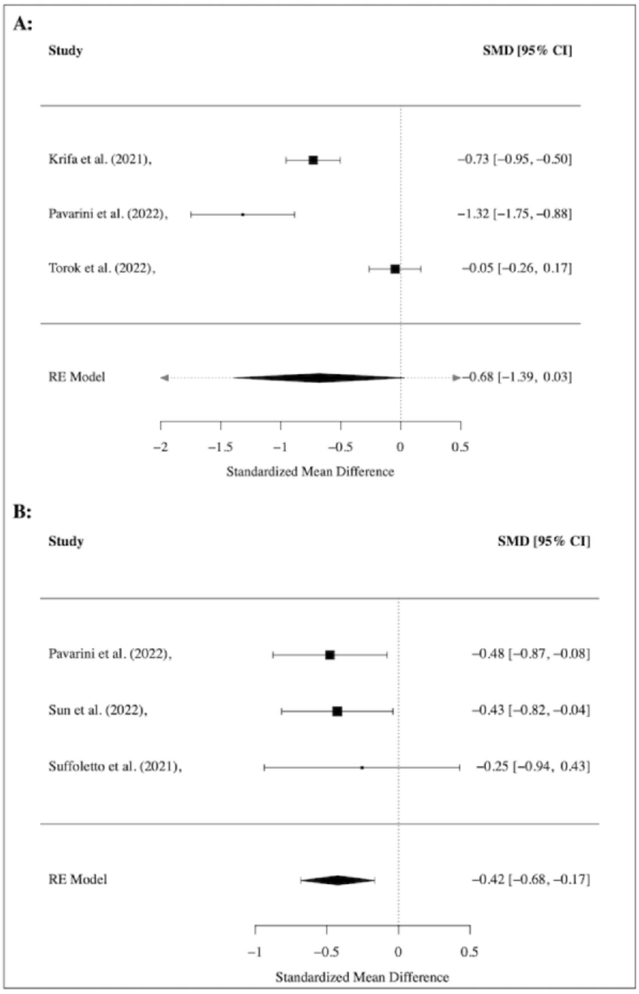
Reduced symptoms of depression were found in several studies [47,55,58,59,61-63,68], with superiority of the intervention group found in some [47,55,58,59,62,63,68]. The studies conducted by Sun et al [61], Simonsson et al [66], and Torok et al [64] found no superior effects of the treatment on depressive symptoms.

Nevertheless, a meta-analysis of 9 studies found a strong treatment effect (SMD=1.31, 95% CI 0.34 to 2.95; I^2=99.63%). The observed outcomes are depicted in a forest plot in Figure 2B.

Mental Well-Being, Quality of Life, Agency or Self-Efficacy

Several studies found increased well-being [47,57,60,65,67,68]. QoL was increased at the 3-month follow-up in 1 study [54]. Self-efficacy or agency was increased in some studies [58,65], while 1 study found no effect [47]. Only 3 studies [64,65,68] were eligible for a meta-analysis analyzing the treatment effect on well-being. However, no significant effect was shown in the meta-analysis (see Figure 3A).

Figure 3. Meta-analysis of treatment effect regarding (A) well-being and (B) social functioning, shown using the overall and individual study standardized mean difference (SMD) and 95% CIs (those that include 0 show nonsignificant effects), where a negative effect size indicates an increase of well-being and social functioning.

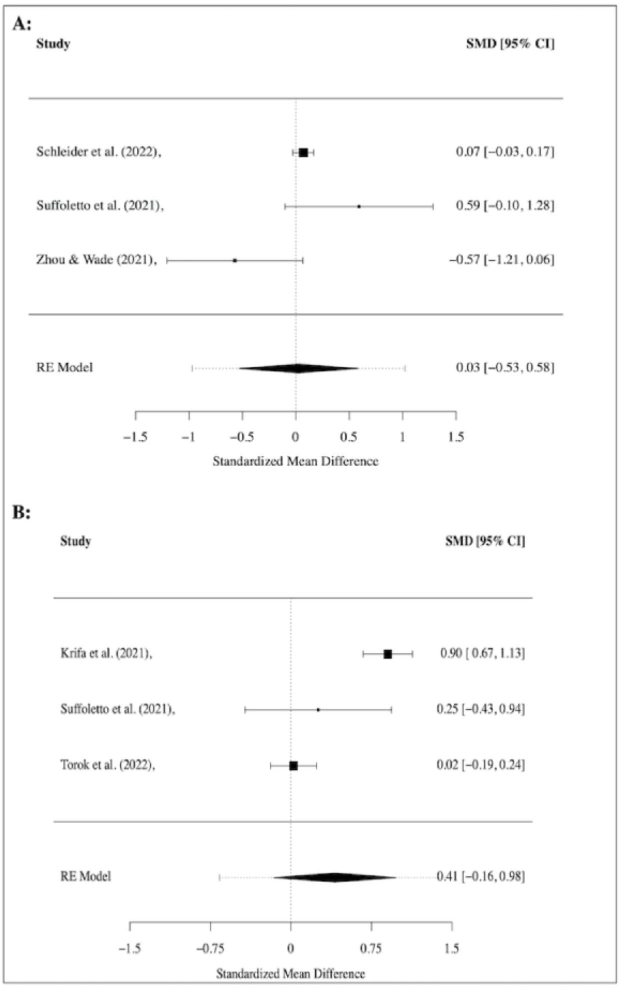


Other Main Outcomes

A study targeting cyberbullying increased bystander intervention in the treatment group [57], while another showed promising results regarding increased social support-giving skills, compassion toward others, and civic engagement, among other outcomes [65]. Tic and obsessive-compulsive symptoms in children and adolescents with Tourette syndrome were equally reduced via videoconference and face-to-face interventions [55]. Using a smartphone app, 1 study was able to significantly reduce suicidal ideation [64], and 1 study targeting disordered eating increased self-compassion through treatment. At the same time, no other effect was found [56].

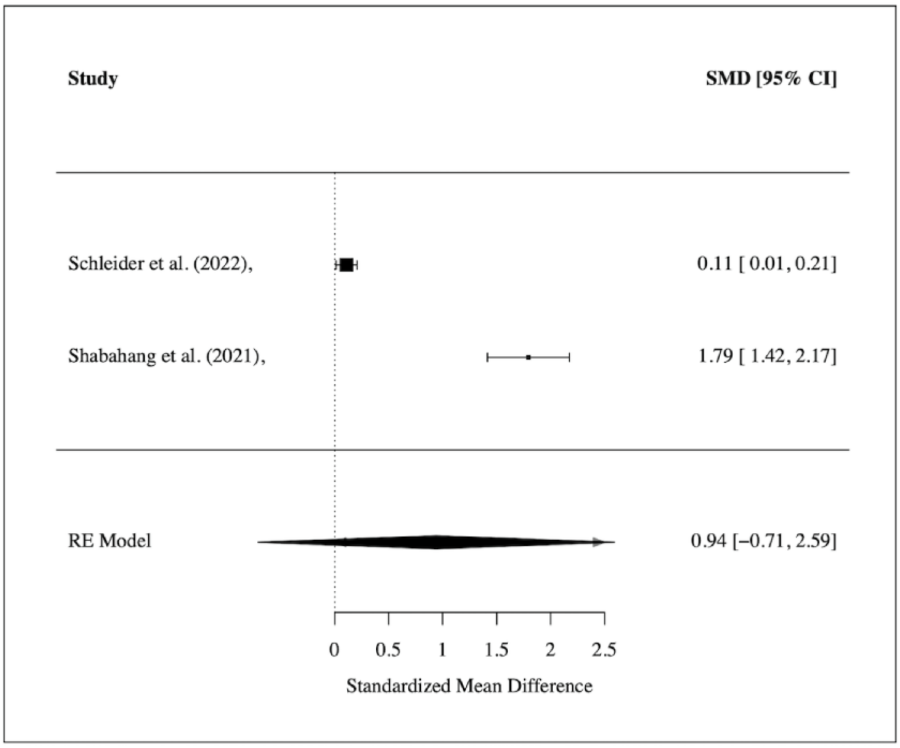
Meta-analyses were conducted for several of these outcomes. Regarding disordered eating (see Figure 4A), psychological stress (see Figure 4B), and COVID-19–related trauma or anxiety (see Figure 5), no significant treatment effects were found across studies, while a significant medium effect (SMD=–0.42, 95% CI –0.68 to –0.17; I^2=0.0%) was found for interventions targeting social functioning in 3 studies [47,61,65] (see Figure 3B). For the outcomes of attention and emotional functioning, more data were needed for the meta-analyses.

Figure 4. Meta-analysis of treatment effect regarding (A) disordered eating and (B) psychological stress, shown using the overall and individual study standardized mean difference (SMD) and 95% CIs (those that include 0 show nonsignificant effects), where a positive effect size indicates a decrease in symptoms.



XSL•FO
RenderX

Figure 5. Meta-analysis of treatment effect regarding COVID-19–related symptoms, shown using the overall and individual study standardized mean difference (SMD) and 95% CIs (those that include 0 show nonsignificant effects), where a positive effect size indicates a decrease in symptoms.



Discussion

Principal Findings

This systematic review and meta-analysis is the first of its kind investigating the effectiveness of online or remote interventions for psychological symptoms and disorders in children, adolescents, and young adults after the onset of the COVID-19 pandemic. We examined 17 studies conducted between the pandemic's start and June 2023 for the impacts of their online interventions. Despite the necessary fast development due to the increased need for remote interventions during the COVID-19 pandemic, the results are promising. All the studies observed positive effects on some of the outcomes they targeted through their remote interventions.

Of the 17 included articles, 16 [19,47,55-68] were RCTs. However, control conditions differed across the RCTs. Only 1 study directly compared an online intervention with a face-to-face intervention [55], which is a control condition that, based on a typology proposed by Goldberg et al [44], provides high comparison strength. The mentioned study found almost equal effects of the online and in-person interventions [55]. Nevertheless, it is expected, due to the great need for fast solutions to deliver interventions amid the ongoing pandemic and obstacles like quarantine, lockdowns, and increased safety measures, that comparisons with face-to-face-interventions were not possible in most cases. Some of the included studies had wait list or no-treatment control groups [19,56,59,62,65,66,68], which can be considered as a control condition with low comparison strength [44]. Others used different content in the control groups [47,57,58,60,61,63,64], mostly providing medium comparison strength [44], or compared different kinds of interventions [58,67] (high comparison strength [44]). These different kinds of control conditions have to be considered when comparing effect sizes of the included studies, and future research should try to use control conditions that provide high comparison strength. Most of the included studies used a videoconferencing system, although several applied interventions that were developed or adapted especially for online delivery.

Meta-analyses on treatment effectiveness yielded significant effects regarding depressive symptoms and medium effects regarding anxiety and social functioning. The results indicate that online or remote interventions show promising results regarding the aforementioned variables. This is a slightly more favorable result than earlier reviews on online interventions or prevention programs with young people conducted before the pandemic. Earlier results were not entirely conclusive, but there were some promising findings for depressive symptoms [69,70] and anxiety [69]. For adults, a review found digitalized CBT interventions to reduce depressive symptoms in pregnant women [41]. However, specifically in the context of COVID-19, a need for more high-quality research has been identified. A systematic review that included only studies with adults published after the onset of the pandemic found some encouraging results for online interventions targeting anxiety and depression [38]. In another review on web-based exercise interventions for adults, the superiority of the interventions over the control conditions was present in only 1 of 3 studies for depressive symptoms and in none for anxiety symptoms [39].

Concerning other variables, the picture is even more unclear: Interventions to improve well-being and reduce psychological stress, disordered eating, and COVID-19–related psychological symptoms did not show significant effects across studies in this meta-analysis. One must consider, however, that only a few studies for each outcome were eligible for these calculations. Only 3 studies could be included in the analyses regarding well-being. One used a Zoom-based intervention focusing on peer support [65], one used an app to reduce suicidal ideation [64], and the third used a self-administered online positive psychology intervention for different mental health outcomes [68]. Thus, in addition to various main interventions using different kinds of online or remote applications, it can be assumed that the baseline state of well-being or differently expressed amount of suffering was quite different between the 3 studies, which might explain the lack of a significant overall treatment effect.

Psychological stress was analyzed in 3 additional studies [47,64,68]; 2 were included in the meta-analysis of interventions for COVID-19–related outcomes [19,58]. There were also 2 different outcomes used in the latter 2 studies: 1 study [58] examined COVID-19–related trauma, while the other [19] focused on COVID-19–related anxiety. However, the systematic review by Bonardi et al [38] found 3 high-quality studies that were able to decrease different COVID-19–related symptoms like anxiety and depressive symptoms in adults postintervention [71] or at least 6 weeks after the intervention [72,73]; this shows that there seems to be some online or remote interventions available for adult participants.

Regarding disordered eating, the lack of overall treatment effect across the studies could indicate that more than remote therapy is needed for eating disorders and symptoms. Eating disorders might require approaches that treat the somatic aspects in a clinical setting to regularly control for treatment compliance [74]. A previous meta-analysis found similar results, with the lowest effectiveness for online interventions for eating disorders [75]. The study by Zhou and Wade [56] compared symptoms before the onset of COVID-19 and during COVID-19 and found more symptoms during the pandemic, underlining the increased need for interventions due to the pandemic. Although disordered eating and body image flexibility decreased in patients entering the study both before and during the pandemic, the impact of the intervention on self-compassion decreased during the pandemic. All in all, for all the variables showing no overall treatment effects, the few studies available suggest that more research is needed before a clear conclusion regarding the effectiveness of remote or online interventions can be drawn for these symptoms.

Most of the included studies based their online interventions primarily on well-studied therapy forms like CBT [19,47,55,57,59,63] and extensions of CBT or therapy forms related to it like acceptance commitment therapy [54], mindfulness-based interventions [61,66,67], and DBT [47,64]. Interventions based on positive psychology [68] and SFBT [62] were also included in the sample. Thus, interventions were developed from evidence-based forms of therapy. No clear superiority of any form of therapy can be found in this sample of studies. Interventions differed according to length, from 1 session [58] to 3 months [47,59]. Nevertheless, even the single-session intervention was effective [58].

Although some studies show promising results regarding interventions for adolescents [55,58-60,62,65] or young adults [19,47,60,61,63,64,66,68], it is more unclear how effective such interventions are for younger children, as only 2 studies focusing on elementary school children [67] or children up to the age of 12 years [54] could be included. In 1 of these studies [67], a philosophy-based intervention was more effective than the mindfulness-based intervention, possibly hinting at a higher effectiveness of more creative approaches when working with younger children.

Most studies recruited through social media, primary care centers, or at a university. However, in 2 articles, schools were involved: 1 study [62] used school referrals, while the other conducted the intervention in elementary school classes [67], thus showing that, especially with younger children, interventions can also be set within the school context, even if online.

It must be noted that most studies were conducted in North America or China. Although it can be assumed that technical opportunities might be equal in most of Europe, it needs to be clarified how the results can be adapted to lower-income countries, where financial aspects might impede technical opportunities.

The digital transition to online or remotely delivered interventions seems essential, not only considering challenging circumstances like the COVID-19 pandemic, which made face-to-face treatment in many cases impossible, but also in light of the ever-increasing numbers of children, adolescents, and young adults experiencing mental health issues or who have psychological disorders. Thus, it is relevant to develop low-threshold interventions [8,9,14,16]. Nevertheless, several factors must be considered: Legal frameworks might need to be adapted for different countries [8], and therapists might need support when transitioning to online interventions [8,27]. Regarding the development of such interventions, studies have shown positive effects by including peer groups in the development process [34,36] and using peers as advisors [35,37].

However, using digital media and smartphones in and of themselves might pose risk factors for children and adolescents: An increase in cyber victimization through media use has been found [57], and young people are more at risk for addictive behavior in general [7]. Problematic behavior has also been discussed for problematic smartphone use [43], which has been found to impede mental well-being and QoL in children and adolescents [76].

The clinical implications of this meta-analysis are both immediate and far-reaching; the results underscore the versatility and applicability of online therapeutic interventions across diverse settings. As the COVID-19 pandemic amplified the demand for remote interventions, the emergence of promising outcomes, despite rapid development, demonstrates the adaptability and resilience of the mental health sector. These findings suggest that online and hybrid therapeutic modalities not only provide a viable alternative to traditional face-to-face sessions but also bridge the accessibility gap. They offer crucial mental health support to those confronted with problems of accessibility rooted in personal, communicative, geographical, or logistical barriers, as well as challenges stemming from limited mobility due to mental or physical disorders. Such restrictions often make traditional therapeutic settings challenging, underscoring the importance of versatile, remote solutions.

The utility of remote interventions and the promising outcomes of these methodologies have transformative potential for various contexts. In educational environments, for instance, schools can leverage online interventions to address the mental health needs of students who may be reluctant or unable to access traditional counseling services [35-37,77]. Primary care settings can also integrate telehealth solutions into their care regimes, ensuring patients have consistent and comprehensive mental health support alongside their physical health needs. Additionally, psychiatric rehabilitation provides supportive care, but the therapeutic effects often decrease after discharge [78]. Implementing online or hybrid care modalities post-inpatient treatment could potentially bolster and prolong the beneficial outcomes of therapy, offering a more sustained therapeutic impact for patients in the long run [79].

Telehealth and hybrid systems can be transformative in delivering mental health services. A hybrid approach, which blends traditional face-to-face therapy with online sessions, can cater to diverse patient needs and preferences, enhancing treatment adherence, accessibility, and comfort. For example, patients might initiate their therapeutic journey via face-to-face consultations and later transition to online sessions for convenience, or vice versa. Schools can adopt similar hybrid models, offering in-person counseling sessions and providing digital platforms for students to access support during out-of-school hours or remote learning periods. Likewise, primary care facilities can offer a combination of in-person consultations with remote follow-ups, ensuring continuity of care. The potential of these strategies will need detailed scientific investigation.

This adaptability was also evident during crises like the Syrian and Ukrainian wars, where online interventions were effectively used to support individuals suffering from trauma and distress in their home countries or while migrating (eg, [80,81]). Therapists and mental health professionals from other countries could remotely provide much-needed assistance, showcasing the potential of such platforms in transcending international borders [82]. The success of these interventions in various crises—health pandemics or geopolitical conflicts—signals the need to reevaluate conventional therapeutic models. A shift toward hybrid care models that combine digital and in-person strategies could be pivotal, especially in future catastrophic events in which immediate physical intervention is hindered [83]. This could ensure that mental health support remains uninterrupted and universally accessible, regardless of geographical and political boundaries.

The universality and adaptability of online interventions suggest a broader shift in the mental health landscape. As the world becomes increasingly interconnected and digital, there is a pressing need to recalibrate therapeutic models. Embracing telehealth and hybrid systems can ensure that mental health support remains robust and adaptable to patients' needs.

Limitations

Although this meta-analysis and literature review is the first to report the effects of online interventions for children, adolescents, and young adults during the pandemic, some limitations of this systematic review and meta-analysis must be noted. We would like to clarify that only studies with experimental or quasiexperimental designs were included in the systematic review. One of the included studies was not an RCT [54]. However, the results of this study are only reported descriptively, as they were not included in the meta-analysis. As for the type of comparison groups, no restrictions were made due to the novelty of the field, aiming to capture a comprehensive range of experimental approaches. We believe this approach provides a more inclusive representation of the current state of research in this domain. Control groups in most studies do not consist of face-to-face interventions due to the pandemic's nature, potentially affecting the validity of conclusions about the true effectiveness of online interventions. A risk of bias assessment was carried out for all included studies, revealing, in some cases, a need for more details regarding allocation concealment and blinding of participants, personnel, and outcome assessors. Proper randomization was only evident in certain cases (see Tables 3 and 4), possibly affecting the overall quality of the studies included. Above all, the various examined outcomes and the different media or renditions of the online or remote interventions mean that no 2 studies in the sample looked at the exact same intervention. Although understandable in a rapidly evolving field like online interventions, more consistent research on specific interventions is necessary for in-depth meta-analyses and subanalyses. The sample is relatively small, leading to even smaller sample sizes of the outcome groups that were analyzed quantitatively. When the literature search was conducted, several other studies matching the search criteria were registered, but results were unavailable.

Conclusion

All in all, the included studies exhibit promising results regarding the implementation of online or app-based interventions for mental health issues for children, adolescents, and young adults. This is relevant not only in times of crises such as the COVID-19 pandemic or catastrophic events but also given the increasing prevalence rates for psychological disorders in these demographics. The results underscore that the digital landscape allows for more straightforward, accessible engagement with young populations, demonstrating equal effectiveness as traditional therapeutic methods. Still, research on this population is limited so far. Notably, online and app-based interventions provide a compelling alternative to face-to-face therapy, showcasing notable efficacy, particularly in addressing symptoms of anxiety and depression and improving social functioning. The commitment from young individuals toward these interventions seems robust and encouraging. Although there is a pressing need for further high-quality research on various interventions with heightened comparability, it is evident that there are tangible, effective alternatives to in-person therapeutic interventions. Considering these findings, incorporating online intervention techniques should be paramount in the future training of clinical psychologists and psychotherapists to ensure they remain adaptive, effective, and relevant in our ever-evolving digital age.

Acknowledgments

The authors appreciate the contribution of NÖ Landesgesundheitsagentur, legal entity of University Hospitals in Lower Austria, for providing the organizational framework to conduct this research. We also want to acknowledge support from the Open Access Publishing Fund of Karl Landsteiner University of Health Sciences, Krems, Austria.

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) checklist.

[DOCX File , 34 KB - mental_v11i1e46637_app1.docx]

Multimedia Appendix 2

Database search strategy in more detail.

[DOCX File , 17 KB - mental_v11i1e46637_app2.docx]

References

1. Costello EJ, Mustillo S, Erkanli A, Keeler G, Angold A. Prevalence and development of psychiatric disorders in childhood and adolescence. Arch Gen Psychiatry 2003 Aug 01;60(8):837-844. [doi: 10.1001/archpsyc.60.8.837] [Medline: 12912767]
2. Fuchs M, Karwautz A. [The epidemiology of mental disorders in youth : A narrative review including Austrian data]. Neuropsychiatr 2017 Sep 30;31(3):96-102. [doi: 10.1007/s40211-017-0238-x] [Medline: 28853032]
3. Merikangas KR, Nakamura EF, Kessler RC. Epidemiology of mental disorders in children and adolescents. Dialogues in Clinical Neuroscience 2022 Apr 01;11(1):7-20. [doi: 10.31887/dcns.2009.11.1/krmerikangas]
4. Polanczyk G, Salum G, Sugaya L, Caye A, Rohde L. Annual research review: A meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. J Child Psychol Psychiatry 2015 Mar;56(3):345-365. [doi: 10.1111/jcpp.12381] [Medline: 25649325]
5. Wilson S, Hicks BM, Foster KT, McGue M, Iacono WG. Age of onset and course of major depressive disorder: associations with psychosocial functioning outcomes in adulthood. Psychol. Med 2014 Jul 10;45(3):505-514. [doi: 10.1017/s0033291714001640]
6. Dahl RE, Allen NB, Wilbrecht L, Suleiman AB. Importance of investing in adolescence from a developmental science perspective. Nature 2018 Feb 21;554(7693):441-450. [doi: 10.1038/nature25770] [Medline: 29469094]
7. Chambers RA, Taylor JR, Potenza MN. Developmental neurocircuitry of motivation in adolescence: a critical period of addiction vulnerability. Am J Psychiatry 2003 Jun;160(6):1041-1052 [FREE Full text] [doi: 10.1176/appi.ajp.160.6.1041] [Medline: 12777258]
8. Eichenberg C. [Online psychotherapy in times of the corona pandemic]. Psychotherapeut (Berl) 2021;66(3):195-202 [FREE Full text] [doi: 10.1007/s00278-020-00484-0] [Medline: 33462530]
9. Magson NR, Freeman JYA, Rapee RM, Richardson CE, Oar EL, Fardouly J. Risk and protective factors for prospective changes in adolescent mental health during the COVID-19 pandemic. J Youth Adolesc 2021 Jan 27;50(1):44-57 [FREE Full text] [doi: 10.1007/s10964-020-01332-9] [Medline: 33108542]
10. Dale R, Budimir S, Probst T, Stippl P, Pieh C. Mental health during the COVID-19 lockdown over the Christmas period in Austria and the effects of sociodemographic and lifestyle factors. Int J Environ Res Public Health 2021 Apr 01;18(7):3679 [FREE Full text] [doi: 10.3390/ijerph18073679] [Medline: 33916019]
11. Huscsava M, Scharinger C, Plener P, Kothgassner O. 'The world somehow stopped moving': impact of the COVID-19 pandemic on adolescent psychiatric outpatients and the implementation of teletherapy. Child Adolesc Ment Health 2022 Sep;27(3):232-237. [doi: 10.1111/camh.12481] [Medline: 34114730]
12. Özlü-Erkilic Z, Kothgassner OD, Wenzel T, Goreis A, Chen A, Ceri V, et al. Does the progression of the COVID-19 pandemic have an influence on the mental health and well-being of young people? A cross-sectional multicenter study. Int J Environ Res Public Health 2021 Dec 04;18(23):1 [FREE Full text] [doi: 10.3390/ijerph182312795] [Medline: 34886520]
13. Pieh C, Plener PL, Probst T, Dale R, Humer E. Assessment of mental health of high school students during social distancing and remote schooling during the COVID-19 pandemic in Austria. JAMA Netw Open 2021 Jun 01;4(6):e2114866 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.14866] [Medline: 34181016]

14. Pierce M, McManus S, Hope H, Hotopf M, Ford T, Hatch SL, et al. Mental health responses to the COVID-19 pandemic: a latent class trajectory analysis using longitudinal UK data. Lancet Psychiatry 2021 Jul;8(7):610-619 [FREE Full text] [doi: 10.1016/S2215-0366(21)00151-6] [Medline: 33965057]
15. Schlack R, Neuperdt L, Junker S, Eicher S, Hölling H, Thom J, et al. Changes in mental health in the German child and adolescent population during the COVID-19 pandemic - Results of a rapid review. J Health Monit 2023 Feb;8(Suppl 1):2-72 [FREE Full text] [doi: 10.25646/10761] [Medline: 36818693]
16. Panchal U, Salazar de Pablo G, Franco M, Moreno C, Parellada M, Arango C, et al. The impact of COVID-19 lockdown on child and adolescent mental health: systematic review. Eur Child Adolesc Psychiatry 2023 Jul 18;32(7):1151-1177 [FREE Full text] [doi: 10.1007/s00787-021-01856-w] [Medline: 34406494]
17. Loades ME, Chatburn E, Higson-Sweeney N, Reynolds S, Shafran R, Brigden A, et al. Rapid systematic review: the impact of social isolation and loneliness on the mental health of children and adolescents in the context of COVID-19. J Am Acad Child Adolesc Psychiatry 2020 Nov;59(11):1218-1239.e3 [FREE Full text] [doi: 10.1016/j.jaac.2020.05.009] [Medline: 32504808]
18. Lee J. Mental health effects of school closures during COVID-19. Lancet Child Adolesc Health 2020 Jun;4(6):421 [FREE Full text] [doi: 10.1016/S2352-4642(20)30109-7] [Medline: 32302537]
19. Shabahang R, Aruguete MS, McCutcheon L. Video-based cognitive-behavioral intervention for COVID-19 anxiety: a randomized controlled trial. Trends Psychiatry Psychother 2021;43(2):141-150 [FREE Full text] [doi: 10.47626/2237-6089-2020-0056] [Medline: 34043898]
20. Gutmann MT, Aysel M, Özlü-Erkilic Z, Popow C, Akkaya-Kalayci T. Mental health problems of children and adolescents, with and without migration background, living in Vienna, Austria. Child Adolesc Psychiatry Ment Health 2019 Sep 10;13(1):35 [FREE Full text] [doi: 10.1186/s13034-019-0295-y] [Medline: 31528201]
21. Haner D, Pepler D. Clients at Kids Help Phone: individual characteristics and problem topics. J Can Acad Child Adolesc Psychiatry 2016;25(3):138-144 [FREE Full text] [Medline: 27924143]
22. Simpson SG, Reid CL. Therapeutic alliance in videoconferencing psychotherapy: a review. Aust J Rural Health 2014 Dec 10;22(6):280-299. [doi: 10.1111/ajr.12149] [Medline: 25495622]
23. Avenevoli S, Swendsen J, He J, Burstein M, Merikangas KR. Major depression in the national comorbidity survey-adolescent supplement: prevalence, correlates, and treatment. J Am Acad Child Adolesc Psychiatry 2015 Jan;54(1):37-44.e2 [FREE Full text] [doi: 10.1016/j.jaac.2014.10.010] [Medline: 25524788]
24. Forman-Hoffman V, McClure E, McKeeman J, Wood CT, Middleton JC, Skinner AC, et al. Screening for major depressive disorder in children and adolescents: a systematic review for the U.S. Preventive Services Task Force. Ann Intern Med 2016 Feb 09;164(5):342. [doi: 10.7326/m15-2259]
25. Engelhardt E, Engels S. Einführung in die Methoden der Videoberatung. e-beratungsjournal 2021;1(2):9-27 [FREE Full text]
26. Finger-Ossinger M. Internetberatung und Psychotherapie: reflektierte Erfahrungen von Psychotherapeut_innen und Klient_innen. Psychotherapie Forum 2021 Dec 07;25(3-4):115-123. [doi: 10.1007/s00729-021-00186-1]
27. Zubala A, Hackett S. Online art therapy practice and client safety: a UK-wide survey in times of COVID-19. International Journal of Art Therapy 2020 Dec 15;25(4):161-171. [doi: 10.1080/17454832.2020.1845221]
28. Felnhofer A, Fischer-Grote L. [Using new media in pediatric psychosomatic medicine]. Monatsschr Kinderheilkd 2021;169(7):628-632 [FREE Full text] [doi: 10.1007/s00112-021-01184-y] [Medline: 33875897]
29. Schultze NG. Erfolgsfaktoren des virtuellen Settings in der psychologischen Internet-Beratung. e-beratungsjournal 2007;1(5):1-8 [FREE Full text]
30. Backhaus A, Agha Z, Maglione ML, Repp A, Ross B, Zuest D, et al. Videoconferencing psychotherapy: a systematic review. Psychol Serv 2012 May;9(2):111-131. [doi: 10.1037/a0027924] [Medline: 22662727]
31. Pretorius C, Chambers D, Coyle D. Young people's online help-seeking and mental health difficulties: systematic narrative review. J Med Internet Res 2019 Nov 19;21(11):e13873 [FREE Full text] [doi: 10.2196/13873] [Medline: 31742562]
32. Naslund JA, Bondre A, Torous J, Aschbrenner KA. Social media and mental health: benefits, risks, and opportunities for research and practice. J Technol Behav Sci 2020 Sep 20;5(3):245-257 [FREE Full text] [doi: 10.1007/s41347-020-00134-x] [Medline: 33415185]
33. Grist R, Porter J, Stallard P. Mental health mobile apps for preadolescents and adolescents: a systematic review. J Med Internet Res 2017 May 25;19(5):e176 [FREE Full text] [doi: 10.2196/jmir.7332] [Medline: 28546138]
34. Thabrew H, Fleming T, Hetrick S, Merry S. Co-design of eHealth interventions with children and young people. Front Psychiatry 2018 Oct 18;9:481 [FREE Full text] [doi: 10.3389/fpsyt.2018.00481] [Medline: 30405450]
35. Jesser A, Schmalwieser S, Mädge A, Culen C, Schrank B, Böckle M. Chancen niederschwelliger online Peer-to-Peer Begleitung für Jugendliche am Beispiel der Online-Plattform OPEN. Psychotherapie Forum 2021 Dec 13;25(3-4):154-160. [doi: 10.1007/s00729-021-00188-z]
36. Mittmann G, Schmalwieser SS, Diendorfer T, Schrank B, Boeckle M. Peer facilitators as core co-developers of an online peer encouragement network (OPEN2chat) for Austrian adolescents. Front Digit Health 2022 Jun 16;4:833006 [FREE Full text] [doi: 10.3389/fdgth.2022.833006] [Medline: 35783598]

37. Schmalwieser S, Jesser A, Mädge AL, Schrank B, Böckle M. open2chat – eine digitale Peer-to-Peer-Begleitung für Jugendliche. Theoretische Zugänge und praktische Lösungen. e-beratungsjournal 2022;18(1):66-77. [doi: 10.48341/8pjb-9S38]
38. Bonardi O, Wang Y, Li K, Jiang X, Krishnan A, He C, et al. Effects of COVID-19 mental health interventions among children, adolescents, and adults not quarantined or undergoing treatment due to COVID-19 infection: a systematic review of randomised controlled trials. Can J Psychiatry 2022 May;67(5):336-350 [FREE Full text] [doi: 10.1177/07067437211070648] [Medline: 35275494]
39. Carneiro L, Rosenbaum S, Ward PB, Clemente FM, Ramirez-Campillo R, Monteiro-Júnior RS, et al. Web-based exercise interventions for patients with depressive and anxiety disorders: a systematic review of randomized controlled trials. Braz J Psychiatry 2022;44(3):331-341 [FREE Full text] [doi: 10.1590/1516-4446-2021-2026] [Medline: 34852034]
40. Valentine AZ, Hall SS, Young E, Brown BJ, Groom MJ, Hollis C, et al. Implementation of telehealth services to assess, monitor, and treat neurodevelopmental disorders: systematic review. J Med Internet Res 2021 Jan 20;23(1):e22619 [FREE Full text] [doi: 10.2196/22619] [Medline: 33326409]
41. Wan Mohd Yunus WMA, Matinolli H, Waris O, Upadhyaya S, Vuori M, Korpilahti-Leino T, et al. Digitalized cognitive behavioral interventions for depressive symptoms during pregnancy: systematic review. J Med Internet Res 2022 Feb 23;24(2):e33337 [FREE Full text] [doi: 10.2196/33337] [Medline: 35195532]
42. Colasante T, Lin L, De France K, Hollenstein T. Any time and place? Digital emotional support for digital natives. Am Psychol 2022;77(2):186-195. [doi: 10.1037/amp0000708] [Medline: 32969678]
43. Cha S, Seo B. Smartphone use and smartphone addiction in middle school students in Korea: Prevalence, social networking service, and game use. Health Psychol Open 2018;5(1):2055102918755046 [FREE Full text] [doi: 10.1177/2055102918755046] [Medline: 29435355]
44. Goldberg SB, Sun S, Carlbring P, Torous J. Selecting and describing control conditions in mobile health randomized controlled trials: a proposed typology. NPJ Digit Med 2023 Sep 30;6(1):181 [FREE Full text] [doi: 10.1038/s41746-023-00923-7] [Medline: 37775522]
45. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009 Aug 18;151(4):264-9, W64 [FREE Full text] [doi: 10.7326/0003-4819-151-4-200908180-00135] [Medline: 19622511]
46. Seiffge-Krenke I. Emerging adulthood. Psychotherapie 2023 Jan 09;68(1):67-76. [doi: 10.1007/s00278-022-00638-2]
47. Suffoletto B, Goldstein T, Gotkiewicz D, Gotkiewicz E, George B, Brent D. Acceptability, engagement, and effects of a mobile digital intervention to support mental health for young adults transitioning to college: pilot randomized controlled trial. JMIR Form Res 2021 Oct 14;5(10):e32271 [FREE Full text] [doi: 10.2196/32271] [Medline: 34647893]
48. Viechtbauer W. Conducting meta-analyses in R with the metafor package. Journal of Statistical Software 2010;36(3):1-48. [doi: 10.18637/jss.v036.i03]
49. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 06;327(7414):557-560 [FREE Full text] [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
50. Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale, N.J: L. Erlbaum Associates; 1988.
51. Faraone SV. Interpreting estimates of treatment effects: implications for managed care. P T 2008 Dec;33(12):700-711 [FREE Full text] [Medline: 19750051]
52. Tufanaru C, Munn Z, Aromataris E, Campbell J, Hopp L. Chapter 3: Systematic reviews of effectiveness. In: Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis. Adelaide, South Australia: Joanna Briggs Institute; 2020.
53. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing URL: https://www.R-project.org [accessed 2023-01-15]
54. Duan W, Kong Y, Bu H, Guan Q, Chen Z, Luo Q, et al. The online strength-informed acceptance and commitment therapy among COVID-19-affected adolescents. Research on Social Work Practice 2022 Feb 11;32(4):465-474. [doi: 10.1177/10497315211067270]
55. Prato A, Maugeri N, Chiarotti F, Morcaldi L, Vicario CM, Barone R, et al. A randomized controlled trial comparing videoconference vs. face-to-face delivery of behavior therapy for youths with Tourette syndrome in the time of COVID-19. Front Psychiatry 2022;13:862422 [FREE Full text] [doi: 10.3389/fpsyt.2022.862422] [Medline: 35686189]
56. Zhou Y, Wade TD. The impact of COVID-19 on body-dissatisfied female university students. Int J Eat Disord 2021 Jul 13;54(7):1283-1288 [FREE Full text] [doi: 10.1002/eat.23521] [Medline: 33851442]
57. Kutok ER, Dunsiger S, Patena JV, Nugent NR, Riese A, Rosen RK, et al. A cyberbullying media-based prevention intervention for adolescents on Instagram: pilot randomized controlled trial. JMIR Ment Health 2021 Sep 15;8(9):e26029 [FREE Full text] [doi: 10.2196/26029] [Medline: 34524103]
58. Schleider JL, Mullarkey MC, Fox KR, Dobias ML, Shroff A, Hart EA, et al. A randomized trial of online single-session interventions for adolescent depression during COVID-19. Nat Hum Behav 2022 Feb 09;6(2):258-268 [FREE Full text] [doi: 10.1038/s41562-021-01235-0] [Medline: 34887544]
59. Nicol G, Wang R, Graham S, Dodd S, Garbutt J. Chatbot-delivered cognitive behavioral therapy in adolescents with depression and anxiety during the COVID-19 pandemic: feasibility and acceptability study. JMIR Form Res 2022 Nov 22;6(11):e40242 [FREE Full text] [doi: 10.2196/40242] [Medline: 36413390]

60. Yeager DS, Bryan CJ, Gross JJ, Murray JS, Krettek Cobb D, H F Santos P, et al. A synergistic mindsets intervention protects adolescents from stress. Nature 2022 Jul 06;607(7919):512-520 [FREE Full text] [doi: 10.1038/s41586-022-04907-7] [Medline: 35794485]
61. Sun S, Lin D, Goldberg S, Shen Z, Chen P, Qiao S, et al. A mindfulness-based mobile health (mHealth) intervention among psychologically distressed university students in quarantine during the COVID-19 pandemic: A randomized controlled trial. J Couns Psychol 2022 Mar;69(2):157-171 [FREE Full text] [doi: 10.1037/cou0000568] [Medline: 34264696]
62. Chen S, Zhang Y, Qu D, He J, Yuan Q, Wang Y, et al. An online solution focused brief therapy for adolescent anxiety : A randomized controlled trial. Asian J Psychiatr 2023 Aug;86:103660 [FREE Full text] [doi: 10.1016/j.ajp.2023.103660] [Medline: 37307701]
63. He Y, Yang L, Zhu X, Wu B, Zhang S, Qian C, et al. Mental health chatbot for young adults with depressive symptoms during the COVID-19 pandemic: single-blind, three-arm randomized controlled trial. J Med Internet Res 2022 Nov 21;24(11):e40719 [FREE Full text] [doi: 10.2196/40719] [Medline: 36355633]
64. Torok M, Han J, McGillivray L, Wong Q, Werner-Seidler A, O'Dea B, et al. The effect of a therapeutic smartphone application on suicidal ideation in young adults: Findings from a randomized controlled trial in Australia. PLoS Med 2022 May 31;19(5):e1003978 [FREE Full text] [doi: 10.1371/journal.pmed.1003978] [Medline: 35639672]
65. Pavarini G, Reardon T, Hollowell A, Bennett V, Lawrance E, Peer Support Young People’s Advisory Group, et al. Online peer support training to promote adolescents' emotional support skills, mental health and agency during COVID-19: Randomised controlled trial and qualitative evaluation. Eur Child Adolesc Psychiatry 2023 Jun 17;32(6):1119-1130 [FREE Full text] [doi: 10.1007/s00787-021-01933-0] [Medline: 35174419]
66. Simonsson O, Bazin O, Fisher SD, Goldberg SB. Effects of an eight-week, online mindfulness program on anxiety and depression in university students during COVID-19: A randomized controlled trial. Psychiatry Res 2021 Nov;305:114222 [FREE Full text] [doi: 10.1016/j.psychres.2021.114222] [Medline: 34601450]
67. Malboeuf-Hurtubise C, Léger-Goodes T, Mageau GA, Joussemet M, Herba C, Chadi N, et al. Philosophy for children and mindfulness during COVID-19: Results from a randomized cluster trial and impact on mental health in elementary school students. Prog Neuropsychopharmacol Biol Psychiatry 2021 Apr 20;107:110260 [FREE Full text] [doi: 10.1016/j.pnpbp.2021.110260] [Medline: 33493652]
68. Krifa I, Hallez Q, van Zyl LE, Braham A, Sahli J, Ben Nasr S, et al. Effectiveness of an online positive psychology intervention among Tunisian healthcare students on mental health and study engagement during the Covid-19 pandemic. Appl Psychol Health Well Being 2022 Nov;14(4):1228-1254. [doi: 10.1111/aphw.12332] [Medline: 34939332]
69. Clarke AM, Kuosmanen T, Barry MM. A systematic review of online youth mental health promotion and prevention interventions. J Youth Adolesc 2015 Jan;44(1):90-113. [doi: 10.1007/s10964-014-0165-0] [Medline: 25115460]
70. Rice SM, Goodall J, Hetrick SE, Parker AG, Gilbertson T, Amminger GP, et al. Online and social networking interventions for the treatment of depression in young people: a systematic review. J Med Internet Res 2014 Sep 16;16(9):e206 [FREE Full text] [doi: 10.2196/jmir.3304] [Medline: 25226790]
71. Wahlund T, Mataix-Cols D, Olofsdotter Lauri K, de Schipper E, Ljótsson B, Aspvall K, et al. Brief online cognitive behavioural intervention for dysfunctional worry related to the COVID-19 pandemic: a randomised controlled trial. Psychother Psychosom 2021;90(3):191-199 [FREE Full text] [doi: 10.1159/000512843] [Medline: 33212440]
72. Kahlon MK, Aksan N, Aubrey R, Clark N, Cowley-Morillo M, Jacobs EA, et al. Effect of layperson-delivered, empathy-focused program of telephone calls on loneliness, depression, and anxiety among adults during the COVID-19 pandemic: a randomized clinical trial. JAMA Psychiatry 2021 Jun 01;78(6):616-622 [FREE Full text] [doi: 10.1001/jamapsychiatry.2021.0113] [Medline: 33620417]
73. Thombs BD, Kwakkenbos L, Levis B, Bourgeault A, Henry RS, Levis AW, et al. Effects of a multi-faceted education and support programme on anxiety symptoms among people with systemic sclerosis and anxiety during COVID-19 (SPIN-CHAT): a two-arm parallel, partially nested, randomised, controlled trial. The Lancet Rheumatology 2021 Jun;3(6):e427-e437. [doi: 10.1016/s2665-9913(21)00060-6]
74. Walsh BT, Fairburn CG, Mickley D, Sysko R, Parides MK. Treatment of bulimia nervosa in a primary care setting. Am J Psychiatry 2004 Mar;161(3):556-561. [doi: 10.1176/appi.ajp.161.3.556] [Medline: 14992983]
75. Barak A, Hen L, Boniel-Nissim M, Shapira N. A comprehensive review and a meta-analysis of the effectiveness of internet-based psychotherapeutic interventions. Journal of Technology in Human Services 2008 Jul 03;26(2-4):109-160. [doi: 10.1080/15228830802094429]
76. Fischer-Grote L, Kothgassner OD, Felnhofer A. The impact of problematic smartphone use on children's and adolescents' quality of life: A systematic review. Acta Paediatr 2021 May 22;110(5):1417-1424 [FREE Full text] [doi: 10.1111/apa.15714] [Medline: 33305437]
77. Zechner KA, Wiesner C, Böckle M. Who cares? #schule verantworten 2022 Dec 23;2(4):105-114. [doi: 10.53349/sv.2022.i4.a278]
78. Owusu E, Oluwasina F, Nkire N, Lawal MA, Agyapong VIO. Readmission of patients to acute psychiatric hospitals: influential factors and interventions to reduce psychiatric readmission rates. Healthcare (Basel) 2022 Sep 19;10(9):1808 [FREE Full text] [doi: 10.3390/healthcare10091808] [Medline: 36141418]

79. Eboreime E, Shalaby R, Mao W, Owusu E, Vuong W, Surood S, et al. Reducing readmission rates for individuals discharged from acute psychiatric care in Alberta using peer and text message support: Protocol for an innovative supportive program. BMC Health Serv Res 2022 Mar 12;22(1):332 [FREE Full text] [doi: 10.1186/s12913-022-07510-8] [Medline: 35279142]
80. Almoshmosh N, Jefee-Bahloul H, Abdallah W, Barkil-Oteo A. Use of store-and-forward tele-mental health for displaced Syrians. Intervention 2020;18(1):66-70. [doi: 10.4103/INTV.INTV_53_19]
81. Gordon D. Amid Russian invasion, new initiative launches free telemedicine for Ukrainians under siege. Forbes. 2022 Mar 21. URL: https://www.forbes.com/sites/debgordon/2022/03/21/amid-russian-invasion-new-initiative-launches-free-telemedicine-for-ukrainians-under-siege/?sh=2817ce964470 [accessed 2023-01-15]
82. Lurie N, Carr BG. The role of telehealth in the medical response to disasters. JAMA Intern Med 2018 Jun 01;178(6):745-746. [doi: 10.1001/jamainternmed.2018.1314] [Medline: 29710200]
83. Rosen A, Gill N, Salvador-Carulla L. The future of community psychiatry and community mental health services. Curr Opin Psychiatry 2020 Jul;33(4):375-390. [doi: 10.1097/YCO.0000000000000620] [Medline: 32452944]

Abbreviations

CBT: cognitive behavioral therapy
DBT: dialectical behavior therapy
IMPACT: Intervention Media to Prevent Adolescent Cyber-Conflict Through Technology
MoST-MH: Mobile Support Tool for Mental Health
P4C: philosophy for children
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis
QoL: quality of life
RCT: randomized controlled trial
SACT: Strength-informed Acceptance and Commitment Therapy
SFBT: solution-focused brief therapy
SMD: standardized mean difference

Edited by J Torous; submitted 20.02.23; peer-reviewed by M Subotic-Kerry, C Leung; comments to author 16.06.23; revised version received 21.09.23; accepted 21.10.23; published 05.02.24.

Please cite as:
Fischer-Grote L, Fössing V, Aigner M, Fehrmann E, Boeckle M
Effectiveness of Online and Remote Interventions for Mental Health in Children, Adolescents, and Young Adults After the Onset of the COVID-19 Pandemic: Systematic Review and Meta-Analysis
JMIR Ment Health 2024;11:e46637
URL: https://mental.jmir.org/2024/1/e46637
doi:10.2196/46637
PMID:38315524

Review

Effectiveness of Virtual Reality–Based Well-Being Interventions for Stress Reduction in Young Adults: Systematic Review

Joy Xu[1]; Areej Khanotia[2]; Shmuel Juni[3]; Josephine Ku[4]; Hana Sami[4]; Vallen Lin[4]; Roberta Walterson[2]; Evelyn Payne[3]; Helen Jo[3]; Parmin Rahimpoor-Marnani[5]

[1]David Geffen School of Medicine at UCLA, Los Angeles, CA, United States
[2]Faculty of Science, University of Ottawa, Ottawa, ON, Canada
[3]Faculty of Arts and Science, University of Toronto, Toronto, ON, Canada
[4]Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada
[5]Faculty of Science, York University, Toronto, ON, Canada

Corresponding Author:
Joy Xu
David Geffen School of Medicine at UCLA
417 Charles E Young Drive West
Los Angeles, CA, 90024
United States
Phone: 1 3109481476
Email: joyjxu@gmail.com

Abstract

Background: Adolescents can be especially vulnerable to various stressors as they are still in their formative years and transitioning into adulthood. Hence, it is important for them to have effective stress management strategies.

Objective: This systematic review investigates current well-being interventions that are aimed at reducing stress among young adults. In particular, interventions using the medium of virtual reality (VR) are explored.

Methods: This mixed methods systematic review follows the PRISMA-P (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Protocols) guidelines, and papers were gathered from databases such as PsycINFO, PubMed, Science Direct, Web of Science, OpenGrey, and Edutopia. Predetermined criteria and specific keywords were used to search for the papers. Search results were screened and extracted with all article screening or extraction delegated among all authors. Any disagreements after reconciliation were settled by a third author. The quality and risk of bias of included studies were assessed using the GRADE (Grading of Recommendations Assessment, Development, and Evaluation) Tool for Quantitative Studies. Studies were analyzed qualitatively.

Results: In total, 20 studies were included, and qualitative analysis was performed to evaluate the effectiveness of VR-based interventions in 3 domains: nature, stress, and academics.

Conclusions: Studies using VR interventions, overall, promoted a reduction in stress and an increase in well-being. The findings suggest that VR may serve as an accessible and affordable medium of stress reduction for students and young adults. Larger sample sizes, and a greater number of included studies, may be required in future directions.

(JMIR Ment Health 2024;11:e52186) doi:10.2196/52186

KEYWORDS

well-being; well-being; virtual reality; VR; stress, nature; academic; student; intervention; young adults; teens; adolescent; stressors; stress management; systematic review; accessible; accessibility; students; affordable

Introduction

The COVID-19 pandemic has impacted millions of lives across the world and provided a spotlight on systemic disparities including present limitations to well-being and accessibility for support. The reverberations of this global phenomenon have reconciled greater focus on adolescents and young adults in particular, as their formative years were inevitably impacted by social distancing. The crucial years of their education were faced with obstacles in course delivery, academic opportunities, and social spheres. Furthermore, stresses were compounded with other hardships such as economic setbacks, limited socialization, and more issues that cumulatively burden one's well-being, especially in the transition into adulthood. Thus, there has been significant effort to identify potential targets for interventions through research. Broadly, interventions aim to help study some change in individual experiences through strategies and processes after a systematic modification [1]. The main objective is to measure the effect of a process or program on certain situations [1]. In this review, virtual reality (VR) interventions refer to programs or treatments that target one or more determinants of health using a VR headset, which displays a visual environment. This review refers to the Canadian Index's definition of Well-being stated as "The presence of the highest possible quality of life in its full breadth of expression focused on but not necessarily exclusive to good living standards, robust health, a sustainable environment, vital communities, an educated populace, balanced time use, …" [2]. Thus, some aspects of an individual's well-being can be measured through the extent to which an individual is stressed. Stress can be understood as a "psychophysiological response" to some form of danger and involves biological components including nervous and hormonal responses to stimuli [3].

This systematic review questions the effectiveness of VR interventions in reducing stress and promoting well-being in students and young adults. Currently, mindfulness-based interventions (MBIs) through digital or computerized means are common, such as mindfulness-based stress reduction, group mindfulness-based intervention, and self-direct mindfulness-based intervention through digital delivery. There is evidence that these interventions have significant improvement in regulating emotion and mindfulness [4]. Similarly, there has been progressive growth in present research demonstrating the efficacy of VR interventions for well-being. VR can be defined as an artificial, 3D digital environment that a user experiences through a computer headset [5]. Although many reviews in the literature summarize the efficacy or need for MBIs delivered on a computer or digital program, there remains a need for a comprehensive review to specifically assess the impact of VR-based interventions on stress and mental well-being among students and young adults.

Currently, the use of nature-based settings alone was found to improve mood to being "good" and "calm" in older adults, without the need for an MBI curriculum [6]. Thus, the papers included in this systematic review use auditory, visual, and even olfactory aids to simulate an MBI to reduce stress and increase mindfulness in young adults, such as integrating MBI techniques within VR or similar tools or resources.

This systematic review aims to understand the effectiveness of current well-being VR interventions in reducing stress among young adults. Due to the nature of learning in the 21st century relying heavily on digital resources, implementing digital tactics is imperative to combating these new sources of stress. In an era where Generation Z (the generation born between 1997 and 2012) is more stressed than previous generations and most are experiencing burnout [7], having digital resources allows for immediate and low-maintenance aid.

Methods

A diverse range of literature was assessed according to the inclusion and exclusion criteria, which were determined prior to the search. These criteria qualified the contextual and scientific relevance of the data by developing a standardized expectation for the literature's content and experimental purpose. The PRISMA-P (Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols; Multimedia Appendix 1 [8]) guideline was used for this qualitative review. There was an inclusion of gray literature in accordance with Assessing the Methodological Quality of Systematic Reviews-2 (AMSTAR-2), as this body of research encompasses material produced outside of traditional commercial or academic publications to reduce publication bias (eg, government databases and preprints). Papers published between 1980 and 2022, which were randomized controlled trials, were gathered from PsycINFO, PubMed, Science Direct, Web of Science, Open Grey, and Edutopia through an initial screening. A specific keyword search was used to gather these papers including the following keywords from each of the search platforms ("wellbeing" OR "well-being" AND "student" AND "virtual" AND "reality" AND "quantitative" AND "intervention"). Duplicates and papers not matching the screening questionnaire were eliminated.

This review considers young adults to be within the age range of 15-40 years. All studies included are published in English. Included studies focused on the analysis of diverse student demographics in the scope of university programs, socioeconomic status, and culture for a robust analysis of student well-being. The interventions under study in this paper include any VR-based environments that can be accessed by a participant through a VR headset, where concepts can be incorporated into a mobile app. Furthermore, stress reduction interventions that use either physiological or psychological questionnaires were included. Exclusion criteria included any research that generated qualitative data and studies published by organizations with conflicts of interest. Studies with interventions geared toward individuals with psychological illnesses and preexisting psychological illnesses were also excluded to reduce confounding factors and external influences or experiences that may uniquely affect the study outcomes in these demographics.

Data for this study were collected and screened using the Covidence software (Veritas Health Innovation). Multiple authors (JX, AK, SJ, JK, HS, VL, RW, EP, HJ, and PRM) screened a collection of research for relevant studies in 2 steps: abstract screening and full-text screening. From a dropdown menu, authors selected a reason for excluding an article (ie, wrong outcomes: Where the paper reports on findings not in line with the research question of systematic reviews, such as pilot studies, no quantitative data, and biological markers of stress). Two authors were assigned to each study. The authors worked independently and reconciled any disagreements and discrepancies after each step. Reconciliation proceeded by a third author who reevaluated the inclusion and exclusion criteria of the study based on the screening or extraction tools independently of the 2 other authors. Following screening, the authors extracted relevant information to be used within this review, which included methods of study design, interventions, subject information, and outcomes. EndNote (McMaster University) was used to combine all studies found and to remove any duplicates. GRADE (Grading of Recommendations Assessment, Development, and Evaluation) was used to assess the quality of the studies. GRADE consists of 4 levels of quality: very low, low, moderate, and high. The 2 authors independently evaluated the included studies using the GRADE methodology. When discrepancies arose regarding the assessed quality levels, the said authors collaborated to resolve them, ultimately reaching a consensus on the final quality rating for those studies.

In this case, all studies that did not appropriately follow the inclusion criteria were marked as wrong outcomes, study interventions, and so forth.

Results

The study selection and screening process is outlined in Figure 1. In total, 401 papers were identified from the database search, out of which 20 were included studies. Among these included studies, 11 were randomized controlled trials and 4 were experimental designs. Both the published and gray literature papers went through abstract screening and full-text screening. After the abstract screening, 60 papers remained in total. After full-text screening, 20 papers were left. Through full-text screening, 20 papers remained with categories of nature, stress, and academic contexts within the scope of well-being within the study designs. Studies primarily focused on undergraduate students with included studies encompassing Canada, Ireland, the United States, China, Italy, Europe, Spain, and Russia. Table 1 shows an overview of diverse studies examining the impact of VR environments, with a focus on the nature setting. Key details include authors, publication year, methods, participant populations, interventions, and main findings related to positive and negative affect schedule scores. The studies demonstrate the varied effects of nature-based VR on well-being across different populations. Table 2 presents a synthesis of various studies investigating the effectiveness of VR interventions on stress management and well-being. The studies, conducted using randomized controlled trials and between-subject designs, explore the impact of VR interventions on stress reduction, emotional regulation, relaxation, and positive affect across diverse participant populations. Table 3 provides a comprehensive overview of studies interventions for stress management and well-being among student populations. It outlines various research methodologies, participant demographics, intervention types, and significant findings, offering insights into the use of VR within educational contexts to promote mental health.

Figure 1. Structural diagram presenting screening and included studies.

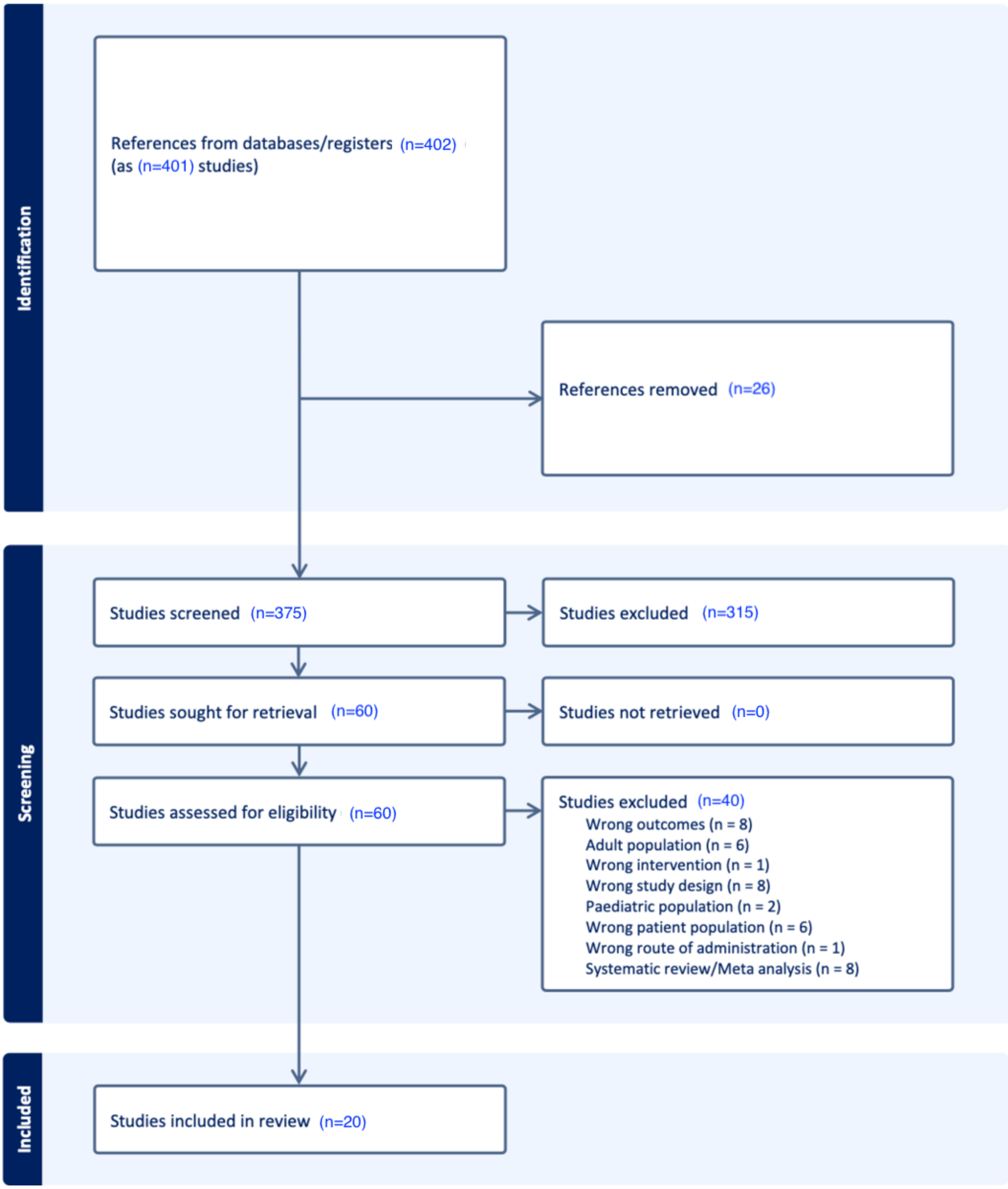


Table 1. Summary of VR[a] environment studies (nature).

Authors and year of publication	Methods	Population	Interventions	Main findings
Valtchanov and Ellard (2010) [9]	Randomized controlled trial	• Undergraduate students aged 18-26 years from the University of Waterloo, Canada • Participants who can read or write English fluently, do not experience "seizures, vertigo, or motion sickness" or severe visual impairment	3 VR settings: • Nature (tropical scenery) • Geometric (assorted 3-dimensional shapes) • Urban (to-scale model of Shibuya station, Tokyo, Japan)	Mean positive affect scores in terms of ZIPERS scores were only significant in the Nature group post immersion in VR, with a mean of 2.21 (SD 0.71) prior to VR compared to 3.03 (SD 0.98) postintervention (*P*<.001). Both geometric and urban VR settings had no significant impact on improving positive effects
O'Meara et al (2020) [10]	Randomized controlled trial	• 18 years old or older • Experience high test anxiety and no treatment • Control: low or healthy level of test anxiety • From University College Cork, Ireland	2 VR settings: • Urban environment • Nature environment	Only the high anxiety group of students significantly benefited from the nature VR intervention within the 4-minute session (*P*=.02). The authors conclude that simulated exposure to nature can therefore reduce negative affect, thereby reducing test anxiety. Conversely, virtual exposure to nature settings did not significantly improve test scores
Browning et al (2020) [11]	Randomized controlled trial	• 18-27 years old, mean age 20 years • University students, from the University of Illinois at Urbana-Champaign, the United States • Excluded if: Diagnosed mood disorder, treated for mental illnesses, hearing impairments, use of alcohol or prescription drugs not normally taken in 24 hours. Intense physical activity in 24 hours	• Control: sit in front of a blank wall • Outdoor group: Forest • VR group: same setting as forest group	Both VR and outdoor interventions have positive and significant results (*P*<.001), indicating high positive effect scores. All 3 conditions reveal a significant decrease in negative affect values (outdoor group *P*=.034; control group *P*<.001, VR *P*=.03). Only the outdoor group reported a statistically significant increase in positive effect for mood effects (*P*=.044)
Gao et al (2019) [12]	Randomized controlled trial	• 120 Chinese college students with mean age of 20.7 years • Myopia degrees <800	6 different VR environment interventions: • Gray • Blue • Open green • Partly open green • Partly closed green • Closed green	Partly open green had the most impact on reducing negative mood, while closed green had the least. Additionally, student preference for environments revealed that blue was most preferred, with gray and closed green as least preferred. A strong positive correlation is revealed between preference for a given environment and positive mood improvement

[a]VR: virtual reality.

Table 2. Overview of studies on VR[a] interventions for stress management and well-being.

Authors and year of publication	Methods	Population	Interventions	Main findings
Villani and Riva (2012) [13]	Between-subject design	36 Italian participants were seen to be on or past the higher quartile on stress	• 6-session stress management intervention • VR experimental: used “ESCAPE” VR • Video experimental group • Audio experimental group	Participants were able to reduce their heart rate across conditions and with time. This was specifically seen in the VR condition, where they were more able to reduce their heart rate which in turn helped their emotional state
Wayment et al (2015) [14]	Randomized controlled trial	32 female first-year undergraduate students	• Audio recording describing the 4 characteristics of the quiet ego • Audio recording with VR of a park scene • Control group: read nature magazines	The VR with the audio Quiet Ego Contemplation reduced the degree to which participants felt “in the moment” as compared to the audio-only group (P<.05)
Villani and Riva (2008) [15]	Randomized controlled trial	36 individuals who were students or office workers from Milan	• Nature VR (park, waterfall, river, garden, etc) • Nature video (park, waterfall, river, garden, etc) • Relaxing audio	VR, video, and audio of nature settings can help induce relaxation and help in stress management, increase positive emotions, enhance self-awareness, and contribute to emotional regulation. Of the 3 interventions, the VR intervention was seen to have the greatest psychological and physiological effects
Villani et al (2007) [16]	Randomized controlled trial	34 female students and 30 male students between the ages of 21 and 28 years from the Catholic University of Milan	• VR with video of tropical islands paired with audio of therapeutic sounds • DVD of relaxing tropical videos • Audio of therapeutic narrative	VR of a relaxing nature environment can enhance the position
Taneja et al (2017) [17]		• Participants who scored 14 or less on the DASS-21[b] test were deemed normal and eligible for the control group • Participants who scored 14 or more were eligible for the stress group	• SCWT[c] task • VR-based stress therapy intervention	PANAS[d] questionnaire: a mean score of 32 for positive affect and a mean score of 15 for negative affect were generated

[a]VR: virtual reality.

[b]DASS-21: Depression Anxiety Stress Scales-21 items.

[c]SCWT: Stroop Color and Word Test.

[d]PANAS: positive and negative affect schedule.

Table 3. Studies on VR[a] interventions for stress management and well-being.

Authors and year of publication	Methods	Population	Interventions	Main findings
Kaplan-Rakowski et al (2021) [18]	Between-subjects randomized design	European-based, university, business students, enrolled in an introductory computer science course, willing to participate in meditation activities	• VR intervention group which used a VR headset for mediation • Video control group which used a monitor for meditation	The mean differences in test scores between examinations before and after intervention were 0.03 for the VR group (SD 0.27; *P*=.01) and –0.19 for the control group (SD 0.42). VR was more effective
Hernández-Ortega et al (2021) [19]	An experimental, analytical, longitudinal, and prospective study	Participants were included if they were enrolled in Practicum I in the 2017-2018 academic year. Additionally, if they were being treated for anxiety with medication then they were excluded	• Cognitive behavioral therapy (digital) • Progressive muscle relaxation (digital)	Participants in IG1[b] showed lower overall scores than those in the control group, which were also statistically significant for the KEZKAK questionnaire (*P*=.019) and state-trait anxiety (*P*=.004). Additionally, no statistical significance was found between IG1 and IG2[c] regarding stress and anxiety
Chen et al (2012) [20]	A randomized controlled trial	• Nursing students from a Spanish public university conducted clinical practices between 2017 and 2019 • Inclusion criteria required enrollment in Practicum I during the 2017-2018 academic year, resulting in 59 participants. Three groups were formed: a control group (n=29), IG1 (n=15) receiving phase I, and IG2 (n=15) undergoing both phases. Students chose control or intervention groups, while random assignment determined IG1 and IG2	Digitally guided meditation sessions	SAS[d] scores showed that postintervention participants reported lower rates of anxiety as compared to preintervention and the control group. On the other hand, SDS[e] scores did not show a great difference between preintervention, postintervention, and the control group
Küchler et al (2020) [21]	Three-armed randomized controlled trial	College students with moderate to low mindfulness levels from universities in Germany, Austria, and Switzerland were recruited. Eligibility criteria include age 18 years or above, enrollment in university or college, proficiency in German, internet access, and moderate to low mindfulness (Freiburg Mindfulness Inventory FMI ≤ 37), with exclusion for ongoing psychotherapy or mindfulness interventions.	1. Studicare meditation—guided and unguided	There are no reported results at this time
Modrego-Alarcon et al (2021) [22]	Three-armed randomized controlled trial	• 18 years of age or older • Enrollment in undergraduate or master's degree course • Enrolled in social sciences or health sciences at the Zaragoza campus or the Calatayud campus • Able to speak and write in Spanish	• Mindfulness-based program that consisted of 90-minute group sessions of mindfulness training • Mindfulness-based program paired with short virtual reality sessions • Relaxation control group	The results showed higher mean stress levels in the control group (17.73, SD 4.42) when compared to both the mindfulness-based program (15.33, SD 4.50), and the mindfulness-based program along with virtual reality (15.75, SD 4.51). VR on its own was unable to prove to be an effective intervention against stress

Authors and year of publication	Methods	Population	Interventions	Main findings
Berezina et al (2022) [23]	Randomized controlled trial	• 62 women and 24 men. Graduate students from Moscow State University of Psychology and Education. Ages 22-53 years • No explicit exclusion criteria	• Stimulating VR scene, followed by relaxing VR scene • Relaxing VR scene, followed by stimulating VR scene • Control group (no interventions)	Stimulating VR scene, then relaxing VR scene: Mean • Stress (female: –1.0; *P*=.00196; male: –0.333333; *P*=.51599) Relaxing VR scene, then stimulating VR scene: Mean • Stress (female: 1.900000; *P*=.00988; male: –0.500000; *P*=.58333) Control: Mean (no reported P values) • Stress (female: –0.9655; male: 0.500000)

[a]VR: virtual reality.

[b]IG1: intervention group 1.

[c]IG2: intervention group 2.

[d]SAS: Self-Rating Anxiety Scale.

[e]SDS: Self-Rating Depression Scale.

Discussion

Measures of Well-Being Through Exposure to Nature

From the included studies that underwent data extraction, a considerable number explored the role of natural environments in a VR setting on well-being. It is well known that exposure to nature can positively improve well-being, such as daily walks in urban parks, hikes, or gardening [24]. In research led by White et al [25], it was revealed that dedicating 2 hours weekly to nature activities correlated with significantly better health and well-being outcomes compared to individuals with no nature exposure. Specifically, those who spent this time in nature had a 20% higher chance of reporting good health and a 33% higher chance of experiencing high well-being. However, access to stimulating biodiversity may not be possible for every individual, such as students in highly urbanized campuses, disabled individuals, and students who live in climates with long winter seasons. Therefore, VR may serve as an accommodation. Browning et al [11] reveal that 6 minutes of VR in an outdoor nature setting results in high positive affect levels postintervention (*P*<.001). The study emphasizes the importance of 360-degree nature videos that are immersive and interactive for adequate similarity to the natural world. O'Meara et al [10] revealed that students with high anxiety can benefit from VR nature exposure by significantly reduced negative affect (*P*=.02). The authors argue that these negative effects can be associated with examination anxiety; thus, VR serves as a well-being tool to reduce student stress.

The significance of the setting in the VR experience, namely nature, over other landscapes is explored within Valtchanov and Ellard's [9] study. A neutral geometric environment, an urban environment, and a natural environment were randomly allocated to students with induced stress. The nature setting increased positive effects significantly (*P*<.001) from a mean of 2.21 pre-VR to 3.03 post-VR compared to no effect of the other settings. Self-reported stress significantly decreased in the natural environment (*P*<.005), unlike the geometric or urban settings [9]. Additionally, differences in the biodiversity of natural environments in VR were also compared within the included studies. Gao et al [12] conclude that despite no significant differences between the VR settings regarding impact on physiological stress, positive mood, or attention, the authors reveal that restorations in negative mood were significantly different (*P*=.03). Partly open green spaces, defined in the study as having a 10%-30% composition of trees and shrubbery in the VR environment, had the highest significant effect on reducing negative mood (at *P*<.01) [12]. To summarize, natural environments presented in the medium of VR can promote positive mood, lowered stress, and overall well-being that can supplement a lack of access that students may face. The included studies feature VR through headsets, such as the Oculus Rift. Although such devices may be cheaper than a cell phone, the average student may not want to invest in a ~US $260 headset for the purpose of accessing well-being interventions [26]. Thus, future directions of implementing VR through cell phones that can be mounted onto cheap headsets, such as the Google Cardboard paired with audio, maybe a cost-efficient but impactful VR intervention with nature for students.

Measures of Well-Being via Stress Levels

Five of the included studies discussed interventions influencing well-being as seen through measures of stress levels. Among many therapy approaches to the reduction of stress and anxiety, relaxation techniques are seen to be very effective [17]. Specifically, the effectiveness of VR therapy was demonstrated through the use of VR environments and relaxing audio. Through the use of the PANSAS questionnaire, a mean score of 32 for positive affect was found as compared to a mean score of 15 for negative affect. Similarly, a study by Plante et al [27] looked at the impacts of VR (paired with aerobic exercise in a laboratory setting) on well-being and stress. Stress and energy were measured under 4 conditions: exercise outdoors, exercise with VR in laboratory conditions, exercise without VR in laboratory conditions, and VR without exercise. Outcomes were measured using Activation-Deactivation Adjective Checklists. It was found that exercising outdoors resulted in the most significant decreases in stress and increases in energy in men

(P<.10) and especially in women (P<.50). Women also felt calmer (P<.50) following the VR intervention, although the findings were not significant for the men. While there is room for more research, this finding strengthens the credibility of VR as an intervention to reduce stress.

The rapid development of technology in recent years has impacted how one perceives, interprets, and organizes their day-to-day lives, which may influence one's health and well-being [28]. When assessing well-being, an aspect that may be considered includes an individual's sense of presence, which can be described as how one experiences and engages in events and situations at the moment [28]. Specifically, in the case of digital environments and simulations, presence can be referred to as an experience where the user is immersed in the VR environment [29]. A study conducted by Villani et al [16] found that VR intervention involving an environment with nature can help induce psychological and physiological effects such as relaxation (P<.005) and lowered anxiety (P<.01). A similar study conducted by Villani and Riva [15] reported that VR can significantly reduce anxiety levels (P<0.05), manage a state of anxiousness (P<.005), and increase relaxation (P<.01). Both studies used the State-Trait Anxiety Inventory to gauge anxiety levels, in addition to employing the Coping Orientation to Problems Experienced Questionnaire to evaluate participant stress levels [15,16]. In the past, researchers have suggested that a greater sense of presence can increase the level of engagement with a digital simulation [29,30]. One study conducted by Mostajeran et al [31] assessed the cognition and stress levels of office workers before and after a VR intervention involving digital plants. It was found that a digital environment with plants (SD 0.82) significantly increased the sense of presence (P<.05) compared to a digital environment where there were no plants (SD 0.87) [31]. Additionally, it was observed that participants performed significantly better on memory tests such as the digit span backward test (P<.05) after the VR intervention [31]. The findings by Mostajeran et al [31] suggest that natural VR environments can help promote a sense of presence, which can have positive effects on productivity levels, as well as mediating stress, which aligns with the findings in this study.

However, not all of the studies showed a benefit from VR. In a study from 2015, a total of 32 female undergraduate students dealing with stress and anxiety in their transition to university completed a quiet ego contemplation intervention, which reminded them of 4 characteristics of the quiet ego: detached awareness, inclusive identity, perspective taking, and growth [14]. They were split into 3 groups: an audio recording describing the 4 characteristics, the same audio recording paired with VR of a park scene, and a control group in which participants perused nature magazines. The participants completed three 15-minute sessions of their assigned condition over 6 weeks. The study found that adding the VR component to quiet ego contemplation reduced the degree to which participants felt "in the moment" (t_{28}=2.66; P<.05). The authors hypothesized that this could be explained by the low-quality headset that was used, which was uncomfortable for participants and by the self-directed aspect of the VR, which may have been more distracting than a guided VR experience would have been.

Measures of Well-Being via Academic Contexts

The final theme extracted from the studies included covered measures of well-being as seen in academic contexts. These studies included participants in academic institutions who are facing many stressors as a result of their academic careers. Two of those studies recruited nursing students; 1 study used cognitive behavioral therapy and progressive muscle relaxation, and the other study administered a mindfulness meditation intervention [19,20]. Ortega et al [19] used the KEZKAK questionnaire and the state-trait anxiety test as a measuring tool for stress levels. They found overall lower scores for both tests among participants who underwent the intervention with P values of .019 for the KEZKAK questionnaire and .004 for the state-trait anxiety test [19]. Similarly, Chen et al [20] used the self-rating anxiety scale as a measurement tool, where lower rates of anxiety were reported for participants in the intervention group. In considering these findings, the strength of digital-based interventions to decrease stress is further reinforced. The authors of both studies argue that these types of intervention are necessary for students to be able to function healthily and excel in their fields [19,20].

A common intervention method among these studies was the use of VR headsets. Kaplan-Rakowski et al [18] used VR headsets in their study for a meditation intervention. This was used for the experimental group, while the control group went through the intervention by watching it on a monitor. Participants went through a pretest and posttest, which entailed basic computer science tasks. It was found that the VR group (0.03) showed a higher mean difference in test scores as compared to the control group (–0.19). Berezina et al's [23] intervention also used a VR headset. Participants in experimental group A watched a stimulating scene, followed by a relaxing scene, while participants in experimental group B watched a relaxing scene followed by a stimulating scene. Finally, there was a control group, which did not watch anything. They found that while both experimental groups experienced a decrease in fatigue, experimental group B showed a less pronounced decrease as compared to experimental group A. The final study on this theme used a VR intervention aimed at reducing stress. There were 3 groups: one which underwent a group mindfulness session, another which went through a VR session, and a control group. Modrego-Alarcón et al [22] found that participants in the mindfulness-based program showed the lowest levels of stress as compared to any other group. Additionally, the VR group scored lower stress levels compared to the control group. Overall, the studies show that VR, well-being interventions are effective in reducing stress levels and promoting relaxation among students, especially those in higher institutions.

Strengths and Limitations

Included studies analyzed within defined categories that supported qualitative analyses of excerpts within niches of productivity that are rarely explored in demographics of young adults. This systematic review provided a diverse understanding of various techniques and research approaches to well-being in psychology in the present literature. However, given the limited research conducted on the demographic of young adults within well-being from a digital standpoint, included studies demonstrated heterogeneity, which precluded meta and statistical analyses; thus, a qualitative analysis of studies was provided. Other limitations included exclusions based on languages that were not English, which do not extend the research boundaries to non–English-speaking demographics. Additionally, given the nuance of recent developments in technology, a lack of longitudinal studies may limit the application of this systematic review. Thus, elements that may be beneficial or consequential to study participants may not be adequately observed as well-being can be a culminated outcome of long-term behaviors.

Conclusions

In conclusion, this systematic review aimed at evaluating the efficacy of VR interventions to promote the well-being of students and young adults. To achieve this, studies were divided into themes of nature, stress, and academic contexts as focuses of interventions. Overall, the included studies reveal that VR interventions pose a promising medium to reduce the stress experienced by young adults and students, which can ultimately improve well-being. These findings reveal that VR may serve as an accessible, affordable tool for students and young adults to promote well-being or lower stress levels. However, there are some limitations to the review. The included studies tend to have smaller sample sizes, which may not be representative of students as a whole. A total of 20 studies were included in the final phase of extraction. Future directions may include expanding the search criteria to include more studies that may have higher sample sizes.

Acknowledgments

We would like to extend our sincere appreciation to all researchers and authors whose valuable contributions have enriched the body of literature in this field. We express our gratitude to Liza Iaralova, Valerie Lo, Cherry Tagra, and Ammal Riaz for their dedicated efforts in conducting literature searches, data extraction, and analysis. Their meticulous work has been instrumental in synthesizing the findings presented in this review. We also acknowledge the academic and institutional support that has facilitated our research endeavors. Lastly, we extend our thanks to the broader scientific community whose collective efforts continue to propel the exploration of novel technologies for enhancing well-being.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA 2020 checklist.
[PDF File (Adobe PDF File), 1080 KB - mental_v11i1e52186_app1.pdf]

References

1. APA Dictionary of Psychology. 2023. URL: https://dictionary.apa.org/ [accessed 2022-07-22]
2. What is wellbeing? University of Waterloo. Canadian Index of Wellbeing. 2016. URL: https://uwaterloo.ca/canadian-index-wellbeing/what-wellbeing [accessed 2022-07-22]
3. Piotrowski N, Hollar D. Magill's Medical Guide. Pasadena, Calif: Salem Press; 2010.
4. Ma Y, She Z, Siu AFY, Zeng X, Liu X. Effectiveness of online mindfulness-based interventions on psychological distress and the mediating role of emotion regulation. Front Psychol 2018;9:2090 [FREE Full text] [doi: 10.3389/fpsyg.2018.02090] [Medline: 30429816]
5. Virtual reality definition & meaning. Merriam-Webster. 2023. URL: https://www.merriam-webster.com/dictionary/virtual%20reality [accessed 2022-07-23]
6. Kalantari S, Xu TB, Mostafavi A, Lee A, Barankevich R, Boot WR, et al. Using a nature-based virtual reality environment for improving mood states and cognitive engagement in older adults: a mixed-method feasibility study. Innov Aging 2022;6(3):igac015 [FREE Full text] [doi: 10.1093/geroni/igac015] [Medline: 35592668]
7. Carnegie M. Are Gen Z the most stressed generation in the workplace? BBC Worklife. 2023. URL: https://www.bbc.com/worklife/article/20230215-are-gen-z-the-most-stressed [accessed 2023-07-22]
8. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
9. Valtchanov D, Ellard C. Physiological and affective responses to immersion in virtual reality: effects of nature and urban settings. J Cybertherapy Rehabil 2010;3(4):359-374 [FREE Full text]
10. O'Meara A, Cassarino M, Bolger A, Setti A. Virtual reality nature exposure and test anxiety. Multimodal Technol Interact 2020;4(4):75 [FREE Full text] [doi: 10.3390/mti4040075]
11. Browning MHEM, Mimnaugh KJ, van Riper CJ, Laurent HK, LaValle SM. Can simulated nature support mental health? Comparing short, single-doses of 360-degree nature videos in virtual reality with the outdoors. Front Psychol 2019;10:2667 [FREE Full text] [doi: 10.3389/fpsyg.2019.02667] [Medline: 32010003]

12. Gao T, Zhang T, Zhu L, Gao Y, Qiu L. Exploring psychophysiological restoration and individual preference in the different environments based on virtual reality. Int J Environ Res Public Health 2019;16(17):3102 [FREE Full text] [doi: 10.3390/ijerph16173102] [Medline: 31455015]
13. Villani D, Riva G. Does interactive media enhance the management of stress? Suggestions from a controlled study. Cyberpsychol Behav Soc Netw 2012;15(1):24-30 [FREE Full text] [doi: 10.1089/cyber.2011.0141] [Medline: 22032797]
14. Wayment HA, Collier AF, Birkett M, Traustadóttir T, Till RE. Brief quiet ego contemplation reduces oxidative stress and mind-wandering. Front Psychol 2015;6(101550902):1481 [FREE Full text] [doi: 10.3389/fpsyg.2015.01481] [Medline: 26483734]
15. Villani D, Riva G. Presence and relaxation: a preliminary controlled study. Psychol J 2008;6(1):7-25 [FREE Full text] [doi: 10.1037/e695432011-140]
16. Villani D, Riva F, Riva G. New technologies for relaxation: the role of presence. Int J Stress Manag 2007;14(3):260-274. [doi: 10.1037/1072-5245.14.3.260]
17. Taneja A, Vishal SB, Mahesh V, Geethanjali B. Virtual reality based neuro-rehabilitation for mental stress reduction. 2017 Presented at: 2017 Fourth International Conference on Signal Processing, Communication and Networking (ICSCN); March 16-18, 2017; Chennai, India URL: https://doi.org/10.1109/ICSCN.2017.8085665 [doi: 10.1109/icscn.2017.8085665]
18. Kaplan-Rakowski R, Johnson KR, Wojdynski T. The impact of virtual reality meditation on college students’ exam performance. Smart Learn Environ 2021;8(21):1-15 [FREE Full text] [doi: 10.1186/s40561-021-00166-7]
19. Ortega RH, Pascual JLG, Araque AMF. Impact of an intervention program with reinforcement on nursing students' stress and anxiety levels in their clinical practices. Nurse Educ Pract 2021;55:103179 [FREE Full text] [doi: 10.1016/j.nepr.2021.103179] [Medline: 34450399]
20. Chen Y, Yang X, Wang L, Zhang X. A randomized controlled trial of the effects of brief mindfulness meditation on anxiety symptoms and systolic blood pressure in Chinese nursing students. Nurse Educ Today 2013;33(10):1166-1172 [FREE Full text] [doi: 10.1016/j.nedt.2012.11.014] [Medline: 23260618]
21. Küchler AM, Schultchen D, Pollatos O, Moshagen M, Ebert DD, Baumeister H. StudiCare mindfulness-study protocol of a randomized controlled trial evaluating an internet- and mobile-based intervention for college students with no and "on demand" guidance. Trials 2020;21(1):975 [FREE Full text] [doi: 10.1186/s13063-020-04868-0] [Medline: 33243300]
22. Modrego-Alarcón M, López-Del-Hoyo Y, García-Campayo J, Pérez-Aranda A, Navarro-Gil M, Beltrán-Ruiz M, et al. Efficacy of a mindfulness-based programme with and without virtual reality support to reduce stress in university students: a randomized controlled trial. Behav Res Ther 2021;142:103866 [FREE Full text] [doi: 10.1016/j.brat.2021.103866] [Medline: 33957506]
23. Berezina T, Temirkanova A, Litvinova A, Kokurin A. Using virtual reality techniques to alleviatecognitive fatigue in graduate students working while in college. Eur J Contemp Educ 2022;11(1):36-46 [FREE Full text] [doi: 10.13187/ejced.2022.1.36]
24. Weir K. Nurtured by nature. Monit Psychol 2020;51(3):50 [FREE Full text]
25. White MP, Alcock I, Grellier J, Wheeler BW, Hartig T, Warber SL, et al. Spending at least 120 minutes a week in nature is associated with good health and wellbeing. Sci Rep 2019;9(1):7730 [FREE Full text] [doi: 10.1038/s41598-019-44097-3] [Medline: 31197192]
26. $50 off second-Gen VR headset - shop now. Meta Quest 2. 2023. URL: https://www.meta.com/ca/quest/products/quest-2/ [accessed 2023-12-24]
27. Plante TG, Cage C, Clements S, Stover A. Psychological benefits of exercise paired with virtual reality: outdoor exercise energizes whereas indoor virtual exercise relaxes. Int J Stress Manag 2006;13(1):108-117. [doi: 10.1037/1072-5245.13.1.108]
28. Zhou Z, Yin D, Gao Q. Sense of presence and subjective well-being in online pet watching: the moderation role of loneliness and perceived stress. Int J Environ Res Public Health 2020;17(23):9093 [FREE Full text] [doi: 10.3390/ijerph17239093] [Medline: 33291458]
29. Grassini S, Laumann K. Questionnaire measures and physiological correlates of presence: a systematic review. Front Psychol 2020;11:349 [FREE Full text] [doi: 10.3389/fpsyg.2020.00349] [Medline: 32265769]
30. Cummings JJ, Bailenson JN. How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. Media Psychol 2015;19(2):272-309 [FREE Full text] [doi: 10.1080/15213269.2015.1015740]
31. Mostajeran F, Steinicke F, Reinhart S, Stuerzlinger W, Riecke BE, Kühn S. Adding virtual plants leads to higher cognitive performance and psychological well-being in virtual reality. Sci Rep 2023;13(1):8053 [FREE Full text] [doi: 10.1038/s41598-023-34718-3] [Medline: 37198210]

Abbreviations

AMSTAR-2: Assessing the Methodological Quality of Systematic Reviews
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
MBI: mindfulness-based intervention
PRISMA-P: Preferred Reporting Items for Systematic Review and Meta-Analysis Protocol
VR: virtual reality

Edited by J Torous; submitted 25.08.23; peer-reviewed by J Ferrer Costa, M Slater; comments to author 01.11.23; revised version received 24.12.23; accepted 01.01.24; published 29.03.24.

Please cite as:
Xu J, Khanotia A, Juni S, Ku J, Sami H, Lin V, Walterson R, Payne E, Jo H, Rahimpoor-Marnani P
Effectiveness of Virtual Reality–Based Well-Being Interventions for Stress Reduction in Young Adults: Systematic Review
JMIR Ment Health 2024;11:e52186
URL: https://mental.jmir.org/2024/1/e52186
doi:10.2196/52186
PMID:38551625

Review

Studies of Social Anxiety Using Ambulatory Assessment: Systematic Review

Javier Fernández-Álvarez[1,2], PhD; Desirée Colombo[1], PhD; Juan Martín Gómez Penedo[3], PhD; Maitena Pierantonelli[4], MSc; Rosa María Baños[4,5,6], Prof Dr; Cristina Botella[1,6], Prof Dr

[1]Department of Basic and Clinical Psychology and Psychobiology, Jaume I University, Castellon de la Plana, Spain

[2]Fundación Aiglé, Buenos Aires, Argentina

[3]Facultad de Psicología, Universidad de Buenos Aires (CONICET), Buenos Aires, Argentina

[4]Polibienestar Research Institute, University of Valencia, Valencia, Spain

[5]Department of Personality, Evaluation, and Psychological Treatments, University of Valencia, Valencia, Spain

[6]Ciber Fisiopatologia Obesidad y Nutricion (CB06/03 Instituto Salud Carlos III), Madrid, Spain

Corresponding Author:
Javier Fernández-Álvarez, PhD
Department of Basic and Clinical Psychology and Psychobiology
Jaume I University
Avda. Vicent Sos Baynat s/n
Castellon de la Plana, 12071
Spain
Phone: 34 964 72 80 0
Email: javferalvarez@gmail.com

Abstract

Background: There has been an increased interest in understanding social anxiety (SA) and SA disorder (SAD) antecedents and consequences as they occur in real time, resulting in a proliferation of studies using ambulatory assessment (AA). Despite the exponential growth of research in this area, these studies have not been synthesized yet.

Objective: This review aimed to identify and describe the latest advances in the understanding of SA and SAD through the use of AA.

Methods: Following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, a systematic literature search was conducted in Scopus, PubMed, and Web of Science.

Results: A total of 70 articles met the inclusion criteria. The qualitative synthesis of these studies showed that AA permitted the exploration of the emotional, cognitive, and behavioral dynamics associated with the experience of SA and SAD. In line with the available models of SA and SAD, emotion regulation, perseverative cognition, cognitive factors, substance use, and interactional patterns were the principal topics of the included studies. In addition, the incorporation of AA to study psychological interventions, multimodal assessment using sensors and biosensors, and transcultural differences were some of the identified emerging topics.

Conclusions: AA constitutes a very powerful methodology to grasp SA from a complementary perspective to laboratory experiments and usual self-report measures, shedding light on the cognitive, emotional, and behavioral antecedents and consequences of SA and the development and maintenance of SAD as a mental disorder.

(JMIR Ment Health 2024;11:e46593) doi:10.2196/46593

KEYWORDS

social anxiety disorder; ambulatory assessment; ecological momentary assessment; intensive longitudinal methods; systematic review; social anxiety; use; qualitative synthesis; emotional; cognitive; behavioral; development; mental disorder; anxiety; mental health; mobile health; mHealth; monitoring; review; assessment; mobile phone

Introduction

Background

Social anxiety (SA) is a normal and adaptive manifestation that all human beings experience in anticipation of a potential interactional threat. Similar to any adaptive response, it is enormously beneficial, particularly in protecting people from potential dangers that may arise from social interactions [1]. However, this adaptive response is sometimes exacerbated, and instead of preparing the individual for optimal performance, it becomes paralyzing, triggering intense fear, catastrophic thoughts, and avoidance behaviors, among other characteristic manifestations. When this pathological response becomes habitual in anticipation and confrontation of social interactions, it is referred to as SA disorder (SAD).

Accordingly, SAD is understood as a prevalent clinical condition characterized by intense fear and avoidance of social situations. SAD is a heterogeneous clinical condition that usually entails high levels of dysfunction in the lives of people who experience it. This heterogeneous nature of SAD is marked by the dynamic deployment of cognitions, emotions, and behaviors in the interaction with others and in different contexts [2].

Ambulatory Assessment

Ambulatory assessment (AA) serves as an umbrella term that includes specific techniques such as experience sampling methods, ecological momentary assessment, or daily retrospective methods. These techniques constitute a research methodology of paramount significance in the field of clinical psychology and psychotherapy, enabling researchers and clinicians to gather in-depth, ecologically valid data from individuals. Unlike traditional assessment methods that rely on retrospective self-reporting, AA involves the repeated real-time measurement and gauging of various aspects of an individual's experiences, behaviors, and physiological responses. The fact that it entails multiple assessments over a certain period makes AA an optimal tool to explore within-person fluctuations and trajectories of these experiences and behaviors.

Moreover, AA circumvents the biases that usual retrospective reports may have because it is usually implemented in momentary assessments or recent retrospective reports (eg, daily diaries [3]). Owing to the variability of psychological processes such as affective and emotional dynamics [4], AA can more precisely determine the fluctuation of symptoms. In addition, owing to the possibility of incorporating sensors and biosensors, AA can provide multimodal assessment, overcoming the biases of self-reports [5].

Over the last few years, there has been an increasing interest in exploring the use of AA in clinical psychology and psychotherapy. Owing to the incorporation of digital technologies, namely, mobile phones, AA has become a very powerful add-on for clinical researchers [6], first and foremost due to the possibility that AA provides of capturing data in real time, thus grasping contextual aspects from naturalistic settings that laboratory-based assessment does not allow for [7,8]. SA symptoms are particularly contextually bound, and thus, it is of utmost relevance to consider the sensitivity of the context [2].

All these characteristics have a central and common pursuit to personalize models of psychopathology [9]. There is no doubt that every person has a certain and unique composition of traits that may lead to functional and dysfunctional states in continuous interaction with the context. Therefore, the revolution of AA is helping foster the creation of tailored models with intensive longitudinal data, which can shed light on the factors that may lead people to experience and behave in adaptive or maladaptive ways.

In this sense, AA proves to be a suitable tool for capturing the co-occurrence of symptoms and psychological processes, offering crucial insights into the antecedents and consequences of SA with enhanced accuracy. This, in turn, facilitates a comprehensive understanding of the factors contributing to the appearance and maintenance of SA and SAD. Given the dimensional nature of SA, AA can shed light on the differences among clinical, subclinical, and healthy participants.

The understanding of SAD in ecological settings may also lead to improving psychological treatments. The insights gained through AA can contribute to refining therapeutic approaches by investigating mechanisms of change [10]. However, to the best of our knowledge, no systematic review has synthesized the available evidence analyzing the strengths and limitations of the current state of the art.

This Study

For all the aforementioned reasons, AA provides a range of advantages that justify its exponential growth as a research methodology in clinical psychology. Although a plethora of research has implemented AA in individuals with SA, these data have not been synthesized yet. Hence, the main aim of this review was to identify studies that used AA to explore SA.

Methods

This review followed the recommendations of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [11]. The full protocol was registered before the data analysis [12].

Literature Search

Scopus, PubMed, and Web of Science literature searches were conducted. The search strategy for PubMed is available in Multimedia Appendix 1 and was adapted to the syntax requirements of each database. No filters were included in the searches. The reference lists of eligible articles were manually reviewed to identify additional relevant publications.

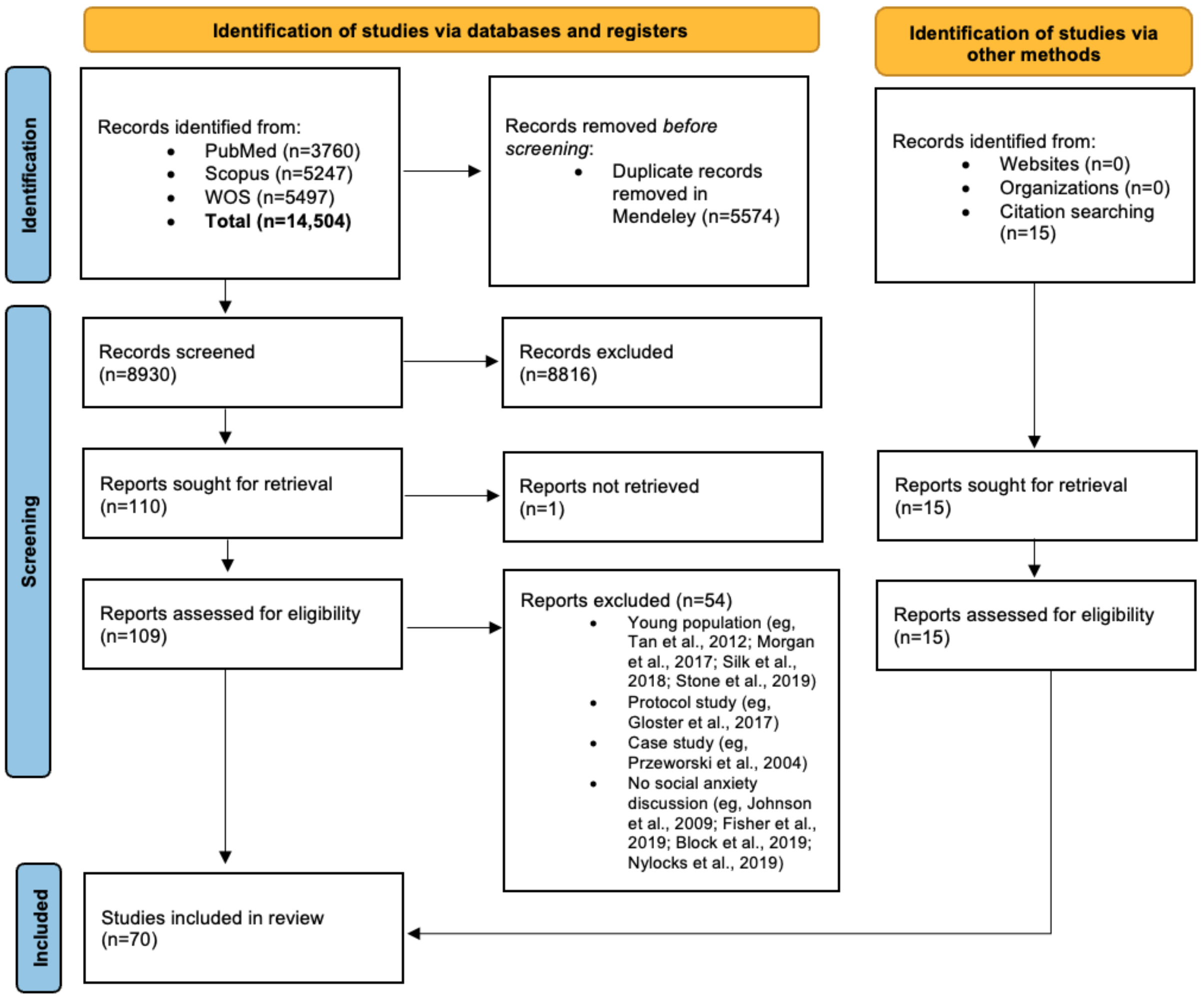
Inclusion Criteria

To be included, the studies had to fulfill the following criteria: (1) studies that focused on participants with SAD or SA symptoms; (2) studies that used AA to explore the dynamics of SA, correlates of SA, or the feasibility of AA for SA assessment; (3) articles that included any sort of discussion regarding SA in AA (this criterion was included given that some articles fulfilled all the previous criteria but featured no specific discussion on SA); (4) articles published in peer-reviewed journals in English; and (5) studies that presented at least one active AA or passive data collected via sensor signals (eg, geo-mapping, accelerometer, or GPS) from a phone or smartwatch multiple times per day. Articles were excluded if they (1) were not empirical studies; (2) used a child or adolescent sample; and (3) did not use a daily diary, experience sampling, or AA approach that included assessment and discussion of SA.

Study Selection and Data Extraction

A database search was conducted until August 31, 2023. The literature search produced a total of 14,504 articles, 8926 (61.54%) of which were retained after removing duplicates. A total of 1.23% (110/8926) of these articles were retrieved after title and abstract screening. The subsequent application of the selection criteria resulted in the inclusion of 50% (55/110) of the articles, adding another 15 studies that were identified through citation searching. In total, we included 70 studies (Figure 1). Studies were independently selected by 2 researchers, and disagreements were resolved through consensus.

Figure 1. Systematic review search flow diagram. WOS: Web of Science.



Using premade tables, the following information was extracted: author and year of publication, study design, follow-up period and sample size and characteristics, aims of the study, measures collected through AA, app name in the case of smartphone-based AA, participation rate (ie, response to recruitment), attrition rate, compliance with AA questions, incentives, outcomes explored, AA measures, frequency of AA, and main findings.

Finally, to summarize the main information, 2 of the authors proposed categories. To form a category, it was required that at least 2 of the included studies addressed the same or a similar topic. To reach an agreement, the resulting categories underwent thorough discussion between the first and fourth authors. This criterion was applied to the "Principal themes explored" section. For the "Emerging topics" section, the criterion for category inclusion was based on relevance regarding addressing a key aspect of the literature that had not been explored before.

Results

Sample Characteristics

Of the 70 studies selected for this systematic review, 29 (41%) used the same sample (Textbox 1). Given that a large proportion of the studies were conducted in university settings, most of the included participants were undergraduate students and, therefore, young adults. Moreover, all the studies (70/70, 100%) were conducted in high-income countries, principally in the United States. A total of 43% (30/70) of the studies included both healthy and clinical populations. In total, 21% (15/70) of the studies were conducted only with clinical participants, whereas 36% (25/70) were conducted with healthy individuals exploring the dynamics of SA. In addition, 33% (23/70) of the studies included comorbid samples, mainly those with other anxiety or depressive disorders. Regarding gender, most studies (57/70, 81%) included more women than men. This reflects the prevalence rates of SAD, which have been shown to be higher in women than in men [13]. Table S1 in Multimedia Appendix 2 [14-81] and Table S2 in Multimedia Appendix 3 [14-81] summarize the characteristics and main findings of the included studies.

Textbox 1. Studies with the same samples.

Studies with the same samples

- Blalock et al [14,15], Farmer and Kashdan [16], Goodman et al [17,18], Kashdan and Farmer [19], and Kashdan et al [20,21]
- Daniel et al [22,82,83] and Beltzer et al [23]
- Daniel et al [24], Daros et al [25], and Ladis et al [26]
- Doorley et al [27,28]
- Goodman et al [29-32]
- Oren-Yagoda et al [33-36]
- O'Toole et al [37,38]
- Villanueva et al [39] and Rinner et al [40]

Methodological Characteristics of the Studies

Sampling Frequency

There were 2 types of sampling identified among the included studies. On the one hand, there were some studies that implemented a daily diary, which typically entails an end-of-day report. On the other hand, other AA studies comprised different numbers of daily assessments and diverse types of prompt contingencies. The duration of the AA ranged from 4 to 35 days. The most commonly implemented study duration was 14 days.

Although a large proportion of the studies (63/70, 90%) used a fixed or random scheduled prompt structure, a range of studies incorporated an event-contingent design [40-45]. Indeed, SA determinants, correlates, and consequences are particularly triggered in interpersonal situations, and event-contingent designs may be suitable for detecting relevant moments. Finally, studies such as those by Daniel et al [24], Helbig-Lang et al [46], and Kashdan et al [20] included a mix of different types of data collection, combining random prompt, event-contingent, and end-of-day records.

Compliance

Compliance rates typically indicate the percentage of prompts or days that the participants completed on average. Compliance rates ranged from 40% to 95%, with average compliance across studies of 73.72% of prompts (SD 15.93%; median 80%).

Statistical Analysis

The vast majority of the studies (61/70, 87%) used hierarchical linear models to account for the nested nature of the collected data. A total of 4% (3/70) of the studies used a combination of hierarchical linear models and structural equation modeling [43,45,79], 1% (1/70) of the studies used machine learning techniques [47], 3% (2/70) of the studies calculated ANOVA models [23,42], and 1% (1/70) of the studies ran ordinary least squares regression [48].

Methodological Design Characteristics

Table 1 summarizes the principal methodological characteristics of each study. Power analysis to calculate the needed sample size (which was rarely conducted; 15/70, 21% of the studies) and the psychometric properties of the AA questions (37/70, 53% of the studies) were the 2 criteria that were reported the least, whereas the percentage of AA compliance (56/70, 80% of the studies) and attrition rates of AA (46/70, 66% of the studies) were the criteria that were reported the most among the studies in this review.

Table 1. Methodological design characteristics.

Study	Compensation	Power analysis for sample size	Attrition rates	Psychometric properties of AA[a]	Compliance with AA
Arch et al [49]	Yes	+[b]	+	–[c]	+
Badra et al [50]	Yes	+	+	+	+
Bailey et al [51]	Yes	+	Unclear	–	+
Battista et al [52]	Yes	–	+	+	+
Beltzer et al [23]	Yes	–	–	+	–
Blalock et al [14]	Yes	–	+	+	+
Blalock et al [15]	No	–	+	–	+
Boukhechba et al [53]	Yes	–	–	–	–
Brown et al [54]	No	–	–	–	+
Brown et al [55]	Yes	–	+	–	–
Buckner et al [56]	No	–	+	+	+
Buckner et al [57]	Yes	–	+	–	+
Chow et al [58]	Yes	–	+	–	+
Čolić et al [59]	No	+	+	–	Unclear
Daniel et al [24]	No	–	+	–	+
Daniel et al [83]	Yes	–	+	–	Unclear
Daniel et al [82]	Yes	–	+	+	–
Daniel et al [22]	No	–	–	+	+
Daros et al [25]	Yes	–	+	–	+
Di Matteo et al [60]	Yes	–	–	+	–
Doorley et al [28]	Yes	–	+	+	Unclear
Doorley et al [27]	Yes	–	+	+	+
Farmer and Kashdan [61]	Yes	–	+	+	+
Farmer and Kashdan [16]	Yes	–	–	+	–
Geyer et al [62]	Yes	–	–	–	+
Goodman et al [63]	Yes	–	+	+	+
Goodman et al [17]	Yes	–	+	+	+
Goodman et al [18]	Yes	+	+	+	+
Goodman et al [32]	Yes	+	–	–	–
Goodman et al [30]	Yes	+	–	–	–
Goodman et al [31]	Yes	–	–	+	–
Goodman et al [29]	Yes	–	+	+	+
Goodman et al [64]	Yes	+	+	–	+
Hannah Lee [65]	Yes	–	+	–	+
Helbig-Lang et al [46]	No	–	–	+	–
Hur et al [66]	Yes	+	+	+	+
Jacobson et al [47]	No	+	+	+	N/A[d]
Jacobson and Bhattacharya [67]	No	–	–	+	–
Kane and Ashbaugh [68]	No	+	–	+	–
Kashdan and Steger [69]	No	–	+	+	–
Kashdan and Collins [70]	Yes	–	–	–	+

Study	Compensation	Power analysis for sample size	Attrition rates	Psychometric properties of AA[a]	Compliance with AA
Kashdan et al [71]	No	−	+	−	+
Kashdan et al [20]	Yes	−	+	+	+
Kashdan et al [21]	Yes	−	+	+	+
Kashdan and Farmer [19]	Yes	−	+	−	+
Katz et al [41]	No	−	−	−	−
Kivity and Huppert [72]	Yes	−	+	−	+
Kim and Kwon [73]	Yes	+	+	−	+
Ladis et al [26]	Yes	−	+	−	+
Lee et al [42]	Yes	−	+	+	+
Nanamori et al [74]	Yes	−	−	+	+
Naragon-Gainey [43]	Yes	−	+	+	+
O'Grady et al [75]	Yes	−	+	−	+
Oren-Yagoda et al [35]	Yes	−	−	+	+
Oren-Yagoda and Aderka [33]	Yes	−	−	+	+
Oren-Yagoda et al [36]	Yes	−	−	+	+
Oren-Yagoda et al [34]	Yes	−	−	+	+
O'Toole et al [37]	Yes	−	+	+	+
O'Toole et al [38]	Yes	−	+	+	+
Papp et al [76]	Yes	−	−	−	+
Park and Naragon-Gainey [45]	Yes	−	+	−	+
Piccirillo and Robedaugh [77]	Yes	−	−	+	−
Reichenberger et al [78]	Yes	−	+	−	+
Rinner et al [40]	No	+	+	−	Unclear
Russell et al [44]	Yes	−	+	−	+
Saulnier et al [79]	Yes	−	−	+	+
Seah et al [48]	Yes	−	+	+	+
Villanueva et al [39]	No	+	+	−	Unclear
Walukevich-Dienst et al [80]	Yes	−	−	−	+
Wilson et al [81]	Yes	−	+	−	+

[a]AA: ambulatory assessment.

[b]Presence.

[c]Absence.

[d]N/A: not applicable.

Principal Themes Explored

Affective and Emotional Dynamics

Naragon-Gainey [43] explored structural models of affect and internalizing symptoms. While the between-person variance in negative affect (NA) and concurrent levels of NA predicted SA, positive affect (PA) did not. This lack of significance contradicted many other AA studies that yielded significant between-person associations between lower levels of PA and higher levels of NA with SA [16,20,30,54,57,58,61,69,70]. Individuals with SAD showed not only higher overall levels of NA but also more emotional instability, which was not the case for PA [16]. These authors even suggested that the interaction between NA and instability could explain the appearance of SAD.

Discrete Emotions

As suggested by Rozen and Aderka [84], there is a wide range of discrete emotions that have not been integrated into classic models of SAD. In that sense, AA allows for a nuanced study of single emotions and how they are interconnected not only with each other but also with other cognitive-affective processes and behaviors.

First, Kashdan and Collins [70] revealed that SA was related to less time spent feeling happy and relaxed and more time spent feeling angry. The results also showed that happy moments were aroused in companion to others. The fact that SA was associated with fewer and less intense positive emotions and more anger episodes was independent of being with others or alone.

Oren-Yagoda et al [35] investigated the role of envy, showing that visual modes of communication are related to elevated envy compared to voice or text communication. In addition, envy predicted subsequent anxiety above and beyond previous anxiety as well as other negative emotions.

Loneliness is a defining emotion of people with SAD, and Oren-Yagoda et al [36] found that this relationship was also confirmed in an ecological setting using AA. A significant association was predicted in certain social situations (ie, negativity, positivity, and meaningfulness). Moreover, the relationship between loneliness and anxiety was shown to be reciprocal in individuals with SAD (loneliness predicted anxiety and anxiety predicted loneliness). This deleterious reciprocity was not found in healthy controls.

Finally, Oren-Yagoda et al [34] investigated the fluctuations of pride in individuals diagnosed with SAD. The results indicated that levels of pride were lower in patients with SAD than in nondiagnosed controls, although when pride was experienced, it predicted a reduction in anxiety levels.

Emotion Regulation

Emotion regulation (ER) was by far the most explored construct among the studies that implemented AA in individuals with SA. This may be explained by the fact that ER is a dynamic process, and AA is very suitable for detecting affect changes and fluctuations in daily life. In addition, not only are ER and SA symptoms highly context sensitive, but mounting research has also shown their interdependence [85], which justifies the importance of exploring how individuals with SAD use ER in daily life.

Discrete Strategies

Several studies explored only 1 strategy. For example, Kashdan et al [20] found that individuals with SAD experienced greater experiential avoidance than healthy individuals. Using the same sample, Kashdan et al [21] found that momentary experiential avoidance presented a stronger association with anxiety during social interactions for individuals with SAD than for individuals without SAD.

Similarly, another study explored the role of emotional suppression, demonstrating that, on days in which SA symptoms increased, the use of suppression tended also to increase [61,69]. The same was found by Beltzer et al [23], who identified that days with high levels of SA and expressive suppression led to fewer positive emotions.

Farmer and Kashdan [61] implemented a 2-week diary, showing that high SA was related to positive emotion suppression, fewer positive social events, and fewer positive emotions on subsequent days. In contrast, low SA was associated with fewer negative social events on the days after cognitive reappraisal was used to reduce distress. However, the use of cognitive reappraisal did not lead to any changes in people with high SA.

Meanwhile, Farmer and Kashdan [16] found that individuals with SAD were 3 times more likely to present acute shifts in NA, which may be a particular experience of this anxiety disorder and not others. This acute shift in NA may lead to experiencing emotions as uncontrollable and threatening, which in turn could explain the higher use of suppression.

Goodman et al [63] showed that alcohol consumption moderated the negative association between SA and a range of healthy social interactions such as laughter or feelings of acceptance; that is, SA was not related to the perceived quality of interpersonal interactions when participants consumed alcohol, suggesting that alcohol consumption may be a reinforcer of SA. These findings suggest that ER plays a central role in the experience of individuals with SAD who try to explicitly control their emotions and are aware of the effort they make to do so. In a different but related study, Goodman and Kashdan [29] found that both anxiety and pain were interfering factors in goal attainment, as well as finding an inverse association between daily meaning in life and perceived emotion-related goal interference.

Substance Use

Alcohol consumption and SAD are highly comorbid given that, among individuals with SAD, it is frequent to resort to alcohol as a coping strategy or, in other words, as an ER strategy. By means of AA, it is possible to detect how these 2 phenomena are interrelated. Therefore, several studies explored this topic. Battista et al [52] examined the relationship between alcohol consumption and SA, revealing that, after each alcoholic drink consumed, SA tended to decrease 2 hours later. Contrary to what the authors expected, this association was not explained by the level of trait SA.

Goodman et al [63] showed that alcohol consumption moderated the negative association between SA and a range of healthy social interactions such as laughter or feelings of acceptance; that is, alcohol consumption SA was no longer related to the perceived quality of interpersonal interactions, suggesting that SA may be a reinforcer of SA. This reinforcement could be either negative (attenuation of anxiety) or positive (better perception of social situations), which was the main finding of Goodman et al [64]. The results of this study yielded evidence supporting the negative reinforcement hypothesis such that people with SAD presented higher coping motives (negative reinforcers) but equal levels of affiliation motives (positive reinforcers).

Walukevich-Dienst et al [80] revealed that consuming substances as a coping strategy in SA is related to heavier consumption, especially on drinking days, which may be a risk factor for the development of an alcohol use disorder. Moreover, when there were SA coping motives, the consequences were more negative compared to days without SA coping motives. Interestingly, levels of SA at baseline were not moderators of these associations, indicating that the coping motive is more important than the antecedent levels of SA.

Kim and Kwon [73] showed that individuals with SAD had a higher increase rate of alcohol craving when they were tense and lonely and experiencing SA in comparison to individuals without a diagnosis of SAD. These results were moderated by the rate of rumination in the SAD group and avoidance in the non-SAD group. O'Grady et al [75] explored the role of social-contextual events in the relationship between trait SA and drinking. The results revealed a positive association between trait SA and drinking on the evenings of days in which the individuals experienced an embarrassing situation. This was significantly higher than in healthy participants, suggesting a behavioral maladaptive ER strategy to cope with the intensified levels of SA.

In addition, Buckner et al [56] studied the association between SA, cannabis use, cannabis craving, and situational variables. This study showed that SA interacted with cannabis craving predicting cannabis use, which sets out a complex relationship between SA and cannabis use and not a simple association (ie, cannabis use as a response to increased state SA).

In the study by Buckner et al [57], they showed that baseline SA was associated with increases in NA throughout the days in which the participants were monitored but was also significantly associated with postquit withdrawal. Participants with higher levels of SA presented more severe postquit withdrawal symptoms as well as an increase in NA during a cessation attempt. For this reason, the authors suggested that those participants may particularly benefit from intervention and treatment strategies.

Finally, Papp et al [76] found that students with higher levels of SA presented a stronger association between NA and prescription misuse. This study included externalizing and internalizing symptoms (depression and SA symptoms), and moderation was shown to be significant only for internalizing symptoms. This may indicate differential patterns of substance use as an ER strategy according to different personality traits.

The Complexity of ER: Polyregulation and Flexibility

Some studies incorporated a range of ER strategies in the AA process, which is in line with the current idea of *polyregulation*, that is, that individuals usually implement a range of strategies simultaneously [86]. For example, Daros et al [25] measured various strategies, and after clustering through a factorial analysis, 2 categories of ER strategies were considered for the analyses: avoidant and engagement strategies. However, the authors did not find any significant results for these 2 macrocategories.

On the basis of the model by Gross, Blalock et al [14] examined both suppression and cognitive reappraisal in SA in comparison to healthy individuals. Both groups presented worse emotional experiences when they suppressed positive versus negative emotions as well as when reappraising negative versus positive emotions. This suggests that suppression may be an adaptive strategy not to feel negative emotions, whereas reappraisal may be effective in increasing positive states. Interestingly, individuals with SAD showed more positive emotions after reappraising negative emotional states to feel fewer negative emotions than the healthy controls.

However, none of these previous examples explicitly framed their studies under the concept of emotion polyregulation, as did Ladis et al [26]. By including 8 strategies (problem-solving, introspection, distraction, acceptance, thought suppression, seeking advice, cognitive reappraisal, and expressive suppression), the authors systematically explored how often polyregulation occurs in daily life. Overall, the results yielded nonsignificant SA correlates of polyregulation, suggesting that it may be more dependent on within-person differences.

In addition to polyregulation, flexibility has been shown to be central in ER literature to distinguish adaptive from maladaptive regulatory processes [87]. Given that flexibility is mostly dependent on context, AA emerges as a very useful tool. O'Toole et al [38] investigated this specific topic in individuals with high and low SA and considered type and intensity as 2 contextual factors. This study revealed that SA moderated the relationship between emotion intensity and experiential avoidance. In individuals with high SA, there was a stronger association between experiential avoidance and specific emotions, such as guilt, nervousness, and sadness.

The study conducted by Beltzer et al [23] presents an illustrative example of how ER can be better explained by contextual triggers than by the adaptative or maladaptive continuum. The authors created an algorithm based on the 10 most used strategies to generate an ecological momentary intervention (EMI) based on contextual triggers. This algorithm was tested with a group of strategies (disengagement, engagement, and aversive cognitive perseveration) and with 10 individual single strategies compared with a random policy and a behavior policy, with ER effectiveness being the observed outcome. The contextual algorithm improved other policies in cases in which the top 10 strategies were considered separately. However, when the strategies were grouped into categories, the algorithm did not outperform the random recommender or the observed ER strategies.

Goodman et al [32] also explored ER flexibility considering 2 components of flexibility: the evaluation of contextual demands and matching regulatory strategies to contextual demands. The results indicated that people with SAD considered momentary assessments to be more anxiety provoking while presenting similar patterns to those of control participants when gauging contextual demands, particularly those related to perceived controllability. That is, some disengagement strategies (rumination, thought suppression, and expressive suppression) were found to be unrelated to perceived controllability. This means that these strategies were used independently of perceived controllability in a certain context. However, contrary to previous results, participants with SAD yielded similar patterns to those of control participants in response to anxiety intensity.

Higher anxiety ratings predicted greater use of all strategies regardless of the type of strategy used. In addition, this study showed that people with SAD may be more prone to use thought suppression but not engagement strategies, which is inconsistent with previous studies.

In other studies, a specific component of the heterogeneous and complex process of ER was explored. Although most clinical research on ER tends to simplify the discussion on putatively maladaptive (eg, expressive suppression) or adaptive (eg, cognitive reappraisal) strategies, there is a range of potential explanatory variables that set a more complex scenario than implementing certain strategies or not. For example, one study [24] investigated the perception of ER effectiveness, which is based on a robust research line that revolves around how goals shape and determine ER deployment and outcomes [88]. Daniel et al [24] found that, depending on the way in which effectiveness is measured (either judgment of effectiveness or change in affect), the results differ. While the judgment of effectiveness indicates that avoidance - oriented ER attempts are less effective than engagement - oriented ER attempts, changes in self-reported effects following ER attempts present the opposite results.

Similarly, Goodman et al [18] explored the extent to which beliefs of ER determine the use of specific patterns. In laboratory settings, De Castella et al [89] demonstrated that individuals with SAD presented low emotional self-efficacy, or the belief that emotions cannot be changed, but the results obtained by Goodman et al [18] provide ecological validity to a result that confirms the interdependency between cognitions and emotions [90].

Another approach is to calculate ER diversity, defined as the variety, frequency, and evenness of the implemented ER. Finally, Daniel et al [22] studied whether ER diversity predicted SA severity. The results showed that diversity within avoidance-oriented strategies was associated with both trait and state SA levels. At a more specific level of analysis, participants who responded more evenly and deployed a vast array of avoidance-oriented strategies more frequently were more prone to belonging to the high-SA group.

Emotion Clarity and Differentiation

Another important aspect that has been of increasing interest in the ER literature is linked to convergent processes such as emotion clarity and differentiation. Emotion clarity is defined as the ability to identify, distinguish, and describe specific emotions [91]. Park and Naragon-Gainey [45] found that lower momentary clarity was related to increases in subsequent momentary internalizing symptoms (ie, anxiety and depressive symptoms). This association was explained by an unsuccessful use of ER.

A total of 4% (3/70) of the studies examined emotion differentiation in individuals with SAD [19,37,48]. Emotion differentiation is conceptualized as the ability to recognize, identify, and label broad emotional experiences into discrete emotion categories [92]. Kashdan and Farmer [19] demonstrated that an increase in SA symptoms is linked to an impairment in negative emotion differentiation (ie, the ability to label and describe differences among negative emotions). In particular, it was found that negative emotion differentiation plays a relevant role in implementing more effective ER strategies. For example, Seah et al [48] conducted 2 studies that showed that negative emotion differentiation moderated the positive relationship between rumination and social avoidance. Similarly, O'Toole et al [37] showed that individuals with both high SA and poor negative emotion differentiation presented the least use of cognitive reappraisal. In addition, individuals with high SA used more suppression strategies despite the ability to differentiate positive emotions.

Perseverative Cognition and Mind Wandering

Although the most common variable that has been studied in individuals with SAD is postevent processing (PEP), there is a great overlap between PEP and rumination. In essence, both revolve around the perseverative thinking of past events, but PEP is strictly related to social interactions, including an important component of the actual interventions that both the individual and the others may have done or said. In total, 6% (4/70) of the studies used AA to examine processes related to PEP and rumination.

Helbig-Lang et al [46] studied individuals diagnosed with SAD to explore predictors of higher PEP levels. The results showed that higher PEP was predicted by self-attention, NA, social performance situations, and the use of safety behaviors. In addition, Badra et al [50] investigated PEP in a nonclinical sample of undergraduate students with high and low SA scores. Although no differences were detected between the groups, PEP was reduced to a single item, and no additional information on the social context was collected, which could have affected the results.

Another study that explored PEP showed that it decreased after a speech task [68]. The between-level average differed from the person-specific trajectories. This was the case not only for the decrease in PEP after the speech task but also in the temporal cascading relationship between PEP and the next measurement of PEP. The level of anxiety in the speech task predicted engagement in PEP, and this activated a more intense experience of the negative balanced memory, indicating the interconnectedness of cognitive-affective processes.

Bailey et al [51] studied perseverative cognition (worry and rumination) using physiological measures and found that there was a higher use of this type of repetitive thinking related to lower heart rate variability after negative social interactions. Finally, potential changes in momentary PEP throughout treatment were explored by Katz et al [41] showing that cognitive behavioral group therapy reduced levels of PEP.

In total, 3% (2/70) of the studies explored mind wandering, which is both an interesting and controversial topic. Traditionally, it was considered that mind wandering was just a negative process given that it was linked to the opposite of having a mindful disposition. However, the latest research has shown that it could be related to less boredom, more creativity, and a better mental health state [93].

In the study by Arch et al [49], the authors investigated differences in the frequency, range of content, and correlates of internal off-task thinking (ie, mind wandering). Relative to on-task thinking, internal off-task thinking was associated with worse mood, more self-focus, and less thought controllability for those with SAD compared to healthy controls. In addition, participants with SAD engaged in internal unrelated task thinking more frequently than those in the control group and presented more unintentional mind wandering on a trait questionnaire.

Specific Interactional Triggers, Patterns, and Activities

Geyer et al [62] investigated the association between NA in social interactions and perceived enjoyment of those interactions to explore specific real-time contributors to negative perceptions often experienced by individuals with SA. The results revealed that this association was more negative when SA was more severe, although the sample consisted of undiagnosed individuals.

Meanwhile, Blalock et al [15] studied the experience of flow in social and nonsocial situations in individuals with SAD and healthy participants. The results were contrary to the hypotheses, revealing that individuals with SAD presented flow more frequently in social situations than healthy participants. The authors explained this unexpected result using the concept of flow, which includes the component of experiencing a challenging situation as a defining feature. It is reasonable to think that people with SAD experience normal social interactions as more challenging than healthy participants.

In one study, the association between sexual activity and SA was explored in nonclinical individuals [71]. As could be expected, sexual activity was influenced by SA such that individuals with SA reported their sexual episodes as less pleasurable and reported being less connected with their partners as well as presenting a lower frequency of sexual activity. Overall, these results suggest that sexual activity is not fulfilling when experiencing SA.

Goodman et al [30] conducted 2 AA studies in which they concluded that, although individuals with SAD present fewer and less satisfying social relationships, they enjoy social interactions when they occur, which might indicate that they are happier with others than alone. In other words, experiencing SA and the relative concern about socializing does not hinder the pleasure of socializing.

These results are aligned with those of other 2 studies conducted on undergraduate students [54,83], which showed that SA was not related to a lower desire to be with others. However, in the study by Brown et al [54], a preference for solitude was found in interactions with unfamiliar people.

Villanueva et al [39] revealed that individuals with SAD presented a higher number of social interactions through their mobile phones than the control group. They were also the group with the least number of social interactions (vs the control group and individuals with depression). In-person interactions (ie, face-to-face) were revealed to be less related to increases in NA and decreases in PA.

In the study by Doorley et al [27], no significant association was found between the medium of communication (ie, digital vs face-to-face communication) and SA. That is, in both media, there was an association between SA and less positive and more negative emotions. Oren-Yagoda and Aderka [33] also explored media of communication in individuals diagnosed with SAD. In this case, the focus was on the media of communication and the associated perceptions and emotions. Individuals with SAD usually preferred to use voice and text media to a greater extent than visual media. However, the authors found that, despite this preference, when visual media were implemented, immediate increases in positive perceptions and emotions were experienced by people with SAD. These results support the idea that the selected medium functions as a safety behavior.

Meanwhile, Russell et al [44] revealed that individuals with SAD presented higher levels of submissive behavior and lower levels of dominant behavior compared to control participants. However, this was true in the presence of anxiety-eliciting cues, which means that there are certain situations that can be perceived as safe environments. In contexts of emotional security, all individuals with SAD presented an enhanced agreeable and deceased quarrelsome behavior, also meaning that there might be situations of security in which individuals with SAD can respond to positive appraisals with enhanced affiliative behavior.

Hur et al [66] found that individuals with SAD benefit more from having close friends, family members, and romantic partners in terms of resulting in lower levels of NA, anxiety, and depression compared to control participants. However, they tend to spend less time with those companions. These results emphasize the intact capacity of individuals with SAD to enhance their mood through social interactions.

In this line, Hannah Lee [65] obtained consistent results, suggesting a tight connection between the levels of SA and the degree of unfamiliarity and judgmentalness in the interactions. Namely, individuals with higher levels of SA presented a stronger association with the 2 processes as a consequence of being more sensitive to experiencing anxiety when facing interactions of the same level of unfamiliarity and judgmentalness compared to people with lower levels of trait SA.

Čolić et al [59] were the first to explore depersonalization and derealization in embarrassing situations. They showed that people with SAD presented more embarrassing social interactions than control participants, and as a result, they also presented more depersonalization and derealization, which can be seen as responses to strong emotions (including embarrassment) as well as attempts to cope with situational challenges.

Cognitive Factors

Social comparison is another important aspect to explore in individuals with SAD given that they usually tend to see themselves with a negative self-image [94,95]. In another study, Brown et al [54] showed that SA was associated with greater self-consciousness, which can also be aligned with the negative self-view that characterizes SAD.

Goodman et al [31] found that SA is related to less favorable and more unstable social comparisons, which can be explained by a negative self-image. Moreover, they demonstrated that, when people with SAD make less favorable social comparisons, they are especially fearful of others’ evaluations.

In a recent study, Brown et al [55] investigated interpersonal distress in heightened SA symptoms as predictors of suicidal ideation. Specifically, this study showed that hurdles to seeking social support and social comparisons mediated suicidal ideation.

Models of SAD have emphasized the central role of fear of evaluation in the appearance and maintenance of this clinical condition. This construct has been included in cognitive models related to attentional biases and negative interpretations of the self [96-98]. Although negative evaluation was considered a core dimension in early models of SAD, fear of positive evaluation has emerged as an important topic in recent years [78,99].

Another study explored anxiety sensitivity cognitive concerns and fear of negative evaluation as 2 potential predictors of SA amplification. Anxiety sensitivity cognitive concerns were shown to uniquely amplify arousal as a consequence of social stress, whereas fear of negative evaluation predicted anxiety fluctuations, indicating that these 2 cognitive constructs may be associated with SA in different ways [79].

By implementing 2 AA studies, Reichenberger et al [78] explored the interaction of both positive and negative evaluation, affect, and stress reactivity. Although the results were not fully in line with the hypotheses, fear of negative evaluation was negatively associated with PA. In addition, the results revealed that the closeness of the relationships was paramount to determine when the interaction was significant, with closer relationships being less anxiety provoking. Consistent with these results, positive and negative feedback seeking has been shown to be higher in individuals with SAD than in healthy participants [81], all of which is aligned with the mounting evidence developed in cross-sectional or laboratory settings. Similarly, Doorley et al [28] demonstrated that self-perceived intense positive events, which are normally reduced in SA, paradoxically provided more psychological benefits (reduced anxiety and motivation toward social situations as well as an increased sense of belonging).

Moreover, Nanamori et al [74] studied triggers of self-focused attention, which is another key component of classic cognitive models of SAD. The results showed that perception of gaze, evaluation, and authority predicted self-focused attention from the observer's perspective, whereas perception of gaze also predicted self-focus on body sensation. Moreover, the perception of positive response and that of a stranger predicted self-focus on body sensation hinged on sex, suggesting that the positive response perception of female participants acted as a predictor of the self-focus on body sensation.

Emerging Topics

Use of AA to Assess Psychological Interventions

Daniel et al [83], Katz et al [41], and Kivity and Huppert [72] implemented AA to explore the course of treatment. In the case of Katz et al [41], PEP, a putative maintenance factor of SA symptoms, was assessed over the course of a cognitive behavioral therapy intervention with a subset of the 60 included participants answering an AA. The intervention yielded significant reductions in both general and momentary PEP, and both types of PEP were significant predictors of SA severity after treatment.

Another study was conducted by Kivity and Huppert [72]. It explored how ER in individuals with high and low SA responded to a practice of cognitive reappraisal using self-report, laboratory tasks, and daily diaries. Although the group with high SA presented lower symptom severity and greater self-efficacy of reappraisal, daily anxiety was not significantly different.

Daniel et al [83] conducted a randomized controlled trial testing whether cognitive bias modification for interpretation could decrease negative interpretation bias. Both the active and control conditions received an AA throughout the treatment period (5 weeks). While the active group also received cognitive bias modification training, the control group only answered the AA. A total of 2 publications were identified from this study, reporting self-report and passive sensing outcome measures. Despite the interesting approach of incorporating multimodal assessment, both analyses yielded nonsignificant results.

Use of Sensors and Biosensors in AA

Over the last few years, new advancements in wearable sensors and biosensors have enabled us to incorporate them into AA studies. In the case of SA, this has led to a considerable body of evidence. Specifically, Bailey et al [51], Boukhechba et al [53], Chow et al [58], Daniel et al [82], Di Matteo et al [60], and Jacobson et al [47] conducted studies using sensors (GPS location and accelerometers) and biosensors (heart rate and heart rate variability).

Bailey et al [51] investigated perseverative cognition in relation to the parasympathetic nervous system, which is considered of utmost relevance in the regulation of stress and emotions [100]. In this study, individuals with both depression and SAD were monitored, and individuals with SAD presented the highest frequency of daily perseverative cognition, which was statistically associated with lower heart rate variability, moderated by negative social interactions. Jacobson and Bhattacharya [67] showed that spending time indoors was associated with anxiety and avoidance symptoms, and this association was significantly higher in individuals with SAD than in those with generalized anxiety disorder.

Boukhechba et al [53] and Chow et al [58] used GPS and Jacobson et al [47] used an accelerometer to demonstrate the capability of passive sensing to predict the severity of SA symptomology according to the level of activity. Moreover, Di Matteo et al [60] designed an app to capture ambient audio, GPS location, screen state, and light sensor data, and this app was shown to be able to identify SAD patterns of behavior in a relatively accurate way.

Exploring Idiographic Comorbidity Patterns

Although a vast array of studies included heterogeneous samples in terms of their diagnosis, only Piccirillo and Rodebaugh [77] had the objective of exploring SAD and major depressive disorder comorbidity, aiming to model person-specific trajectories of cognitive-affective and behavioral dimensions related to these disorders. By including only cisgender women with comorbid SAD and major depressive disorder, this study showed the utmost relevance in disentangling between-person, within-person, and person-specific patterns. For example, loneliness was revealed to be a common predictor of depressive mood and social avoidance at the group level; however, this was not the case when examining the idiographic networks.

Transcultural Differences

Only 1% (1/70) of the studies examined potential variations between different cultural groups. Lee et al [42] explored differences between European Americans and Asian Americans, showing that both groups experienced the same number of anxious events during social situations but Asian Americans presented more negative emotions in those moments.

Discussion

Principal Findings

Our review of 70 original studies using AA to explore SA showed that this methodology provides valuable real-time information on the momentary association of SA with several variables, such as context, affective dynamics, emotional states and regulation, social interactions, and other consequences and antecedents. This comprehensive understanding can contribute to better insights into the appearance and maintenance of symptoms and of this clinical disorder.

Principal Themes Explored

Aligned with the burgeoning literature on AA, affect and emotional dynamics emerged as the most studied topics. These investigations revealed a trend of an increase in NA levels leading to a heightened experience of SA, as well as a growing attention to positive emotions and PA deficits in individuals with SAD. This review also supports the notion that negative emotions and affect are not enough to distinguish normal from pathological SA. As demonstrated by Park and Naragon-Gainey [45], AA may help shed light on the structural models of affect both between and within individuals’ variances, evidence that traditional cross-sectional research or long-term longitudinal research may not capture.

Most interestingly, most of the studies exploring affect trends explored them coupled with ER strategies, consistently extending the vast literature in this regard. Exacerbated NA and PA and dysfunctional strategies to cope with them form a dysfunctional pattern that may be responsible for the appearance and maintenance of SA and SAD [85]. The studies revealed that the interpersonal encounters of individuals with SAD may differ from those of controls in terms of ER use and type of ER strategy. Specifically, both intra- and interpersonal regulatory mechanisms have been shown to be associated with increasing levels of SA. In clinical populations, what was shown to influence the levels of SA was not the use of certain strategies but rather the lack of effectiveness of their use. However, this was not the case in healthy populations. Accordingly, a potential difference between SA symptoms in healthy and clinical populations may lie on the effectiveness or underuse of ER strategies.

In line with the mounting evidence exploring suppression and both experiential and behavioral avoidance in individuals with SAD, this systematic review showed coherent and robust results across the included studies concerning the maladaptive use of these 2 strategies. People with SAD may present an overreliance on the use of suppression and avoidance [85], resulting in a range of negative outcomes such as an increase in NA and a decrease in PA.

In contrast, cognitive reappraisal, a putatively effective strategy, does not yield straightforward results. The problem with individuals with SAD is more the ineffective use of cognitive reappraisal rather than the scarce use of this strategy, although the studies did not seem to coherently yield conclusive results.

Taken together, the results on affective dynamics and ER indicate how appropriate AA can be to study these processes, especially in the case of SA and SAD. AA may be particularly helpful in disentangling between- and within-person effects, which the literature demonstrates can have different or even contrary results, especially in the context of psychological interventions [101].

As a key takeaway message, AA shows that individuals with high SA symptoms or a diagnosis of SAD may not use a narrower repertoire of ER strategies but rather implement that repertoire with less skillfulness or less ability to identify when it is appropriate to implement a certain strategy. In this sense, it is essential to continue exploring the role of polyregulation and flexibility to identify which specific facet of ER contributes as a mechanism of action of SAD.

In that vein, substance use can be seen as a maladaptive behavior that functions as a behavioral strategy to regulate emotions and cope with situations that elicit symptoms of SA. This is particularly recurrent in SA-provoking situations, constituting a reinforcement cycle that operates similarly to other safety behaviors. In particular, alcohol consumption may function as a negative reinforcer, attenuating the negative self-perceived quality of interpersonal encounters and anxiety levels. When this occurs, alcohol use becomes a rapidly established maladaptive behavior with negative consequences.

In addition, maladaptive cognitions, emotional mechanisms, and behaviors were found to be activated when levels of anxiety increased. More specifically, cognitive aspects such as beliefs in capacity, effectiveness in regulating emotions, and the ability to differentiate emotions appear to be relevant in explaining how SA is activated.

As a general takeaway message, there is ample evidence showing the mutual directionality between cognitive and affective or emotional facets and behaviors in the appearance and maintenance of SAD. The several sections in which the studies were categorized constitute just one way of organizing the information. However, many of these studies can be understood as forms of cognition or ER or specific interactional patterns. A clear example is alcohol consumption, which is a behavior that, in the context of SAD, can be understood as an ER strategy.

In that sense, ER is currently a trending topic in AA, but it is important to integrate this increasing amount of knowledge into classic cognitive models. This is pertinent for psychopathological developments in general, and SAD is not an exception. Over the years, the most influential developments in SAD have been cognitive and behavioral models [96,98]. Paradoxically, in this review, cognitive processes remained an underexplored area. An example of this is mental imagery, which has been shown to be a transdiagnostic process that explains the appearance and maintenance of a range of clinical conditions, including SAD [94]; however, there is a dearth of studies on this crucial construct.

Emerging Topics

An additional line of research that needs to be further explored is comorbidity to explore the mutual dependency of certain groups of signs and symptoms. However, given the lack of network analyses, this mutual dependency was not explored in depth. For example, there was only one study exploring suicidal ideation and attempts despite the ample existing literature on AA in suicide research [102] and the strong association between suicide and SA [103]. In the same vein, the relationship among personality, personality pathology, and SA is a relevant topic in contemporary psychopathology that could be further explored using AA strategies. Indeed, there is a growing body of evidence exploring personality and personality pathology dynamics and states, which should be considered in future studies of SA [9].

Another issue worth discussing revolves around the incorporation of AA into psychological interventions. This is an increasingly used practice and may be well integrated with routine outcome monitoring procedures that have been shown to yield significant effects when implemented in both controlled and naturalistic interventions [104]. Routine outcome monitoring is an increasingly implemented strategy that can connect research and practice in unprecedented ways. With that aim, it is necessary to create simple visualization interfaces to feed back the trajectories of certain patient variables. Some efforts have already been made in this direction [105]. If these endeavors are further developed, they can be used by clinicians as a clinical tool, and at the same time, researchers can collect naturalistic data.

Another topic with a lot of potential is the incorporation of behavioral and physiological processes by means of sensors and biosensors. Multimethod measurements that incorporate both passive and active assessments can be of tremendous relevance to harness the affordances of each approach. Together with the development of machine learning algorithms, the proliferation of EMIs is more plausible. This is very important to enhance the personalization of possible treatments.

Regarding data collection, there are now software solutions that are opening up unprecedented opportunities for future research. Older studies usually included PDAs or similar devices, which implies not only spending more resources to implement an AA study but also some degree of training in order for participants to be able to use these devices. Currently, there are studies that harness existing survey platforms such as *Qualtrics* to program either random or fixed prompts without the need for any specifically developed software.

Given that most of the studies were conducted in the United States and the rest were conducted in other Western high-income countries, the results should be generalized to other contexts. Cultural and contextual factors are determinant in all psychopathological conditions, and SAD is not an exception [106]. The proliferation of open-source platforms (eg, m-Path [107]) will permit the dissemination of this methodology to researchers without large budgets, such as researchers from low- and middle-income countries. This is crucial to guarantee that knowledge is not restricted to certain populations, fundamentally populations from Western, educated, industrialized, rich, and democratic countries. In addition, most of the studies (54/70, 77%) paid the participants to enhance the compliance rates, which turned out to be in line with the average compliance in AA literature (approximately 75% [108]). However, this should be considered in future research on AA that seeks to increase the external validity by means of ecological designs.

Methodological Design

Regarding the methodological design of the studies included in this systematic review, there are important aspects to discuss. Most of the studies were well designed; advanced statistical strategies were applied; and, accordingly, most of this research was published in journals with a high impact factor. However, there is a methodological pitfall in AA research that revolves around the lack of psychometrically sound instruments, usually because of trying to reduce participant burden as much as possible. According to Hopwood et al [109], four aspects are essential when discussing the theoretical and methodological implications for AA research: (1) How should time be scaled? (2) How many assessments are needed? (3) How frequently should assessments be conducted? and (4) When should the assessments occur? Researchers using AA methods to conduct research on SA and SAD should carefully consider these questions both theoretically and empirically. Moreover, there is a wider consensus on the need to conduct more theory-driven hypothesis testing [110]. All these methodological aspects should be considered with caution together with the importance of increasing the transparency of reporting the results of AA research [111].

Power analysis was revealed as a weak methodological aspect, with many of the studies not calculating the required sample sizes to anticipate the number of needed participants. Potential limitations concerning the quality of the studies seem to be related to the lack of clear guidelines and standards, which have only recently started to emerge [112].

Regarding the data analysis, most of the studies used multilevel or hierarchical linear models [113]. Indeed, in cases in which ANOVAs or ordinary least squares models were used instead of hierarchical models, the results should be interpreted with more caution. They do not account for the dependency of the data, and in longitudinal assessments such as AA in which data are essentially nested, using these strategies may yield inaccurate pictures of the data [114]. In addition, AAs generally entail mounting random missing data, and multilevel mixed models are appropriate to deal with that data structure.

Future studies need to incorporate new modalities of data analysis that might provide more complex information to understand the dynamics of SA. For example, multilevel network analyses [115] would allow for shedding light not only on the nested structure of the symptoms but also on the interconnectedness at every moment of the individuals' experiences and behaviors. In addition, new-generation time-series analyses such as the time-varying change point autoregressive models would allow for the detection of gradual and abrupt changes in SA markers over the course of the AAs [116]. Furthermore, the recently developed dynamic structural equation modeling method [117] is particularly suitable for intensive longitudinal data from AA. This method allows for a more accurate estimation of individual differences in means and autoregressive effects from AA data. Finally, machine learning strategies will be paramount not only to build predictive models that can better explain SA but also to implement EMIs based on people's needs [118]. In the field of SAD, there is a dearth of studies on EMIs, which is surprising given the ample evidence that has been found using AA.

Limitations

The results of this review should be considered in light of certain limitations. The first limitation concerns the inclusion criteria. Gray literature, including dissertations and preprint depositories, was not considered. Given the growing interest in this topic, we may have missed other relevant studies from these sources. However, this decision had the main aim of ensuring the rigor of including articles that had undergone a peer review process. In addition, we only included revised articles published in English, excluding articles published in different languages.

A second limitation is that this is the first synthesis that summarizes the literature on AA for SA, but further quantitative syntheses (ie, meta-analyses) should be conducted on the specific topics identified in this study. Thus, a qualitative review is a first step that contributes to taking stock of the principal topics studied in the field of SA and AA, but no conclusive statements should be drawn.

Conclusions

This systematic review shows that AA constitutes a very powerful modality to grasp SA from a complementary perspective to laboratory experiments and usual self-report measures. Over the last few years, mounting research has been conducted showing important trends that are shedding light on the understanding of SA and SAD using this ecological tool that is revolutionizing the field.

Acknowledgments

This work was supported by Margarita Salas postdoctoral contracts MGS/2021/37 (UP2021-021) and MGS/2022/18 (UP2022-024) financed by the European Union NextGenerationEU. This study was also funded by the Prometeo Programme grant for Research Groups of Excellence (CIPROM/2021/041—Project "IMPULSA"), Conselleria d'Innovació, Universitats, Ciència i Societat Digital, Generalitat Valenciana. Finally, this work was supported by the Generalitat Valenciana through the Santiago Grisolía predoctoral program (GRISOLIAP/2021/009).

Data Availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Syntax used in database searches.
[DOCX File , 13 KB - mental_v11i1e46593_app1.docx]

Multimedia Appendix 2
Characteristics of the included studies.
[DOCX File , 41 KB - mental_v11i1e46593_app2.docx]

Multimedia Appendix 3
Social anxiety disorder ambulatory assessment research design overview.
[DOCX File , 32 KB - mental_v11i1e46593_app3.docx]

Multimedia Appendix 4
PRISMA Checklist.
[PDF File (Adobe PDF File), 80 KB - mental_v11i1e46593_app4.pdf]

References

1. Gilboa-Schechtman E, Shachar I, Sahar Y. Positivity impairment as a broad-based feature of social anxiety. In: Weeks JW, editor. The Wiley Blackwell Handbook of Social Anxiety Disorder. Hoboken, NJ: John Wiley & Sons; 2014:409-432.
2. Morrison AS, Heimberg RG. Social anxiety and social anxiety disorder. Annu Rev Clin Psychol 2013 Mar 28;9(1):249-274. [doi: 10.1146/annurev-clinpsy-050212-185631] [Medline: 23537485]

3. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008 Apr;4(1):1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
4. Kuppens P, Oravecz Z, Tuerlinckx F. Feelings change: accounting for individual differences in the temporal dynamics of affect. J Pers Soc Psychol 2010 Dec;99(6):1042-1060. [doi: 10.1037/a0020962] [Medline: 20853980]
5. Trull TJ, Ebner-Priemer UW. Using experience sampling methods/ecological momentary assessment (ESM/EMA) in clinical assessment and clinical research: introduction to the special section. Psychol Assess 2009 Dec;21(4):457-462 [FREE Full text] [doi: 10.1037/a0017653] [Medline: 19947780]
6. Colombo D, Suso-Ribera C, Fernández-Álvarez J, Cipresso P. Technology based ecological momentary assessment in clinical psychology. In: Asmundson JG, editor. Comprehensive Clinical Psychology. New York, NY: Elsevier Academic Press; 2022:106-120.
7. Fonseca-Pedrero E, Ródenas-Perea G, Pérez-Albéniz A, Al-Halabí S, Pérez M, Muñiz J. La hora de la evaluación ambulatoria. Pap Psicol 2021;42(2):21-28. [doi: 10.23923/pap.psicol.2983]
8. Palmier-Claus JE, Myin-Germeys I, Barkus E, Bentley L, Udachina A, Delespaul PA, et al. Experience sampling research in individuals with mental illness: reflections and guidance. Acta Psychiatr Scand 2011 Jan;123(1):12-20. [doi: 10.1111/j.1600-0447.2010.01596.x] [Medline: 20712828]
9. Wright AG, Woods WC. Personalized models of psychopathology. Annu Rev Clin Psychol 2020 May 07;16:49-74. [doi: 10.1146/annurev-clinpsy-102419-125032] [Medline: 32070120]
10. Myin-Germeys I, Kasanova Z, Vaessen T, Vachon H, Kirtley O, Viechtbauer W, et al. Experience sampling methodology in mental health research: new insights and technical developments. World Psychiatry 2018 Jun 24;17(2):123-132 [FREE Full text] [doi: 10.1002/wps.20513] [Medline: 29856567]
11. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009 Jul 21;339(jul21 1):b2535 [FREE Full text] [doi: 10.1136/bmj.b2535] [Medline: 19622551]
12. Social anxiety is experienced in the wild: a protocol for a systematic review of studies using ambulatory assessment. Open Science Framework. URL: https://osf.io/47hc8/ [accessed 2024-02-17]
13. Asher M, Asnaani A, Aderka IM. Gender differences in social anxiety disorder: a review. Clin Psychol Rev 2017 Aug;56:1-12. [doi: 10.1016/j.cpr.2017.05.004] [Medline: 28578248]
14. Blalock DV, Kashdan TB, Farmer AS. Trait and daily emotion regulation in social anxiety disorder. Cogn Ther Res 2015 Dec 8;40(3):416-425. [doi: 10.1007/s10608-015-9739-8]
15. Blalock DV, Kashdan TB, McKnight PE. High risk, high reward: daily perceptions of social challenge and performance in social anxiety disorder. J Anxiety Disord 2018 Mar;54:57-64. [doi: 10.1016/j.janxdis.2018.01.006] [Medline: 29421373]
16. Farmer AS, Kashdan TB. Affective and self-esteem instability in the daily lives of people with generalized social anxiety disorder. Clin Psychol Sci 2014 Mar 01;2(2):187-201 [FREE Full text] [doi: 10.1177/2167702613495200] [Medline: 25821659]
17. Goodman FR, Kashdan TB, Stiksma MC, Blalock DV. Personal strivings to understand anxiety disorders: social anxiety as an exemplar. Clin Psychol Sci 2018 Nov 14;7(2):283-301. [doi: 10.1177/2167702618804778]
18. Goodman FR, Kashdan TB, İmamoğlu A. Valuing emotional control in social anxiety disorder: a multimethod study of emotion beliefs and emotion regulation. Emotion 2021 Jun;21(4):842-855. [doi: 10.1037/emo0000750] [Medline: 32191093]
19. Kashdan TB, Farmer AS. Differentiating emotions across contexts: comparing adults with and without social anxiety disorder using random, social interaction, and daily experience sampling. Emotion 2014 Jun;14(3):629-638 [FREE Full text] [doi: 10.1037/a0035796] [Medline: 24512246]
20. Kashdan TB, Farmer AS, Adams LM, Ferssizidis P, McKnight PE, Nezlek JB. Distinguishing healthy adults from people with social anxiety disorder: evidence for the value of experiential avoidance and positive emotions in everyday social interactions. J Abnorm Psychol 2013 Aug;122(3):645-655. [doi: 10.1037/a0032733] [Medline: 23815396]
21. Kashdan TB, Goodman FR, Machell KA, Kleiman EM, Monfort SS, Ciarrochi J, et al. A contextual approach to experiential avoidance and social anxiety: evidence from an experimental interaction and daily interactions of people with social anxiety disorder. Emotion 2014 Aug;14(4):769-781 [FREE Full text] [doi: 10.1037/a0035935] [Medline: 24749634]
22. Daniel KE, Larrazabal MA, Boukhechba M, Barnes LE, Teachman BA. State and trait emotion regulation diversity in social anxiety. Clin Psychol Sci 2023 Sep 28;11(5):894-909 [FREE Full text] [doi: 10.1177/21677026231151956] [Medline: 37981951]
23. Beltzer ML, Ameko MK, Daniel KE, Daros AR, Boukhechba M, Barnes LE, et al. Building an emotion regulation recommender algorithm for socially anxious individuals using contextual bandits. Br J Clin Psychol 2022 Jan 14;61 Suppl 1(S1):51-72. [doi: 10.1111/bjc.12282] [Medline: 33583059]
24. Daniel KE, Baee S, Boukhechba M, Barnes LE, Teachman BA. Do I really feel better? Effectiveness of emotion regulation strategies depends on the measure and social anxiety. Depress Anxiety 2019 Dec 25;36(12):1182-1190. [doi: 10.1002/da.22970] [Medline: 31652383]
25. Daros AR, Daniel KE, Meyer MJ, Chow PI, Barnes LE, Teachman BA. Impact of social anxiety and social context on college students' emotion regulation strategy use: an experience sampling study. Motiv Emot 2019 May 27;43(5):844-855. [doi: 10.1007/s11031-019-09773-x]

26. Ladis I, Toner ER, Daros AR, Daniel KE, Boukhechba M, Chow PI, et al. Assessing emotion polyregulation in daily life: who uses it, when is it used, and how effective is it? Affect Sci 2023 Jun 08;4(2):248-259. [doi: 10.1007/s42761-022-00166-x] [Medline: 37304559]
27. Doorley JD, Volgenau KM, Kelso KC, Kashdan TB, Shackman AJ. Do people with elevated social anxiety respond differently to digital and face-to-face communications? Two daily diary studies with null effects. J Affect Disord 2020 Nov 01;276:859-865 [FREE Full text] [doi: 10.1016/j.jad.2020.07.069] [Medline: 32739703]
28. Doorley JD, Goodman FR, Disabato DJ, Kashdan TB, Weinstein JS, Shackman AJ. The momentary benefits of positive events for individuals with elevated social anxiety. Emotion 2021 Apr;21(3):595-606 [FREE Full text] [doi: 10.1037/emo0000725] [Medline: 31944786]
29. Goodman FR, Kashdan TB. The most important life goals of people with and without social anxiety disorder: focusing on emotional interference and uncovering meaning in life. J Posit Psychol 2021 Nov 11;16(2):272-281 [FREE Full text] [doi: 10.1080/17439760.2019.1689423] [Medline: 34239597]
30. Goodman FR, Rum R, Silva G, Kashdan TB. Are people with social anxiety disorder happier alone? J Anxiety Disord 2021 Dec;84:102474 [FREE Full text] [doi: 10.1016/j.janxdis.2021.102474] [Medline: 34509949]
31. Goodman FR, Kelso KC, Wiernik BM, Kashdan TB. Social comparisons and social anxiety in daily life: an experience-sampling approach. J Abnorm Psychol 2021 Jul;130(5):468-489 [FREE Full text] [doi: 10.1037/abn0000671] [Medline: 34472884]
32. Goodman FR, Daniel KE, Eldesouky L, Brown BA, Kneeland ET. How do people with social anxiety disorder manage daily stressors? Deconstructing emotion regulation flexibility in daily life. J Affect Disord Rep 2021 Dec;6:100210. [doi: 10.1016/j.jadr.2021.100210]
33. Oren-Yagoda R, Aderka IM. The medium is the message: effects of mediums of communication on perceptions and emotions in social anxiety disorder. J Anxiety Disord 2021 Oct;83:102458. [doi: 10.1016/j.janxdis.2021.102458] [Medline: 34343785]
34. Oren-Yagoda R, Paz N, Aderka IM. Pride in social anxiety disorder. J Anxiety Disord 2023 Oct;99:102759. [doi: 10.1016/j.janxdis.2023.102759] [Medline: 37639972]
35. Oren-Yagoda R, Schwartz M, Aderka IM. The grass is always greener: envy in social anxiety disorder. J Anxiety Disord 2021 Aug;82:102445. [doi: 10.1016/j.janxdis.2021.102445] [Medline: 34271331]
36. Oren-Yagoda R, Melamud-Ganani I, Aderka IM. All by myself: loneliness in social anxiety disorder. J Psychopathol Clin Sci 2022 Jan;131(1):4-13. [doi: 10.1037/abn0000705] [Medline: 34843268]
37. O'Toole MS, Jensen MB, Fentz HN, Zachariae R, Hougaard E. Emotion differentiation and emotion regulation in high and low socially anxious individuals: an experience-sampling study. Cogn Ther Res 2014 Mar 13;38(4):428-438. [doi: 10.1007/s10608-014-9611-2]
38. O'Toole MS, Zachariae R, Mennin DS. Social anxiety and emotion regulation flexibility: considering emotion intensity and type as contextual factors. Anxiety Stress Coping 2017 Nov 30;30(6):716-724. [doi: 10.1080/10615806.2017.1346792] [Medline: 28662586]
39. Villanueva J, Meyer AH, Miché M, Wersebe H, Mikoteit T, Hoyer J, et al. Social interaction in major depressive disorder, social phobia, and controls: the importance of affect. J Technol Behav Sci 2019 Dec 10;5(2):139-148. [doi: 10.1007/s41347-019-00121-x]
40. Rinner MT, Meyer AH, Mikoteit T, Hoyer J, Imboden C, Hatzinger M, et al. General or specific? The memory-experience gap for individuals diagnosed with a major depressive disorder or a social phobia diagnosis, and individuals without such diagnoses. Memory 2019 Oct 16;27(9):1194-1203. [doi: 10.1080/09658211.2019.1640252] [Medline: 31311430]
41. Katz DE, Cassin S, Weerasinghe R, Rector NA. Changes in post-event processing during cognitive behavioural therapy for social anxiety disorder: a longitudinal analysis using post-session measurement and experience sampling methodology. J Anxiety Disord 2019 Aug;66:102107. [doi: 10.1016/j.janxdis.2019.102107] [Medline: 31284123]
42. Lee MR, Okazaki S, Yoo HC. Frequency and intensity of social anxiety in Asian Americans and European Americans. Cultur Divers Ethnic Minor Psychol 2006 Apr;12(2):291-305. [doi: 10.1037/1099-9809.12.2.291] [Medline: 16719578]
43. Naragon-Gainey K. Affective models of depression and anxiety: extension to within-person processes in daily life. J Affect Disord 2019 Jan 15;243:241-248. [doi: 10.1016/j.jad.2018.09.061] [Medline: 30248635]
44. Russell JJ, Moskowitz DS, Zuroff DC, Bleau P, Pinard G, Young SN. Anxiety, emotional security and the interpersonal behavior of individuals with social anxiety disorder. Psychol Med 2010 May 12;41(3):545-554. [doi: 10.1017/s0033291710000863]
45. Park J, Naragon-Gainey K. Daily experiences of emotional clarity and their association with internalizing symptoms in naturalistic settings. Emotion 2019 Aug;19(5):764-775. [doi: 10.1037/emo0000482] [Medline: 30148374]
46. Helbig-Lang S, von Auer M, Neubauer K, Murray E, Gerlach AL. Post-event processing in social anxiety disorder after real-life social situations - an ambulatory assessment study. Behav Res Ther 2016 Sep;84:27-34. [doi: 10.1016/j.brat.2016.07.003] [Medline: 27442226]
47. Jacobson NC, Summers B, Wilhelm S. Digital biomarkers of social anxiety severity: digital phenotyping using passive smartphone sensors. J Med Internet Res 2020 May 29;22(5):e16875 [FREE Full text] [doi: 10.2196/16875] [Medline: 32348284]

48. Seah TH, Aurora P, Coifman KG. Emotion differentiation as a protective factor against the behavioral consequences of rumination: a conceptual replication and extension in the context of social anxiety. Behav Ther 2020 Jan;51(1):135-148. [doi: 10.1016/j.beth.2019.05.011] [Medline: 32005331]
49. Arch JJ, Wilcox RR, Ives LT, Sroloff A, Andrews-Hanna JR. Off-task thinking among adults with and without social anxiety disorder: an ecological momentary assessment study. Cogn Emot 2021 Mar 19;35(2):269-281. [doi: 10.1080/02699931.2020.1830751] [Medline: 33076778]
50. Badra M, Schulze L, Becker ES, Vrijsen JN, Renneberg B, Zetsche U. The association between ruminative thinking and negative interpretation bias in social anxiety. Cogn Emot 2017 Sep;31(6):1234-1242. [doi: 10.1080/02699931.2016.1193477] [Medline: 27279528]
51. Bailey T, Shahabi L, Tarvainen M, Shapiro D, Ottaviani C. Moderating effects of the valence of social interaction on the dysfunctional consequences of perseverative cognition: an ecological study in major depression and social anxiety disorder. Anxiety Stress Coping 2019 Mar 22;32(2):179-195 [FREE Full text] [doi: 10.1080/10615806.2019.1570821] [Medline: 30667270]
52. Battista SR, MacKinnon SP, Sherry SB, Barrett SP, MacNevin PD, Stewart SH. Does alcohol reduce social anxiety in daily life? A 22-day experience sampling study. J Soc Clin Psychol 2015 Jun;34(6):508-528. [doi: 10.1521/jscp.2015.34.6.508]
53. Boukhechba M, Chow P, Fua K, Teachman BA, Barnes LE. Predicting social anxiety from global positioning system traces of college students: feasibility study. JMIR Ment Health 2018 Jul 04;5(3):e10101 [FREE Full text] [doi: 10.2196/10101] [Medline: 29973337]
54. Brown LH, Silvia PJ, Myin-Germeys I, Kwapil TR. When the need to belong goes wrong: the expression of social anhedonia and social anxiety in daily life. Psychol Sci 2007 Sep 06;18(9):778-782. [doi: 10.1111/j.1467-9280.2007.01978.x] [Medline: 17760772]
55. Brown BA, Rottenberg J, Goodman FR. Social anxiety and interpersonal risk for suicidal ideation: a longitudinal daily diary analysis. Suicide Life Threat Behav 2023 Dec 28;53(6):968-980. [doi: 10.1111/sltb.12996] [Medline: 37638764]
56. Buckner JD, Crosby RD, Wonderlich SA, Schmidt NB. Social anxiety and cannabis use: an analysis from ecological momentary assessment. J Anxiety Disord 2012 Mar;26(2):297-304 [FREE Full text] [doi: 10.1016/j.janxdis.2011.12.006] [Medline: 22246109]
57. Buckner JD, Langdon KJ, Jeffries ER, Zvolensky MJ. Socially anxious smokers experience greater negative affect and withdrawal during self-quit attempts. Addict Behav 2016 Apr;55:46-49 [FREE Full text] [doi: 10.1016/j.addbeh.2016.01.004] [Medline: 26790139]
58. Chow PI, Fua K, Huang Y, Bonelli W, Xiong H, Barnes LE, et al. Using mobile sensing to test clinical models of depression, social anxiety, state affect, and social isolation among college students. J Med Internet Res 2017 Mar 03;19(3):e62 [FREE Full text] [doi: 10.2196/jmir.6820] [Medline: 28258049]
59. Čolić J, Bassett TR, Latysheva A, Imboden C, Bader K, Hatzinger M, et al. Depersonalization and derealization in embarrassing social interactions: an experience sampling study in social phobia, major depression and controls. J Anxiety Disord 2020 Mar;70:102189. [doi: 10.1016/j.janxdis.2020.102189] [Medline: 32070861]
60. Di Matteo D, Fotinos K, Lokuge S, Mason G, Sternat T, Katzman MA, et al. Automated screening for social anxiety, generalized anxiety, and depression from objective smartphone-collected data: cross-sectional study. J Med Internet Res 2021 Aug 13;23(8):e28918 [FREE Full text] [doi: 10.2196/28918] [Medline: 34397386]
61. Farmer AS, Kashdan TB. Social anxiety and emotion regulation in daily life: spillover effects on positive and negative social events. Cogn Behav Ther 2012;41(2):152-162 [FREE Full text] [doi: 10.1080/16506073.2012.666561] [Medline: 22428662]
62. Geyer EC, Fua KC, Daniel KE, Chow PI, Bonelli W, Huang Y, et al. I did ok, but did I like it? Using ecological momentary assessment to examine perceptions of social interactions associated with severity of social anxiety and depression. Behav Ther 2018 Nov;49(6):866-880. [doi: 10.1016/j.beth.2018.07.009] [Medline: 30316486]
63. Goodman FR, Stiksma MC, Kashdan TB. Social anxiety and the quality of everyday social interactions: the moderating influence of alcohol consumption. Behav Ther 2018 May;49(3):373-387. [doi: 10.1016/j.beth.2017.10.002] [Medline: 29704967]
64. Goodman FR, Brown BA, Silva GM, Bradford DE, Tennen H, Kashdan TB. Motives and consequences of alcohol use in people with social anxiety disorder: a daily diary study. Behav Ther 2022 Jul;53(4):600-613. [doi: 10.1016/j.beth.2022.01.005] [Medline: 35697425]
65. Hannah Lee J. Perceptions towards an interaction partner predict social anxiety: an ecological momentary assessment study. Cogn Emot 2021 Dec 30;35(8):1479-1498. [doi: 10.1080/02699931.2021.1969339] [Medline: 34455927]
66. Hur J, DeYoung KA, Islam S, Anderson AS, Barstead MG, Shackman AJ. Social context and the real-world consequences of social anxiety. Psychol Med 2019 Aug 19;50(12):1989-2000. [doi: 10.1017/s0033291719002022]
67. Jacobson NC, Bhattacharya S. Digital biomarkers of anxiety disorder symptom changes: personalized deep learning models using smartphone sensors accurately predict anxiety symptoms from ecological momentary assessments. Behav Res Ther 2022 Feb;149:104013 [FREE Full text] [doi: 10.1016/j.brat.2021.104013] [Medline: 35030442]

68. Kane L, Ashbaugh AR. Ecological momentary assessment of post-event processing in between two speech tasks: relationships with cognitive and affective factors involved in the maintenance of social anxiety. Behav Res Ther 2022 Dec;159:104208. [doi: 10.1016/j.brat.2022.104208] [Medline: 36252291]
69. Kashdan TB, Steger MF. Expanding the topography of social anxiety. An experience-sampling assessment of positive emotions, positive events, and emotion suppression. Psychol Sci 2006 Feb 06;17(2):120-128. [doi: 10.1111/j.1467-9280.2006.01674.x] [Medline: 16466419]
70. Kashdan TB, Collins RL. Social anxiety and the experience of positive emotion and anger in everyday life: an ecological momentary assessment approach. Anxiety Stress Coping 2010 May;23(3):259-272. [doi: 10.1080/10615800802641950] [Medline: 19326272]
71. Kashdan TB, Adams L, Savostyanova A, Ferssizidis P, McKnight PE, Nezlek JB. Effects of social anxiety and depressive symptoms on the frequency and quality of sexual activity: a daily process approach. Behav Res Ther 2011 May;49(5):352-360. [doi: 10.1016/j.brat.2011.03.004] [Medline: 21463854]
72. Kivity Y, Huppert JD. Does cognitive reappraisal reduce anxiety? A daily diary study of a micro-intervention with individuals with high social anxiety. J Consult Clin Psychol 2016 Mar;84(3):269-283. [doi: 10.1037/ccp0000075] [Medline: 26795939]
73. Kim S, Kwon JH. The impact of negative emotions on drinking among individuals with social anxiety disorder in daily life: the moderating effect of maladaptive emotion regulation strategies. Cogn Ther Res 2019 Oct 09;44(2):345-359. [doi: 10.1007/s10608-019-10045-8]
74. Nanamori M, Tomita N, Kametani C, Matsuda N, Kumano H. Triggers of self-focused attention: an ecological momentary assessment study. Biopsychosoc Med 2023 Apr 23;17(1):16 [FREE Full text] [doi: 10.1186/s13030-023-00273-6] [Medline: 37087445]
75. O'Grady MA, Cullum J, Armeli S, Tennen H. Putting the relationship between social anxiety and alcohol use into context: a daily diary investigation of drinking in response to embarrassing events. J Soc Clin Psychol 2011 Jun;30(6):599-615 [FREE Full text] [doi: 10.1521/jscp.2011.30.6.599] [Medline: 21874086]
76. Papp LM, Kouros CD, Witt HK, Curtin JJ, Blumenstock SM, Barringer A. Real-time momentary mood as a predictor of college students' prescription drug misuse in daily life: direct links and the moderating role of background mental health. Exp Clin Psychopharmacol 2022 Dec;30(6):787-796. [doi: 10.1037/PHA0000500]
77. Piccirillo ML, Rodebaugh TL. Personalized networks of social anxiety disorder and depression and implications for treatment. J Affect Disord 2022 Feb 01;298(Pt A):262-276 [FREE Full text] [doi: 10.1016/j.jad.2021.10.034] [Medline: 34699851]
78. Reichenberger J, Smyth JM, Blechert J. Fear of evaluation unpacked: day-to-day correlates of fear of negative and positive evaluation. Anxiety Stress Coping 2018 Mar 09;31(2):159-174. [doi: 10.1080/10615806.2017.1396826] [Medline: 29119832]
79. Saulnier KG, Saulnier SJ, Allan NP. Cognitive risk factors and the experience of acute anxiety following social stressors: an ecological momentary assessment study. J Anxiety Disord 2022 May;88:102571. [doi: 10.1016/j.janxdis.2022.102571] [Medline: 35487044]
80. Walukevich-Dienst K, Calhoun BH, Fairlie AM, Cadigan JM, Patrick ME, Lee CM. Using substances to cope with social anxiety: associations with use and consequences in daily life. Psychol Addict Behav 2023 Jun;37(4):581-591. [doi: 10.1037/adb0000899] [Medline: 36442020]
81. Wilson GA, Koerner N, Antony MM. An examination of feedback seeking in individuals with social anxiety disorder, generalized anxiety disorder, or no history of mental disorder using a daily diary method. J Cogn Psychother 2018 Apr 01;32(1):15-37. [doi: 10.1891/0889-8391.32.1.15] [Medline: 32746411]
82. Daniel KE, Mendu S, Baglione A, Cai L, Teachman BA, Barnes LE, et al. Cognitive bias modification for threat interpretations: using passive Mobile Sensing to detect intervention effects in daily life. Anxiety Stress Coping 2022 May 31;35(3):298-312 [FREE Full text] [doi: 10.1080/10615806.2021.1959916] [Medline: 34338086]
83. Daniel KE, Daros AR, Beltzer ML, Boukhechba M, Barnes LE, Teachman BA. How anxious are you right now? Using ecological momentary assessment to evaluate the effects of cognitive bias modification for social threat interpretations. Cogn Ther Res 2020 Mar 11;44(3):538-556. [doi: 10.1007/s10608-020-10088-2]
84. Rozen N, Aderka IM. Emotions in social anxiety disorder: a review. J Anxiety Disord 2023 Apr;95:102696. [doi: 10.1016/j.janxdis.2023.102696] [Medline: 36878132]
85. Dryman MT, Heimberg RG. Emotion regulation in social anxiety and depression: a systematic review of expressive suppression and cognitive reappraisal. Clin Psychol Rev 2018 Nov;65:17-42. [doi: 10.1016/j.cpr.2018.07.004] [Medline: 30064053]
86. Ford BQ, Gross JJ, Gruber J. Broadening our field of view: the role of emotion polyregulation. Emot Rev 2019 Jun 23;11(3):197-208. [doi: 10.1177/1754073919850314]
87. Aldao A, Sheppes G, Gross JJ. Emotion regulation flexibility. Cogn Ther Res 2015 Jan 1;39(3):263-278. [doi: 10.1007/s10608-014-9662-4]
88. Tamir M, Mitchell C, Gross JJ. Hedonic and instrumental motives in anger regulation: research report. Psychol Sci 2008 Apr 01;19(4):324-328. [doi: 10.1111/j.1467-9280.2008.02088.x] [Medline: 18399883]
89. De castella K, Goldin P, Jazaieri H, Ziv M, Heimberg RG, Gross JJ. Emotion beliefs in social anxiety disorder: associations with stress, anxiety, and well‐being. Aust J Psychol 2020 Nov 20;66(2):139-148 [FREE Full text] [doi: 10.1111/ajpy.12053]

90. McRae K, Zarolia P. Cognition and emotion in emotion dysregulation. In: Beauchaine TP, Crowell SE, editors. The Oxford Handbook of Emotion Dysregulation. Oxford, UK: Oxford University Press; 2018:39-52.
91. Gohm CL, Clore GL. Individual differences in emotional experience: mapping available scales to processes. Pers Soc Psychol Bull 2016 Jul 02;26(6):679-697. [doi: 10.1177/0146167200268004]
92. Barrett LF, Gross J, Christensen TC, Benvenuto M. Knowing what you're feeling and knowing what to do about it: mapping the relation between emotion differentiation and emotion regulation. Cogn Emot 2001 Nov;15(6):713-724. [doi: 10.1080/02699930143000239]
93. Ottaviani C, Couyoumdjian A. Pros and cons of a wandering mind: a prospective study. Front Psychol 2013;4:524 [FREE Full text] [doi: 10.3389/fpsyg.2013.00524] [Medline: 23966964]
94. McEvoy PM, Erceg-Hurn DM, Barber KC, Dupasquier JR, Moscovitch DA. Transportability of imagery-enhanced CBT for social anxiety disorder. Behav Res Ther 2018 Jul;106:86-94. [doi: 10.1016/j.brat.2018.05.007] [Medline: 29779855]
95. Hirsch CR, Clark DM, Mathews A, Williams R. Self-images play a causal role in social phobia. Behav Res Ther 2003 Aug;41(8):909-921. [doi: 10.1016/s0005-7967(02)00103-1] [Medline: 12880646]
96. Clark DM, Wells A. A cognitive model of social phobia. PsycNET. URL: https://psycnet.apa.org/record/1995-98887-004 [accessed 2024-09-15]
97. Moscovitch DA. What is the core fear in social phobia? A new model to facilitate individualized case conceptualization and treatment. Cogn Behav Pract 2009 May;16(2):123-134. [doi: 10.1016/j.cbpra.2008.04.002]
98. Rapee RM, Heimberg RG. A cognitive-behavioral model of anxiety in social phobia. Behav Res Ther 1997 Aug;35(8):741-756. [doi: 10.1016/s0005-7967(97)00022-3] [Medline: 9256517]
99. Weeks JW, Howell AN. Fear of positive evaluation: the neglected fear domain in social anxiety. In: Weeks JW, editor. The Wiley Blackwell Handbook of Social Anxiety Disorder. Hoboken, NJ: John Wiley & Sons; 2014.
100. Mather M, Thayer JF. How heart rate variability affects emotion regulation brain networks. Curr Opin Behav Sci 2018 Feb;19:98-104 [FREE Full text] [doi: 10.1016/j.cobeha.2017.12.017] [Medline: 29333483]
101. Zilcha-Mano S. Toward personalized psychotherapy: the importance of the trait-like/state-like distinction for understanding therapeutic change. Am Psychol 2021 Apr;76(3):516-528. [doi: 10.1037/amp0000629] [Medline: 32658495]
102. Sedano-Capdevila A, Porras-Segovia A, Bello HJ, Baca-García E, Barrigon ML. Use of ecological momentary assessment to study suicidal thoughts and behavior: a systematic review. Curr Psychiatry Rep 2021 May 18;23(7):41. [doi: 10.1007/s11920-021-01255-7] [Medline: 34003405]
103. Buckner JD, Lemke AW, Jeffries ER, Shah SM. Social anxiety and suicidal ideation: test of the utility of the interpersonal-psychological theory of suicide. J Anxiety Disord 2017 Jan;45:60-63 [FREE Full text] [doi: 10.1016/j.janxdis.2016.11.010] [Medline: 27940416]
104. de Jong K, Conijn JM, Gallagher RA, Reshetnikova AS, Heij M, Lutz MC. Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: a multilevel meta-analysis. Clin Psychol Rev 2021 Apr;85:102002 [FREE Full text] [doi: 10.1016/j.cpr.2021.102002] [Medline: 33721605]
105. Bringmann LF, van der Veen DC, Wichers M, Riese H, Stulp G. ESMvis: a tool for visualizing individual experience sampling method (ESM) data. Qual Life Res 2021 Nov 22;30(11):3179-3188 [FREE Full text] [doi: 10.1007/s11136-020-02701-4] [Medline: 33222049]
106. Soto JA, Dawson-Andoh NA, BeLue R. The relationship between perceived discrimination and generalized anxiety disorder among African Americans, Afro Caribbeans, and non-Hispanic Whites. J Anxiety Disord 2011 Mar;25(2):258-265 [FREE Full text] [doi: 10.1016/j.janxdis.2010.09.011] [Medline: 21041059]
107. Mestdagh M, Verdonck S, Piot M, Niemeijer K, Tuerlinckx F, Kuppens P, et al. m-Path: an easy-to-use and highly tailorable platform for ecological momentary assessment and intervention in behavioral research and clinical practice. PsyArXiv. Preprints posted online January 25, 2022 [FREE Full text] [doi: 10.31234/OSF.IO/UQDFS]
108. Wrzus C, Neubauer AB. Ecological momentary assessment: a meta-analysis on designs, samples, and compliance across research fields. Assessment 2023 Apr 11;30(3):825-846 [FREE Full text] [doi: 10.1177/10731911211067538] [Medline: 35016567]
109. Hopwood CJ, Bleidorn W, Wright AG. Connecting theory to methods in longitudinal research. Perspect Psychol Sci 2022 May 05;17(3):884-894. [doi: 10.1177/17456916211008407] [Medline: 34739347]
110. Kuppens P, Dejonckheere E, Kalokerinos EK, Koval P. Some recommendations on the use of daily life methods in affective science. Affect Sci 2022 Jun 19;3(2):505-515. [doi: 10.1007/s42761-022-00101-0] [Medline: 36046007]
111. Hall M, Scherner PV, Kreidel Y, Rubel JA. A systematic review of momentary assessment designs for mood and anxiety symptoms. Front Psychol 2021 May 17;12:642044 [FREE Full text] [doi: 10.3389/fpsyg.2021.642044] [Medline: 34079492]
112. Trull TJ, Ebner-Priemer UW. Ambulatory assessment in psychopathology research: a review of recommended reporting guidelines and current practices. J Abnorm Psychol 2020 Jan;129(1):56-63. [doi: 10.1037/abn0000473] [Medline: 31868388]
113. Raudenbush SW, Bryk AS. Hierarchical Linear Models: Applications and Data Analysis Methods. 2nd edition. Thousand Oaks, CA: Sage Publications; 2002.
114. Gómez Penedo JM, Muiños R, Hirsch P, Roussos A. La aplicación de modelos lineales jerárquicos para el estudio de la eficacia en psicoterapia. Rev Argent Cienc Comport 2019 Apr 24;11(1):25. [doi: 10.32348/1852.4206.v11.n1.20412]

115. Bringmann LF, Vissers N, Wichers M, Geschwind N, Kuppens P, Peeters F, et al. A network approach to psychopathology: new insights into clinical longitudinal data. PLoS One 2013 Apr 4;8(4):e60188 [FREE Full text] [doi: 10.1371/journal.pone.0060188] [Medline: 23593171]
116. Hehlmann MI, Schwartz B, Lutz T, Gómez Penedo JM, Rubel JA, Lutz W. The use of digitally assessed stress levels to model change processes in CBT - a feasibility study on seven case examples. Front Psychiatry 2021 Mar 9;12:613085 [FREE Full text] [doi: 10.3389/fpsyt.2021.613085] [Medline: 33767638]
117. Hamaker EL, Asparouhov T, Brose A, Schmiedek F, Muthén B. At the frontiers of modeling intensive longitudinal data: dynamic structural equation models for the affective measurements from the COGITO study. Multivariate Behav Res 2018 Apr 06;53(6):820-841. [doi: 10.1080/00273171.2018.1446819] [Medline: 29624092]
118. Delgadillo J. Machine learning: a primer for psychotherapy researchers. Psychother Res 2021 Jan 29;31(1):1-4 [FREE Full text] [doi: 10.1080/10503307.2020.1859638] [Medline: 33375922]

Abbreviations

AA: ambulatory assessment
EMI: ecological momentary intervention
ER: emotion regulation
NA: negative affect
PA: positive affect
PEP: postevent processing
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SA: social anxiety
SAD: social anxiety disorder

Edited by J Torous; submitted 17.02.23; peer-reviewed by K Daniel, T Ranjan; comments to author 03.09.23; revised version received 28.01.24; accepted 07.02.24; published 04.04.24.

Please cite as:
Fernández-Álvarez J, Colombo D, Gómez Penedo JM, Pierantonelli M, Baños RM, Botella C
Studies of Social Anxiety Using Ambulatory Assessment: Systematic Review
JMIR Ment Health 2024;11:e46593
URL: https://mental.jmir.org/2024/1/e46593
doi:10.2196/46593
PMID:38574359

Original Paper

Digital Mental Health for Schizophrenia and Other Severe Mental Illnesses: An International Consensus on Current Challenges and Potential Solutions

Katharine A Smith[1,2,3], MD, PhD; Amy Hardy[4,5], DClinPsy, PhD; Anastasia Vinnikova[6], MSc; Charlotte Blease[7,8], PhD; Lea Milligan[6†]; Diego Hidalgo-Mazzei[9,10,11,12], MD, PhD; Sinéad Lambe[2,13], DClinPsy; Lisa Marzano[14], PhD; Peter J Uhlhaas[15,16], PhD; Edoardo G Ostinelli[1,2,3], MD; Gerard Anmella[9,10,11,12], MD, PhD; Caroline Zangani[1,2,3], MD; Rosario Aronica[17,18], MD; Bridget Dwyer[8], BS; John Torous[8*], MBI, MD; Andrea Cipriani[1,2,3*], MD, PhD

[1]Department of Psychiatry, University of Oxford, Oxford, United Kingdom
[2]Oxford Health NHS Foundation Trust, Warneford Hospital, Oxford, United Kingdom
[3]Oxford Precision Psychiatry Lab, NIHR Oxford Health Biomedical Research Centre, Oxford, United Kingdom
[4]Department of Psychology, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom
[5]South London & Maudsley NHS Foundation Trust, London, United Kingdom
[6]MQ Mental Health Research, London, United Kingdom
[7]Participatory eHealth and Health Data Research Group, Department of Women's and Children's Health, Uppsala University, Uppsala, Sweden
[8]Division of Digital Psychiatry, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, United States
[9]Department of Psychiatry and Psychology, Institute of Neuroscience, Hospital Cli´nic de Barcelona, Barcelona, Spain
[10]Bipolar and Depressive Disorders Unit, Digital Innovation Group, Institut d'Investigacions Biome` diques August Pi i Sunyer (IDIBAPS), Barcelona, Spain
[11]Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM), Instituto de Salud Carlos III, Madrid, Spain
[12]Department of Medicine, School of Medicine and Health Sciences, Institute of Neurosciences, University of Barcelona, Barcelona, Spain
[13]Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom
[14]School of Science and Technology, Middlesex University, London, United Kingdom
[15]Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, United Kingdom
[16]Department of Child and Adolescent Psychiatry, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany
[17]Psychiatry Unit, Department of Neurosciences and Mental Health, Ospedale Maggiore Policlinico Ca' Granda, Fondazione Istituto di Ricovero e Cura a Carattere Scientifico, Milan, Italy
[18]Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy
[†]deceased
[*]these authors contributed equally

Corresponding Author:
Katharine A Smith, MD, PhD
Department of Psychiatry
University of Oxford
Warneford Hospital
Warneford Lane
Oxford, OX3 7JX
United Kingdom
Phone: 44 1865 902135
Email: katharine.smith@psych.ox.ac.uk

Abstract

Background: Digital approaches may be helpful in augmenting care to address unmet mental health needs, particularly for schizophrenia and severe mental illness (SMI).

Objective: An international multidisciplinary group was convened to reach a consensus on the challenges and potential solutions regarding collecting data, delivering treatment, and the ethical challenges in digital mental health approaches for schizophrenia and SMI.

Methods: The consensus development panel method was used, with an in-person meeting of 2 groups: the expert group and the panel. Membership was multidisciplinary including those with lived experience, with equal participation at all stages and coproduction of the consensus outputs and summary. Relevant literature was shared in advance of the meeting, and a systematic search of the recent literature on digital mental health interventions for schizophrenia and psychosis was completed to ensure that the panel was informed before the meeting with the expert group.

Results: Four broad areas of challenge and proposed solutions were identified: (1) user involvement for real coproduction; (2) new approaches to methodology in digital mental health, including agreed standards, data sharing, measuring harms, prevention strategies, and mechanistic research; (3) regulation and funding issues; and (4) implementation in real-world settings (including multidisciplinary collaboration, training, augmenting existing service provision, and social and population-focused approaches). Examples are provided with more detail on human-centered research design, lived experience perspectives, and biomedical ethics in digital mental health approaches for SMI.

Conclusions: The group agreed by consensus on a number of recommendations: (1) a new and improved approach to digital mental health research (with agreed reporting standards, data sharing, and shared protocols), (2) equal emphasis on social and population research as well as biological and psychological approaches, (3) meaningful collaborations across varied disciplines that have previously not worked closely together, (4) increased focus on the business model and product with planning and new funding structures across the whole development pathway, (5) increased focus and reporting on ethical issues and potential harms, and (6) organizational changes to allow for true communication and coproduction with those with lived experience of SMI. This study approach, combining an international expert meeting with patient and public involvement and engagement throughout the process, consensus methodology, discussion, and publication, is a helpful way to identify directions for future research and clinical implementation in rapidly evolving areas and can be combined with measurements of real-world clinical impact over time. Similar initiatives will be helpful in other areas of digital mental health and similarly fast-evolving fields to focus research and organizational change and effect improved real-world clinical implementation.

(JMIR Ment Health 2024;11:e57155) doi:10.2196/57155

KEYWORDS

digital; mental health; severe mental illness; consensus; lived experience; ethics; user-centered design; patient and public involvement; mobile phone

Introduction

Background

Addressing the shortfall in service provision in mental health is a key challenge [1] that has been brought into even sharper focus in the aftermath of the COVID-19 pandemic. The pandemic showed that a shift to digital platforms to deliver synchronous mental health services could be rapidly implemented [2] and can be an acceptable format for many clinicians, patients, and carers [3,4]. However, asynchronous digital approaches (eg, measuring symptoms using digital phenotyping or ecological momentary assessment or providing partially automated therapies using digital platforms) show even more potential to increase capacity and outcomes [5]. These innovations allow patients to undertake a variety of clinically relevant tasks (such as self-monitoring or therapy tasks) outside the in-person clinical encounter. This can be completed independently or with the support of digital navigators or technicians [6] and could reduce the need for specialist clinician support. However, despite their potential, these approaches are mostly still in development and have often proved to be challenging to implement in real-world clinical and community settings [6].

An additional challenge for mental health services is to ensure that the potential benefits of digital approaches are applied in the areas of greatest need, particularly because the pandemic has worsened preexisting disparities in mental health care [7]. Increasing access to care for people with severe mental illnesses (SMIs) such as psychosis and bipolar disorder is a priority worldwide for mental health, particularly in low- and middle-income countries (LMICs) [8]. Although people with SMI experience major health inequalities and have a life expectancy 10 to 25 years shorter than that of the general population [9,10], less research on digital interventions has been conducted in this patient group compared with other conditions [11]. Assumptions that people with SMI will not be able or willing to engage in technology-augmented assessment and treatment have combined with the known element of digital exclusion due to the intersection of SMI and socioeconomic inequities, resulting in a shortfall of development in this area [12,13].

Internet access among people with SMI is increasing [14,15], and ownership of smartphones is now more common than ownership of computers [16]. However, there are still barriers to digital access and digital literacy in this group [17,18]. Even with a device, many people with SMI have insufficient economic resources to maintain consistent access or may lack the confidence or skills to use it to its full ability [12,18,19]. Although there are examples of training programs in digital skills and confidence for people with SMI [20,21] and evidence that they are willing and able to engage effectively with digital mental health [22-25], this is often a neglected area [11].

Study Objectives

In this study, we used a consensus method to identify the challenges and potential solutions in the development and implementation of digital mental health in the care pathways of people with schizophrenia and other SMIs, and consider how new research could help fill the gap between need and service provision to improve patients' outcomes. To explore this, we convened an expert international multidisciplinary group to focus specifically on the complexities of collecting data, delivering treatment, and the ethical challenges in this area. We focused on SMI as a group where there is arguably the greatest need. While SMI is a broad term covering mental illnesses causing serious functional impairment, we focused primarily on evidence and examples of digital mental health interventions for schizophrenia and psychosis, and related risks such as suicide.

Methods

Overview

We used the consensus development panel (or consensus development conference [CDC]) approach [26,27] and followed the methodology described and used by the US National Institutes of Health [28] and the World Health Organization [26,29]. The CDCs were developed by the National Institutes of Health [26] and we chose this as it is a particularly effective consensus method for identifying broad areas of challenge and potential solutions. This is in contrast to alternative consensus approaches (such as the Delphi or nominal group techniques) that aim to achieve specific criteria or protocols [30]. Therefore, the CDC is a particularly relevant method for a rapidly developing area such as digital mental health [6]. In addition, it enables a multidisciplinary approach and can moderate potential bias from a group of individual experts using several strategies, such as the inclusion of a separate panel of nonexpert participants (hereafter, "the panel").

Central to the methodology of the CDC is a face-to-face meeting between the expert group and the panel involving an interactive method to develop a consensus. Panel members are provided with evidence by the expert group. The panel members ask questions to clarify and then deliberate on the issue directed by their chairperson in the process to reach a consensus [26].

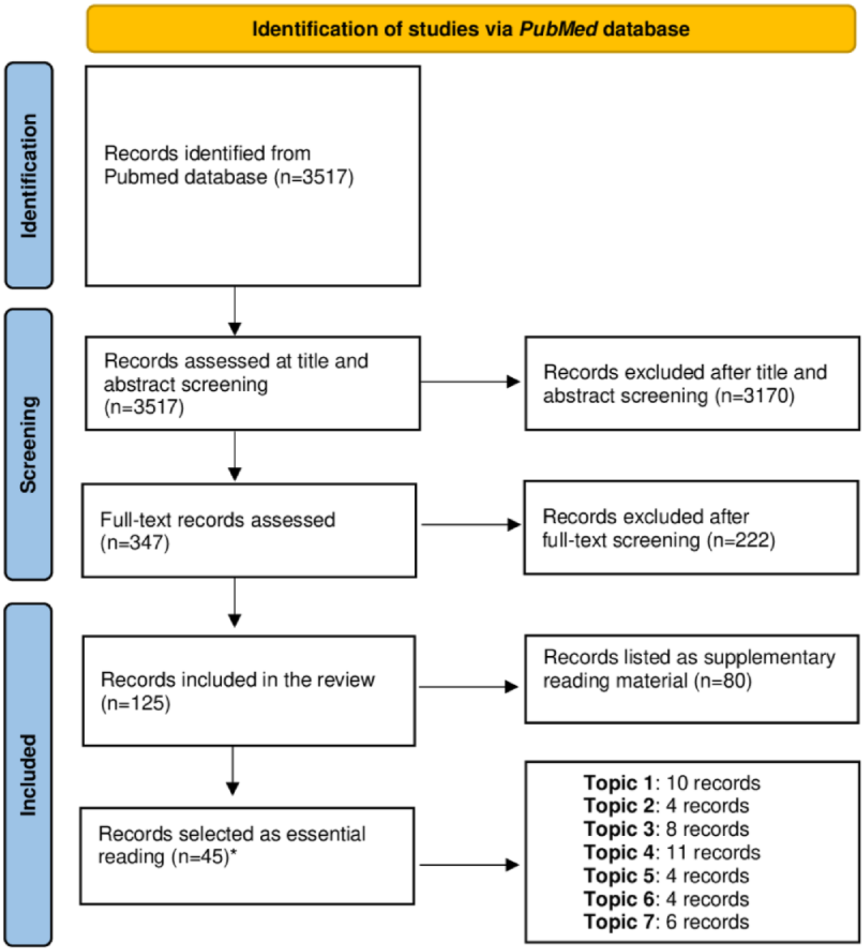
In this study, the panel also increased their knowledge of the field in advance of the meeting by conducting a literature review using PubMed to search for terms relevant to the main themes identified by the experts (see Textbox 1 and Figure 1 for details). This preliminary work identified the areas of recent development, uncertainties, or challenges that formed the agenda for the questions to be addressed in the face-to-face meeting.

Textbox 1. Literature review.

Methods

- The panel members conducted a literature review of the records published over the last 5 years (from January 1, 2018) on PubMed, and searched for articles relevant to the themes identified by the expert group in the area of digital technologies and severe mental illness (SMI; focusing for this review on schizophrenia and psychosis).
- The seven identified topics suggested by the experts were (1) digital markers and personalized interventions, (2) patient and public involvement and engagement perspectives on SMI, (3) digital technologies and SMI, (4) virtual reality approaches for SMI, (5) technology and mobile health for suicide prevention, (6) digital approaches to empathy in clinician-patient interactions, and (7) the use of web-based screening to detect emerging psychosis.
- The panel used a broad search strategy for papers relevant to schizophrenia, psychosis, and digital mental health. The search was conducted on September 17, 2023, using the following search strategy: *("Schizophrenia Spectrum and Other Psychotic Disorders" [MeSH Terms] OR ("schizophreni*"[Title/Abstract] OR "psychos*"[Title/Abstract] OR "psychotic"[Title/Abstract])) AND ("digital*"[Title/Abstract] OR "smartphone*"[Title/Abstract] OR "mobile*"[Title/Abstract] OR "virtual*"[Title/Abstract] OR "internet*"[Title/Abstract]) AND 2018/01/01:3000/12/31[Date – Publication].*
- The screening process was completed independently by the 7 panelists and any queries were resolved through team discussion. Papers were eligible for inclusion if they were relevant to schizophrenia or psychosis and digital mental health. The search resulted in 3517 records, 3170 (90.13%) of which were excluded after title and abstract screening.
- At the full-text screening, 125 records met the inclusion criteria. Of these 125 records, 45 (36%) were further selected as essential reading by the panel and shared with all the panelists, including a selection of up to 5 articles suggested by each member of the expert group. The remaining 64% (80/125) of the records were listed as supplementary reading material for consultation purposes (see the PRISMA [Preferred Reporting Items for Systematic Reviews and Meta-Analyses] flow diagram [31] in Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart. *Two of the records were each allocated to two topics as they were relevant to both.



The Consensus Meeting

The meeting was held in Rome over 2 days in October 2023. The meeting was facilitated by JT and AC who led the expert group, and by the chair of the panel (KAS) who also recorded and summarized the meeting. Each expert gave a brief presentation, including shared slides, methodology, analysis of data, and relevant citations, followed by a whole-group discussion led by a different member of the panel for each talk. At the end of each day a summary was prepared in discussion with the whole group, and the structure of the consensus (challenges and potential solutions) was agreed upon at the end of the second day. Consensus was defined as either fully met or not met at all, with the outcome being transparently reported [32,33]. At the end of the meeting, the whole group engaged in plenary discussion to identify the key themes and structure the results.

The Expert Group

The 9 experts (AC, AH, CB, DH, SL, L Milligan, L Marzano, JT, and PU) encompassed expertise in a variety of specialist areas within SMI and digital mental health (including virtual reality, coproduction and co-design, suicide prevention, web-based screening and early intervention, digital approaches to empathy and the therapeutic relationship, ethical issues, and lived experience). The group composition was gender-balanced, and professional backgrounds included psychiatry, psychology, methodology, evidence synthesis, patient and public involvement (PPI), and ethics. The expert group was international (including members from Germany, Italy, Ireland, Spain, Sweden, the United Kingdom, and the United States). In advance of the meeting the experts were asked to provide information on their area of expertise as a short abstract supported by up to 5 references that they considered to be key in the last 2 to 3 years, and this was circulated to the expert group and panel. This was supplemented by a systematic search of the literature conducted by the panel (Textbox 1).

The Panel

The panel was composed of 7 members (GA, RA, BD, EGO, KAS, AV, and CZ), including early-career and more experienced clinicians and researchers (at a different level of expertise) and an individual with lived experience. The panel members were chosen because they were well informed or experienced in the field of SMI but had no particular expertise in any one area of digital psychiatry. The panel was also international (including members from Italy, Paraguay, Spain, the United Kingdom, and the United States).

In preparation for the meeting, the panel members summarized the best available recent evidence on the 7 topics to be discussed. This was achieved by completing a systematic search of the recent literature in the areas to be covered by the meeting (see Textbox 1 and Figure 1 for details). All panelists were asked to read the selected papers, and each panel member led the group discussion on one expert talk to facilitate equal contributions from members of the expert group and the panel.

Reflexivity Statement

The meeting was convened by JT and AC, who selected the expert group to represent a balance of professional backgrounds, areas of specialist digital mental health expertise, lived experience, and gender. Panel members were suggested by members of the expert group and through professional contacts. JT and AC were assisted by KAS in the organization and preparation of the meeting. The logistics of the meeting were supported externally by Angelini Pharma, but they did not have any input in the design of the meeting, identification or selection of the expert group or panel, agenda of the meeting, discussions, consensus, or output. We acknowledge that the shared knowledge and experiences of the expert group and panel may have had an impact on the interpretation of the data.

Ethical Considerations

Ethical approval was not sought for this study as it did not involve research on human participants. The consensus was conference based, and all attendees offered contributions to the research topic in an open environment where talks were voluntary. No personal details were solicited or reported, and all presentations were based on expertise, including the lived experience expert coinvestigator who spoke in her capacity as an expert in how patients with SMI navigate health systems.

Results

Overview

The group achieved full consensus on all the outputs of the meeting, identified 4 broad themes, and divided each consensus theme into the current challenges and potential solutions, with examples from existing projects for SMI. A summary of the results is presented in Textbox 2.

Textbox 2. Summary of consensus themes, challenges, and potential solutions.

Theme 1: user involvement for real coproduction

Challenges:

- User engagement:
  - Most apps are discarded within days
  - There is a lack of consensus or standards for assessing and monitoring engagement and attrition
  - Most of the research and development of digital tools focuses on conditions outside severe mental illness and is conducted and published in higher-income countries (not low- and middle-income countries)
  - Interventions need to be at a younger age and across the life span

Potential solutions:

- Patient and public involvement and active engagement in human-centered design:
  - True representation of diverse populations
  - Coproduction
  - Patient and public involvement and engagement which is integral within the study
- Personalized approaches
- Early interventions

Theme 2: research methodology

Challenges:

- There is a lack of validated and agreed standards for digital biomarkers or behavioral data

Potential solutions:

- Agreed standards for digital mental health research
- Transparent data processing pipelines and data sharing
- Measuring harms
- A focus on prevention and longer-term strategies
- Mechanistic research in digital mental health interventions

Theme 3: regulation, governance, and funding

Challenges:

- Achieving sustainable funding
- Unstable regulation

Potential solutions:

- Sustainability planning in funding calls and applications
- Governance change
- Guidelines for best practice and research standards

Theme 4: implementation in real-world clinical settings

Challenges:

- Translating research findings into the clinical setting

Potential solutions:

- Collaboration between multidisciplinary groups
- Training and assessment in clinics

Theme 1: User Involvement for Real Coproduction

Current Challenges

User Engagement

User engagement is a key issue as it represents a potential mechanism of change for improving clinical outcomes in digital health interventions [34,35]. Even in the general population, most apps are discarded within days of first use [36], and recently some have labeled engagement as the "Achilles Heel" of digital therapeutics for mental health, with attrition also of concern [37]. An additional complication is that engagement and adherence are often conflated; participants may disengage for multiple reasons (including the attainment of their goals rather than lack of adherence). The concept of optimal use may be a more representative indicator but needs to be defined a priori in research studies [38]. However, there is also a lack of consensus or standards for assessing and monitoring engagement with and attrition from digital tools, limiting progress in the field [39]. In addition, most of the research on and development of digital tools focuses on conditions outside SMI [11]. Digital technologies, particularly smartphone apps, have the potential to be novel tools for managing SMIs, especially in LMICs [40]. However, most research studies in this area are conducted and published in higher-income countries, and samples may not be representative of the wider SMI population [41]. Individual reports from researchers in LMICs suggest that engagement is also challenging in these settings [42].

Interventions at a Younger Age and Across the Life Span

Most mental disorders begin during adolescence [43] and are frequently preceded by subthreshold symptoms, suggesting that this period is critical for early intervention and diagnosis [44]. While this age group may be particularly suited to digital interventions (eg, 46% of individuals aged 13-17 years in the United States reported "almost constant" use of social networking services [45]), effective digital interventions may need to be targeted earlier than in previous studies and tailored to the needs of young people [46,47]. Youth is not the only stage of the life span in which focus is needed. In general, older people are most likely to be digitally excluded, and enduring or treatment-resistant psychosis has received less focus, funding, and innovation for digital approaches [18].

Potential Solutions

PPI and Engagement in Human-Centered Design

A potential solution to the challenges of user engagement with digital tools is to actively involve patients and participants in the design and development of digital tools and all stages of the research process (see Textbox 3 for further information from a lived experience perspective).

Not only does this active engagement and involvement result in research that is more relevant and useful for users, but it also meets the essential rights of users to be included in the development of interventions that will affect their lives and those closest to them [49]. However, even when users have been included, there is often a lack of true representation of diverse populations [50], resulting in digital solutions that do not provide sufficient options for different individual preferences and life circumstances. PPI and engagement (PPIE) within digital development and research studies can be highly variable, and although the aim is coproduction, for true design *with* and not *for* participants, PPIE needs to be integral within the structures, policies, and processes of the study team so that power is shared [51-54]. This will only occur by building on the existing methods for public engagement using innovative new models encompassing the breadth of patient voices alongside industry, regulators, and academics [55,56].

While "ticking the box" of PPIE is relatively easy, PPIE that is truly representative of the target population and integrated into all phases of the design and research process is much more challenging and therefore costly. However, there are examples of integration of a wider range of patient populations throughout the design, prototype, and evaluation phases (eg, SlowMo and gameChange [see Textbox 4 for details] and mindLAMP [57]) and examples of global applications across LMICs as well as higher-income countries [58]. Textbox 4 provides some more details on 2 examples.

Consideration of the specific characteristics of involved PPIE members is also important. In the move toward an improved user-centered PPIE involvement, researchers should consider involving a broader range of patient experiences. PPIE involvement in product and study development often comes from "expert" PPIE members. While this is helpful ("expert" PPIE members have often had experience with other interventions and can contribute specialist knowledge), input from diverse experiences enhances the breadth of perspectives and understanding throughout the development phases. A risk inherent in participatory design is that the voices of seldom-heard groups are neglected, and so purposive recruitment of PPIE to ensure representation is recommended. This addresses potential hurdles such as digital exclusion (ie, in terms of access and skills) and the monitoring of the negative and positive effects of proposed design solutions for a range of stakeholders [6,41].

Textbox 3. Lived experience perspectives.

Overview

- The inclusion of an individual with lived experience as part of the multidisciplinary panel approach was an important way to ensure that the patient voice was heard and included during the meeting and interwoven throughout the discussions. It was also an opportunity to provide an alternative perspective beyond the parameters of academic research and study.
- The user perspective shared in the meeting itself drew upon a number of key challenges in the current model of mental health care (in the United Kingdom), namely, (1) respect for patient autonomy and voice, (2) sensitivity to cultural barriers, and (3) accessibility of interventions and their adaptability to patients' lives. While this is the experience of only one individual, it does raise important considerations and can inform 3 key user perspective takeaways from the meeting to guide further work in the study of digital mental health for severe mental illness (SMI).

Patient autonomy and voice

- Respect for patient autonomy and listening to the voices of patients are crucially important in the research and development of digital mental health interventions, just as they are in the delivery of care. The lived experience shared during the meeting indicated that there was a feeling of a lack of control and respect for patient decision-making capacity in the provision of care. There was a sentiment expressed that, by stressing the importance of the patient's voice in the research and production stage of mental health interventions, this may subsequently set the tone for real-world application.
- As discussed in this paper, there is already an identified greater need for genuine coproduction to facilitate this, with coproduction being defined as an equal weight placed on the involvement and accountability of those with lived experience and of academics and experts [48]. A challenge in coproduction continues in current methodologies, language, and ways of working in academia that are not always easily translated or transferred to those outside this field. There needs to be more to ensure that the platform for shared work is an equal one, with opportunities for both parties to learn from one another and ensure that coproduction is suitably accessible for genuine equal input in knowledge production. During the meeting, it was considered from a user perspective that, when principles of coproduction in research are fully realized, this may improve patient uptake and also strengthen patient autonomy. With respect to the latter, this can arise because greater insight into the needs of patients among digital health designers and clinician researchers should be much clearer as a result of patient input as well as shifting standards in mental health care more broadly.

Cultural barriers

- Related to the aforementioned, there is a need for understanding regarding cultural barriers and stigma that act as a deterrent to access to mental health interventions and care for different users. There is a plethora of reasons why culture can be a barrier, from traditional roles and responsibilities of an individual that draw on available time, to education on mental illness (or a lack thereof) and particularly negative stigmatization. The user perspective shared during the meeting indicated that many currently available interventions do not offer sufficient nuance or flexibility in recognizing the challenges to access faced by many communities. This is linked to existing literature that often acknowledges the limitation of research being undertaken with a focus on Western-centric, middle-class, digitally literate populations despite high proportions of global mental illness being found in low- and middle-income countries. Both in the early stages of research and development and in the deployment of existing interventions, it is necessary to reflect on which communities have the greatest needs to be heard and understood.

Accessibility and adaptability

- For a service user or individual with lived experience of SMI, a best-case scenario is one in which all barriers to accessing an intervention have been removed. An important barrier for many will be flexibility of access. Converging both the points on patient autonomy and overcoming cultural barriers, increased flexibility in access to interventions takes away a number of crucial obstacles. It is here that digital mental health has an opportunity to have a significant impact. Whether owing to the need for discretion, the weight of other time-consuming responsibilities (particularly in the case of high-functioning individuals with SMI, as shared during the discussion), or cultural stigma, the opportunity for adaptability not only serves a practical purpose but also provides a notion of empowerment and a sense of control for the individual.

Textbox 4. Examples of human-centered design in the development and evaluation of digital therapeutics for psychosis.

SlowMo: integrating an interventionist causal approach and inclusive human-centered design to develop a next-generation cognitive behavioral therapy for psychosis (CBTp)

- Background
  - An evidence-based causal reasoning mechanism in paranoia (“fast thinking”) has an antidote in slow thinking (ie, belief flexibility) to improve paranoia and promote living well [59].
  - Proof of concept randomized experiments, and a feasibility randomized controlled trial (RCT) showed CBTp targets fast thinking and promotes slow thinking and improved paranoia [60-63].
- Patient and public involvement and engagement (PPIE) and an inclusive, human-centered design
  - Problem: the need to improve access, experience, and outcomes for the National Institute for Health and Care Excellence (NICE) recommended CBTp, particularly among marginalized groups [64].
  - Method: UK Design Council’s Double Diamond, using ethnographic methods to define the design problem and iteratively co-designing solutions and testing prototypes with purposive sampling of users (n=18). Interdisciplinary collaboration with the Helen Hamlyn Centre for Design, the Royal College of Art, King’s College, London, software developers, and National Health Service (NHS) Trusts.
  - Solution: a redesigned version of the therapy, SlowMo, tailored to users’ needs which supports self-monitoring, provides accessible and memorable information, is enjoyable and trustworthy, promotes personalization, and provides flexible interpersonal support.
    - A web application supports the delivery of sessions with a therapist, which is synchronized with a native mobile app for use in daily life, addressing access and data protection concerns.
    - Responsive, touch screen technology supports personalization and visualization of thoughts and thinking habits.
    - Audiovisual lived experience stories provide engaging interpersonal support.
- Multisite RCT and process evaluation (SlowMo1 [25]; N=361)
  - This was the first digitally supported therapy for paranoia to demonstrate efficacy (improved effect size compared to conventional CBTp was achieved in half of the minimum number of recommended sessions [65]), and the mechanism of change. A process evaluation of therapy experience demonstrated high rates of therapy uptake and adherence [66], a coproduced qualitative interview study supported the user experience and mechanism of change [67], and a user experience study showed that SlowMo bridged the “digital divide” as poorer digital literacy in Black people and older people did not translate to the user experience of SlowMo [19].
  - A PPIE evaluation indicated valued outcomes in the RCT [68].
- Future research includes implementation, effectiveness, and cost-effectiveness study (SlowMo2; ongoing)
  - Interdisciplinary co-design of software for implementation incorporating lived experience findings from the RCT.
  - Lived experience coapplicant and leadership.
  - Substantive posts for lived experienced researchers and purposive recruitment of a representative lived experience advisory panel.
  - The NICE early value assessment has recommended SlowMo for use in the NHS for the treatment of paranoia in adults with psychosis while more evidence is generated [69].

gameChange: automated virtual reality (VR) cognitive therapy for treating agoraphobic avoidance and distress in patients with psychosis

- Target
  - Agoraphobic avoidance in psychosis
- Defining the clinical problem
  - A survey of 1809 people with psychosis [70] showed two-thirds had levels of anxious avoidance comparable to agoraphobia.
- Steps of user involvement
  - Aimed to create a VR therapy to help people with psychosis feel safer, more confident, and in control in everyday situations.
  - The design brief: 3 hours of novel VR experiences with graded levels of difficulty. Automated through a virtual coach, users would be guided through 6 scenarios and given opportunities to drop their defenses and test their fear beliefs.
  - A person-centered design approach was used involving people with psychosis at each stage of development, including choosing scenarios, selecting tasks, testing prototypes, and user testing [71].
  - In total, >100 people with psychosis provided ≥500 hours of input.
  - gameChange is Conformité Européenne (CE) marked as a Class I active medical device.

- Clinical testing and implementation
  - Clinical testing in a multisite RCT [72].
  - Health economics evaluation was embedded in the clinical trial [73].
  - gameChange is approved for use in the NHS while more evidence is generated to treat severe agoraphobic avoidance in people with psychosis aged ≥16 years (with the support of a mental health professional) [74].

Personalized Approaches

Facilitating personalization of digital approaches increases their ability to meet the diverse needs of people in real-world clinical populations [10]. Incorporating patient preferences in the broad development of tools is essential, but integrating patient preferences into the operation of the digital tool also enables genuine personalization so that the tool is closely matched to the individual requirements and preferences of the user [75]. A concrete step toward personalized approaches can include increasing the pool of data that informs the use of an intervention. One route to this increased body of data would be to facilitate data sharing from all studies in a specific area (see also the *Theme 2: Research Methodology* section). While there are specific ethical challenges in data sharing, particularly in digital mental health studies where sensitive personal data are often collected, there are well-defined principles to guide best practice [76]. Despite this, digital health data sharing remains less common, although there is a clear recognition of the need for this [76].

Early Interventions

Early intervention for psychosis has the potential to identify and improve the outcome for individuals who meet clinical high-risk state for psychosis (CHR-P) or first-episode psychosis (FEP) criteria, but most individuals with CHR-P who later develop psychosis are not currently detected during the prodromal phase [77]. Most individuals with FEP use web-based resources, and 76% responded favorably to the possibility of receiving web-based mental health support [78] and so could be identified through web-based approaches [79]. However, such tools need to be appropriate and appealing for this age group—social media and low-threshold entry points may be useful to extend the early intervention approach outside established care pathways [80]. In addition, evidence-based digital tools and resources are needed to guide parents, carers, and the young person's wider support networks (such as teachers or social workers). Young people also need to be actively involved in coproduction to ensure the highest-quality equitable outcomes [81].

Theme 2: Research Methodology

Current Challenges

A key challenge for all areas of digital mental health is the lack of validated and internationally agreed standards for digital biomarkers or objective behavioral data obtained from patients' personal devices [6,82]. This means that studies vary in how they assess digital biomarkers and interventions, and in what and how they measure change, engagement, adherence, improvement, or potential harms. For instance, with some exceptions, studies have not examined adverse events [38]. This is needed for safety measurements in digital interventions not only among people with SMI but also in health care more broadly. In 2023, regulatory bodies such as the US Food and Drug Administration (FDA) issued draft guidance outlining how verification and validation of digital health technologies should be approached [83], including some challenges that are more prominent in digital health, such as data privacy and confidentiality. However, a 2023 review of FDA approvals across all domains of health care, including digital health technologies, suggested a lack of scientific rigor across studies and the need for higher-quality research [84]. Because there is significant variability between studies, current research is not easy to replicate or validate, leading to reduced confidence in the results and the robustness of the evidence base supporting digital interventions. In addition, digital mental health research in recent years has focused mainly on reactive interventions that address immediate or short-term needs, whereas longer-term and preventative approaches or time series–aware methods often remain lacking [85].

Potential Solutions

Agreed Standards for Digital Mental Health Research

Differences regarding study standards and data quality need to be addressed at an international level with agreed standards for studies and their subsequent publication. The homogenization of standards could mitigate fragmented approaches, and regulatory agencies and funders could insist on such standards [86]. This could prompt researchers to align research protocols in digital mental health so that there is transparent reporting of the data collected and the frequency with which they are collected (which would allow for comparison and combination of data sets). This would also facilitate the understanding of the relative usefulness of, for example, particular machine learning paradigms (as these are most accurate on data sets that are similar to those they have been trained on), and this would require separate data processing standards [87,88].

Data Sharing

To enable the comparison and combination of data sets, standards should require the sharing of data via open-source documentation (eg, the study by Bent et al [89]). While this can generate some potential ethical challenges (Textbox 5), there are already international examples demonstrating data sharing [90,91]. The older 2016 CrossCheck study of people with schizophrenia using a smartphone app to collect digital phenotyping data also offers an open data set that has enabled numerous publications advancing computational methods of symptom and relapse prediction [92].

Textbox 5. Applying a biomedical ethics framework in the use of digital tools for severe mental illness (SMI).

Respect for autonomy

- Digital phenotyping involves continuous, real-time multimodal streaming of data from smartphones and other internet-enabled devices. As noted, the insights offered by these data may be highly sensitive, uncovering, for example, an internet browsing history, purchasing patterns, or levels of social contact. Conceivably, such data may also create legal exposures. For example, an estimated 1 in 4 people with SMI also experience substance use disorders [93], which, in some circumstances, could heighten risks of criminal behavior or illicit drug purchasing or use. For clinicians to access these digital data streams to yield benefits for patients, health care users must provide fully informed consent regarding the whys, hows, and consequences of sharing digital information, including which data may yield the most valuable insights and how health care providers will store and use this information.

Beneficence

- Digital tools have the potential to offer faster, less expensive, and more accessible care at scale. Many digital interventions, such as apps, also afford patients with SMI opportunities to engage with technology without the fear or risks of stigmatization that may arise in clinic visits. While some tools already offer considerable promise, for the benefits of these applications to be optimized, greater research is needed to explore implementation to increase uptake (themes 1, 2, and 4).

Harms

- Clinicians are duty bound to “first, do no harm,” yet, currently, there is a deficit of research into the potential adverse effects of digital health tools for SMI (theme 2). Identifying when tools lead to incidents of harm, including self-harm or discontinuation of treatment is critical to ensure safety. For example, preoccupation with checking health-tracking data via downloadable apps and wearable devices might increase anxiety, especially among some subpopulations of patients with SMI.

Justice

- Artificial intelligence–powered digital tools rely on patient involvement, and if data sets are not representative of the populations in which they will be used, these technologies may not be as useful or could be harmful for these groups. Furthermore, to benefit all patients, digital tools need to be accessible to everyone [94]. While digital divides are narrowing, the most vulnerable patients—including those with SMI—are often more likely to live with lower incomes, meaning that they are still less likely to own digital devices, to have access to broadband, or have acquired the digital literacy skills necessary to partake in technology use and reap the benefits [95]. In many countries, advancing digital health research is now a priority [96]; however, without concerted efforts to improve the distribution and access to these tools, inequities will persist or potentially increase. Aimed at improving digital literacy among patients, including those with SMI, the Digital Opportunities for Outcomes in Recovery Services program has been deployed in many community settings, including in the United Kingdom and the United States [20].

Professional-patient relationships

- Used effectively, digital tools are unlikely to replace human relationships, including in care settings. Moreover, for data to be effectively interpreted and understood, a deeper and more honest and trusting partnership between patients and clinicians is imperative. For example, the insights gleaned from digital phenotyping are fallible and require context—only patients with lived experience can assist in offering the situational knowledge needed to furnish a deeper understanding of what the data show and how they might be harnessed in preventative care [94]. Conversely, as noted previously, owing to the nature of data gathering, there are also multiple new opportunities for these technologies to undermine trust in the fiduciary clinician-patient relationship (see the previous points). Regulatory policies, civic debate and patient involvement, and health professional ethical awareness must strive to keep abreast of advances (themes 1 and 3).

Measuring Harms

For transparency, standards should also require that potential adverse events or harms, as well as benefits, are identified and measured (eg, the International Collaboration for Harmonising Adverse Events Reporting in Technology for Schizophrenia (iCHARTS) network by Bucci et al [97]). In addition, preregistration of protocols [98] should also be implemented to increase the publication of “negative” findings. This would balance the known bias toward the publication of “positive results”—a trend already noted in biomarker research for bipolar disorder [99]. Such approaches could better inform researchers about when to build interventions using existing platforms and when novel platforms are needed. Harm need not be limited to classic symptom exacerbation and can also include loss of privacy, inequality or discrimination, social or personal loss, and even physical injury. These could be measured using a combination of both self-report and investigator-assessed scales.

Prevention and Longer-Term Strategies

Digital mental health research in recent years has focused mainly on reactive interventions in the short term, but longer-term or preventative interventions may be equally or more important, particularly in areas such as SMI and suicide prevention. Examples include the use of web-based platforms in suicide prevention while tackling harmful content that could promote or encourage suicide and self-harm [100-103], and web-based screening in youth mental health to support the identification of high-risk individuals or groups.

Mechanistic Research in Digital Mental Health

Research on the mechanisms of digital mental health is needed, specifically more focused research on the potential moderators or mediators of effects. Even engagement itself may require mechanistic research as it has proven to be a challenging construct to improve upon. For example, a recent review of digital therapeutics suggested that the field could benefit from the application of clinical pharmacology principles from the drug development field, such as a stepwise and progressive focus on engagement and adherence, proxy of effects, and clinical end points [104]. This would need funders to launch specific calls for this type of research (as this is intricate and costly) and is explored further in theme 3 in the following section.

Theme 3: Regulation, Governance, and Funding

Current Challenges

Funding

Funding is a major issue specifically in digital mental health as the process is often fragmented [105] and may only cover the initial development of digital interventions. The rest of the pathway is also often lengthy and costly, but is critical in the road toward implementation and needs to incorporate the processes of regulation and market transfer. A particularly important area to assess is the implementation of a mental health digital product in the real-world health care setting. This will involve a comprehensive analysis of implementation costs, integration into the existing IT infrastructures of the institution, and regulatory and privacy compliance, but this may not have been adequately assessed at the beginning of the process. A result is that many digital health technologies are not iterated and sustained, with a 2022 review finding that nearly half of apps created for schizophrenia research in the last decade are no longer accessible or supported [11]. These challenges in funding impede replication and impair the research-to-clinical translation of digital mental health tools (see further discussion in theme 4). An area of additional challenge is that research funding has usually not considered the complexity of science (including the need for significant user input throughout the research cycle, as explored in theme 1). In addition, digital interventions have technological complexities that have a significant impact on costs, such as the need to maintain apps on an ongoing basis, provide updates, and implement new and regular cybersecurity and data protection measures.

Regulation

These challenges are complicated by a lack of health care regulation specifically tailored for the digital space. For example, in the United States, many digital interventions fall within FDA regulations that are challenging when applied to this area, and so many interventions do not come to market. There are also other extra regulatory issues that need to be formally addressed regarding safety and data protection, including privacy and security [106]. For example, data collection in digital phenotyping involves tracking patients beyond traditional health information and can involve data such as social media posts, geolocation, and telephone and SMS text message traffic (among other data) that can provide revealing insights about the daily lives of individuals [107,108]. Studies also show that there are significant limitations with clinicians' awareness of the ethical considerations regarding artificial intelligence–powered innovations in health care [109-113] (Textbox 5). Furthermore, health laws have not kept up with digital technologies, although authorities have recently made efforts to regulate technology while also aiming to protect civil liberties and rights to privacy [107,108].

Potential Solutions

Solutions include advocating for governance change, but this would need to ensure the involvement of and contributions from multiple stakeholders (patients and carers, technology experts, researchers, clinicians, health technology assessment agencies, and regulators). One option as a model might be a roundtable discussion with all relevant stakeholders, such as that at the recent UK AI Safety Summit [114,115]. Governance change could also encompass standards for the level of evidence required, including controlling for digital placebo effects and demonstrating savings in cost [116]. In line with governance change, funding in this area needs to be more adaptive to recognize the particular needs in this space (eg, commitments to funding until completion or funding the people rather than the project [117]). As an intermediate step, guidelines for best practices and research standards that funders and journals may enforce would also shift the field in a positive direction. In terms of ethical concerns, including how to navigate evolving regulations, clinicians will require greater training, and patients will need more guidance and advice on the benefits and potential risks of these digital tools in health care settings (Textbox 5). Brief training interventions should be considered for all stakeholders. For example, among patients, the use of "digital navigators"—peer supporters who can offer patients advice on how to download and use apps, where to find information about privacy considerations, and the evidence base for these tools—has been pioneered in outpatient psychiatry with notable success [118].

Theme 4: Implementation in Real-World Clinical Settings

Current Challenges

Translation of research findings into the clinical setting is a significant challenge. In general, many digital health studies fail to reach the market or translate into real-world clinical care, and even when they do, rates of adoption can be low [119]. For example, in a study of smartphone apps for schizophrenia, <10 of those identified from a search of interventions from the research literature were easily accessible to the population, and the picture was similar in a parallel search of marketplace apps [11]. These were also few in number and lacked frequent updates (average time since last update 1121 days). Even where there is engagement in the research setting, sustaining this in the clinic, outside of the constraints and incentives of clinical trials, is challenging, particularly over the longer term. For example, digital health tools with 44% to 99% completion rates in research studies translate to only 1% to 28% in actual clinical use [120].

There are several potential reasons for this:

1. Researchers are often not trained in the technology skills and application of the implementation science methods required to ensure the tailoring and ongoing development of interventions for real-world uptake.
2. Even once digital interventions have been fully developed, individual clinicians may be reluctant to implement them in the clinic. Uptake requires behavior change from clinicians, who may be resistant due to lack of training,

skills, or confidence in digital interventions, regulatory issues, and cautions and anxieties about adopting commercial therapy products.
3. Patients, carers, and the public may also be reluctant due to digital inclusion issues such as digital literacy and confidence (especially for some groups, such as older patients or patients who are severely ill [121]) as well as concerns about privacy and digital coercion [122].

In addition, the focus of digital mental health research may contribute to difficulties in translation into real-world settings:

1. Many research studies focus on “replacing” the clinician, with apps designed for stand-alone use, whereas the research base suggests that blended or augmented care, in line with patient preference, is more beneficial [16,118,123].
2. Many studies focus on general outcomes rather than the mechanisms of change, and without registration of studies, there is a risk of inefficient use of funding resources.
3. The focus of digital mental health research also tends to be on biological and psychological parameters, whereas social mechanisms are equally important [10].
4. For the user, there is a lack of guidance on which interventions might be better or worse. This is a long-standing problem [124] and reflects the lack of evidence base. For example, assessments of smartphone tools for suicide prevention have identified a wide variety of approaches but also include apps with potentially harmful content [125,126].

Potential Solutions

Successful development, implementation, and application in real-world clinical settings will require collaboration between multidisciplinary groups who have not traditionally been brought together, for example, a wider range of academics with complementary expertise, clinicians, user groups and industrial designers, software developers, commercial partners, and regulatory specialists. Skills and confidence in using digital interventions will be a key element. Training schedules for clinicians in using and integrating digital tools into their clinical practice have already been defined, and these need to be integrated into clinical training at all levels from core training to specialist academic programs [127], with health care organizations creating sustained budgets to fund digital inclusion schemes. Patient and user confidence is also important, and there are examples of effective training schemes for SMI (eg, the study by Hoffman et al [20]) and of interventions to help bridge the technological divide, such as digital navigators [118] and specialized youth mental health workers [44].

Workable and scalable solutions will rest in augmenting the in-person consultation using a blended approach. In addition, the focus needs to be on increasing the capacity of clinical services, such as by augmenting service provision with support workers who can deliver protocolized interventions. Implementation science frameworks can be used to guide the assessment of site readiness and evaluate their ability to successfully introduce, implement, and sustain digital technology use [119,128].

Solutions will also depend on looking at social interventions, which may be population focused as well as targeted at the individual or specific group level. Examples include wide-reaching digital training, awareness and antistigma campaigns [129], and automated solutions to reduce access and exposure to lethal means of suicide [130-132]. Although in the past there have been difficulties in formal guidance for the user [121], the American Psychiatric Association app evaluation framework offers a viable alternative [133].

Discussion

Overview

In this paper, we have illustrated and discussed the complexities of collecting data, delivering treatment, and the ethical challenges of digital mental health in the care pathways of people with SMI. During the process of the consensus meeting and the consensus recommendations, we implemented a thematic approach focusing broadly on digital interventions for psychosis. This enabled us to concentrate on an area of significant need where digital health innovation has the potential to be safe and effective [134]. However, we also found that, in taking this thematic approach, we identified broader issues, and the solutions proposed can apply to other fields of mental health.

Potential Limitations

There are some potential limitations to our approach. While we conducted a systematic review of the literature, we restricted this to one source (PubMed) and to articles published in the last 5 years on psychosis and schizophrenia. The primary aim of this review was to ensure that the panel was informed before the meeting, but it is possible that this approach may have missed some relevant publications (eg, on other diagnoses within SMI, such as bipolar disorder). In addition, as with all consensus meetings, there are no standard guidelines for identifying expertise. Although we selected participants to represent a diverse spectrum of views, the reliability of consensus opinions is dependent on the specialist knowledge and experiences of those who participated. We aimed to encompass a wide variety of expertise in the expert group and panel to ensure that many different perspectives were heard. These included clinical psychology, psychotherapy, psychiatry, philosophy of medicine and health care ethics, health services research, social sciences, health informatics, digital health care, and mental health charity management. We included an expert with lived experience of SMI in the panel but recognize that we could have included more, which we take on board for future consensus studies. The structured in-person nature of the meeting may also have unintentionally excluded the opinions of experts, particularly those with lived experience of SMI, who are not willing or able to engage in that format. However, we had 2 PPI contributors within the process (one in the panel and one in the expert group) who made material contributions throughout; to the literature review, presentation and discussion of the evidence, formation of consensus, and coproduction of the paper. In this way, we aimed to engage a high-quality PPI coproduction rather than focusing purely on the number of PPI members involved. In addition, by adopting and building on new features of the consensus method used in our previous work [6] (eg, including a panel separate from the expert group and including PPIE members in the expert group and panel), we have strengthened international and multidisciplinary discussion on this important but often overlooked area [134].

Principal Findings

The consensus meeting identified a number of broad recommendations in this field:

1. A new approach to research in digital mental health is needed that is different from the standard pathways used in pharmacological and psychological intervention research.
2. This new approach requires internationally agreed standards for reporting research and open data access to allow for true collaboration and enable easy validation of biomarkers and replication of interventions.
3. This could be facilitated by the development of shared protocols for research to be carried out in multiple recruiting sites.
4. Research should place an equal emphasis on social and population factors [135] as well as biological and social factors in the etiology and maintenance of symptoms and risk in SMI.
5. Successful implementation and application of digital mental health in real-world clinical settings will require new and evolving collaborations between academics, clinicians, people with lived experience of SMI, industrial designers, software developers, and regulatory specialists.
6. The uniqueness of the digital space in clinical research means that it requires a different approach, focusing not just on the translational pathway of research in isolation but also on the business model; the "product"; and the impact or value of the intervention for all relevant stakeholders, from patients to clinicians, health care organizations, and society at large.
7. This new approach may prompt a possible conflict of values as the "product" (the digital mental health intervention) needs to be economically viable so that it can be scaled and sustained while also ensuring effectiveness, safety, and compliance with medical technology quality standards and health care regulation.
8. Potential harms are just as important to record as in other areas of research (such as with pharmacological or psychological interventions) but may be more hidden and need to be actively sought out and logged.
9. Funding structures need to be adjusted to the new elements required within digital mental health research and must be sufficient to support ongoing product development in line with regulatory requirements and allow for representative PPIE.
10. Funding streams will need to recognize that not all the tasks required can be performed by a single person or group and they may need a "relay" approach between stakeholders (ie, projects led by clinician academics at the proof-of-concept, feasibility, and efficacy stages, with increased commercialization and regulatory input as products move into implementation and market).
11. In addition, more fundamental organizational changes are needed to underpin necessary changes to funding and research study approaches. Participants with lived experience and academic experts do not always "speak the same language" [136], so awareness of differences in expression and the need to work together to solve health problems is necessary to minimize the impact of power imbalances and promote coproduction [137]. This can be achieved by developing safe spaces to create and share knowledge [138] and allow for opportunities for researchers and participants with lived experience to learn and enrich their own expertise from the experience of informed participation and collaboration [49]. This improved communication could provide the platform to create new models of care to deliver digital services, which will also require adjustments to organizational structure, policies, and membership.
12. The ethical components of digital mental health are also crucial. This is not just in terms of trust and trustworthiness in digital mental health but also in managing patient expectations, "ownership of their own health," and future developments.

Conclusions

In this study, our approach, which combined an international expert meeting with PPIE throughout the process, consensus methodology, discussion, and publication, was a fruitful way to reach expert consensus and focus directions for future research and clinical implementation, especially in rapidly evolving fields. We improved and expanded our approach and showed how to integrate research evidence with a process of measuring real-world clinical impact over time [139]. To enhance our scope, future meetings should directly involve stakeholders in health technology assessment and representatives from regulatory agencies and industry and encompass researchers working in health care ethics and policy in different countries, including from the Global South. Similar initiatives should be repeated regularly in digital mental health and adopted also by researchers in other fields to focus research and organizational change to effect real-world clinical implementation.

Acknowledgments

We acknowledge especially the contribution of our co-author and colleague, Lea Milligan, CEO of MQ Mental Health Research, who sadly passed away unexpectedly on the 15th of April 2024 following a sudden illness. He made a significant and much-valued contribution to this research project as an active member of the expert panel and co-author, and will be hugely missed

KAS, CZ, EGO, and AC are supported by the National Institute for Health and Care Research (NIHR) Oxford Health Clinical Research Facility. AC and EGO are supported by an NIHR Research Professorship (grant RP-2017-08-ST2-006), the NIHR

Oxford and Thames Valley Applied Research Collaboration, and the NIHR Oxford Health Biomedical Research Centre (grant BRC-1215-20005). AC is supported by Wellcome (Global Alliance for Living Evidence on Anxiety, Depression, and Psychosis [GALENOS] project). EGO is supported by the Brasenose College Senior Hulme Scholarship. AH is partly funded by the NIHR Maudsley Biomedical Research Centre at the South London and Maudsley National Health Service Foundation Trust and King's College London. LM is partly funded by Wellcome and the Medical Research Council Health Data Research United Kingdom. SL is supported by an NIHR Doctoral Fellowship (NIHR301483) and the NIHR Oxford Health Biomedical Research Centre (BRC-1215-20005). RA is supported by the Italian Ministry of University and Research via the Residency Training Programme in Psychiatry. GA is supported by a Juan Rodés 2023 grant (JR23/00050), a Rio Hortega 2021 grant (CM21/00017) and Acción Estratégica en Salud - Mobility (M-AES) fellowship (MV22/00058), from the Spanish Ministry of Health financed by the Instituto de Salud Carlos III (ISCIII) and co-financed by the Fondo Social Europeo Plus (FSE+). GA thanks the support of the Spanish Ministry of Health financed by the Instituto de Salud Carlos III (ISCIII) and co-financed by the European Social Fund+ (ESF+) (JR23/00050, MV22/00058, CM21/00017); the ISCIII (PI21/00340, PI21/00169); the Milken Family Foundation (PI046998); the Fundació Clínic per a la Recerca Biomèdica (FCRB) - Pons Bartan 2020 grant (PI04/6549), the Sociedad Española de Psiquiatría y Salud Mental (SEPSM); the Fundació Vila Saborit; and the Societat Catalana de Psiquiatria i Salut Mental (SCPiSM). PJU was supported by the Wellcome Trust Innovator Grant 215793/Z/19/Z. This paper is based on a meeting held in Rome in October 2023. The meeting and the logistics of the event were supported by Angelini Pharma, who did not have any influence on the content of the discussion, the outcome, or the preparation of this manuscript. The views expressed are those of the authors and not necessarily those of the UK National Health Service, the NIHR, or the UK Department of Health and Social Care.

Data Availability

All data generated or analyzed during this study are included in this published article.

Authors' Contributions

AC and JT developed the concept for the study and the format of the consensus meeting and chaired the expert group. KAS assisted with the consensus meeting, chaired the panel, wrote the first draft of the manuscript, and coordinated the editing and updates of subsequent drafts. All coauthors attended and actively contributed to the meeting and the process of consensus formation, agreed on consensus statements, and reviewed and approved the final draft of the manuscript.

Conflicts of Interest

KAS, AH, CZ, RA, and LM have no conflicts of interest to declare. AC has received research and consultancy fees from the Italian Network for Pediatric Clinical Trials, the Cariplo Foundation, Lundbeck, and Angelini Pharma outside the submitted work. EGO has received research and consultancy fees from Angelini Pharma outside the submitted work. GA has received honoraria or consulting fees from Angelini Pharma, Casen Recordati, Janssen-Cilag, Lundbeck, Lundbeck-Otsuka, Rovi, and Viatris outside the submitted work. PJU has received honoraria from Boehringer Ingelheim and research funding from Lundbeck and Lilly United Kingdom. SL has received consultancy fees from OxfordVR, a University of Oxford spin-out company that is commercializing the gameChange treatment. JT is cofounder of a mental health company called Precision Mental Wellness and is the editor-in-chief of *JMIR Mental Health* at the time of publication. DH has received CME-related honoraria and served as consultant for Abbott, Angelini, Ethypharm Digital Therapy, Janssen-Cilag, Lundbeck and Viatris outside the submitted work.

References

1. Mahase E. Government must get a "firmer grip" on mental health crisis, says watchdog. BMJ 2023 Feb 09;380:324 [FREE Full text] [doi: 10.1136/bmj.p324] [Medline: 36758976]
2. Hong JS, Sheriff R, Smith K, Tomlinson A, Saad F, Smith T, et al. Impact of COVID-19 on telepsychiatry at the service and individual patient level across two UK NHS mental health trusts. Evid Based Ment Health 2021 Nov;24(4):161-166 [FREE Full text] [doi: 10.1136/ebmental-2021-300287] [Medline: 34583940]
3. Zangani C, Ostinelli EG, Smith KA, Hong JS, Macdonald O, Reen G, et al. Impact of the COVID-19 pandemic on the global delivery of mental health services and telemental health: systematic review. JMIR Ment Health 2022 Aug 22;9(8):e38600 [FREE Full text] [doi: 10.2196/38600] [Medline: 35994310]
4. Sheriff R, Hong JS, Henshall C, D'Agostino A, Tomassi S, Stein HC, et al. Evaluation of telepsychiatry during the COVID-19 pandemic across service users, carers and clinicians: an international mixed-methods study. BMJ Ment Health 2023 Aug;26(1):e300646 [FREE Full text] [doi: 10.1136/bmjment-2022-300646] [Medline: 37567731]
5. Currey D, Torous J. Increasing the value of digital phenotyping through reducing missingness: a retrospective review and analysis of prior studies. BMJ Ment Health 2023 Feb;26(1):e300718 [FREE Full text] [doi: 10.1136/bmjment-2023-300718] [Medline: 37197799]
6. Smith KA, Blease C, Faurholt-Jepsen M, Firth J, Van Daele T, Moreno C, et al. Digital mental health: challenges and next steps. BMJ Ment Health 2023 Feb;26(1):e300670 [FREE Full text] [doi: 10.1136/bmjment-2023-300670] [Medline: 37197797]

7. Patel V, Saxena S, Lund C, Kohrt B, Kieling C, Sunkel C, et al. Transforming mental health systems globally: principles and policy recommendations. Lancet 2023 Aug;402(10402):656-666 [FREE Full text] [doi: 10.1016/s0140-6736(23)00918-2]
8. Asher L, Fekadu A, Hanlon C. Global mental health and schizophrenia. Curr Opin Psychiatry 2018 May;31(3):193-199. [doi: 10.1097/YCO.0000000000000404] [Medline: 29474265]
9. Fiorillo A, de Girolamo G, Simunovic IF, Gureje O, Isaac M, Lloyd C, et al. The relationship between physical and mental health: an update from the WPA working group on managing comorbidity of mental and physical health. World Psychiatry 2023 Feb;22(1):169-170 [FREE Full text] [doi: 10.1002/wps.21055] [Medline: 36640405]
10. O'Connor RC, Worthman CM, Abanga M, Athanassopoulou N, Boyce N, Chan LF, et al. Gone Too Soon: priorities for action to prevent premature mortality associated with mental illness and mental distress. Lancet Psychiatry 2023 Jun;10(6):452-464 [FREE Full text] [doi: 10.1016/s2215-0366(23)00058-5]
11. Kwon S, Firth J, Joshi D, Torous J. Accessibility and availability of smartphone apps for schizophrenia. Schizophrenia (Heidelb) 2022 Nov 16;8(1):98 [FREE Full text] [doi: 10.1038/s41537-022-00313-0] [Medline: 36385116]
12. Kozelka EE, Acquilano SC, Al-Abdulmunem M, Guarino S, Elwyn G, Drake RE, et al. Documenting the digital divide: identifying barriers to digital mental health access among people with serious mental illness in community settings. SSM Mental Health 2023 Dec;4:100241 [FREE Full text] [doi: 10.1016/j.ssmmh.2023.100241]
13. Spanakis P, Lorimer B, Newbronner E, Wadman R, Crosland S, Gilbody S, et al. Digital health literacy and digital engagement for people with severe mental ill health across the course of the COVID-19 pandemic in England. BMC Med Inform Decis Mak 2023 Sep 26;23(1):193 [FREE Full text] [doi: 10.1186/s12911-023-02299-w] [Medline: 37752460]
14. Firth J, Cotter J, Torous J, Bucci S, Firth JA, Yung AR. Mobile phone ownership and endorsement of "mHealth" among people with psychosis: a meta-analysis of cross-sectional studies. Schizophr Bull 2016 Mar 22;42(2):448-455 [FREE Full text] [doi: 10.1093/schbul/sbv132] [Medline: 26400871]
15. Iliescu R, Kumaravel A, Smurawska L, Torous J, Keshavan M. Smartphone ownership and use of mental health applications by psychiatric inpatients. Psychiatry Res 2021 May;299:113806 [FREE Full text] [doi: 10.1016/j.psychres.2021.113806] [Medline: 33667947]
16. Eisner E, Berry N, Bucci S. Digital tools to support mental health: a survey study in psychosis. BMC Psychiatry 2023 Oct 07;23(1):726 [FREE Full text] [doi: 10.1186/s12888-023-05114-y] [Medline: 37803367]
17. Middle R, Welch L. Experiences of digital exclusion and the impact on health in people living with severe mental illness. Front Digit Health 2022 Nov 22;4:1004547 [FREE Full text] [doi: 10.3389/fdgth.2022.1004547] [Medline: 36483535]
18. Spanakis P, Wadman R, Walker L, Heron P, Mathers A, Baker J, et al. Measuring the digital divide among people with severe mental ill health using the essential digital skills framework. Perspect Public Health 2024 Jan;144(1):21-30 [FREE Full text] [doi: 10.1177/17579139221106399] [Medline: 35929589]
19. Hardy A, Ward T, Emsley R, Greenwood K, Freeman D, Fowler D, et al. Bridging the digital divide in psychological therapies: observational study of engagement with the SlowMo mobile app for paranoia in psychosis. JMIR Hum Factors 2022 Jul 01;9(3):e29725 [FREE Full text] [doi: 10.2196/29725] [Medline: 35776506]
20. Hoffman L, Wisniewski H, Hays R, Henson P, Vaidyam A, Hendel V, et al. Digital opportunities for outcomes in recovery services (DOORS): a pragmatic hands-on group approach toward increasing digital health and smartphone competencies, autonomy, relatedness, and alliance for those with serious mental illness. J Psychiatr Pract 2020 Mar;26(2):80-88 [FREE Full text] [doi: 10.1097/PRA.0000000000000450] [Medline: 32134881]
21. Camacho E, Torous J. Impact of digital literacy training on outcomes for people with serious mental illness in community and inpatient settings. Psychiatr Serv 2023 May 01;74(5):534-538 [FREE Full text] [doi: 10.1176/appi.ps.20220205] [Medline: 36164771]
22. García-Estela A, Cantillo J, Angarita-Osorio N, Mur-Milà E, Anmella G, Pérez V, et al. Real-world implementation of a smartphone-based psychoeducation program for bipolar disorder: observational ecological study. J Med Internet Res 2022 Feb 02;24(2):e31565 [FREE Full text] [doi: 10.2196/31565] [Medline: 35107440]
23. Hidalgo-Mazzei D, Nikolova VL, Kitchen S, Young AH. Internet-connected devices ownership, use and interests in bipolar disorder: from desktop to mobile mental health. Digit Psychiatry 2019 May 28;2(1):1-7 [FREE Full text] [doi: 10.1080/2575517x.2019.1616476]
24. Green JB, Rodriguez J, Keshavan M, Lizano P, Torous J. Implementing technologies to enhance coordinated specialty care framework: implementation outcomes from a development and usability study. JMIR Form Res 2023 Oct 03;7:e46491 [FREE Full text] [doi: 10.2196/46491] [Medline: 37788066]
25. Garety P, Ward T, Emsley R, Greenwood K, Freeman D, Fowler D, et al. Effects of SlowMo, a blended digital therapy targeting reasoning, on paranoia among people with psychosis: a randomized clinical trial. JAMA Psychiatry 2021 Jul 01;78(7):714-725 [FREE Full text] [doi: 10.1001/jamapsychiatry.2021.0326] [Medline: 33825827]
26. Arakawa N, Bader LR. Consensus development methods: considerations for national and global frameworks and policy development. Res Social Adm Pharm 2022 Jan;18(1):2222-2229 [FREE Full text] [doi: 10.1016/j.sapharm.2021.06.024] [Medline: 34247949]
27. Fouad Y, Dufour JF, Zheng MH, Bollipo S, Desalegn H, Grønbaek H, et al. The NAFLD-MAFLD debate: is there a consensus-on-consensus methodology? Liver Int 2022 Apr;42(4):742-748 [FREE Full text] [doi: 10.1111/liv.15197] [Medline: 35182007]

28. Ferguson JH. The NIH consensus development program. the evolution of guidelines. Int J Technol Assess Health Care 1996;12(3):460-474. [Medline: 8840666]
29. WHO handbook for guideline development. World Health Organization. 2014. URL: https://www.who.int/publications/i/item/9789241548960 [accessed 2024-04-15]
30. Waggoner J, Carline JD, Durning SJ. Is there a consensus on consensus methodology? Descriptions and recommendations for future consensus research. Acad Med 2016 May;91(5):663-668. [doi: 10.1097/ACM.0000000000001092] [Medline: 26796090]
31. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ 2021 Mar 29;372:n160 [FREE Full text] [doi: 10.1136/bmj.n160] [Medline: 33781993]
32. Gattrell W, Logullo P, van Zuuren EJ, Price A, Hughes EL, Blazey P, et al. ACCORD (ACcurate COnsensus Reporting Document): a reporting guideline for consensus methods in biomedicine developed via a modified Delphi. PLoS Med 2024 Jan;21(1):e1004326 [FREE Full text] [doi: 10.1371/journal.pmed.1004326] [Medline: 38261576]
33. Humphrey-Murto S, Varpio L, Gonsalves C, Wood TJ. Using consensus group methods such as Delphi and nominal group in medical education research. Med Teach 2016 Nov 12;39(1):14-19 [FREE Full text] [doi: 10.1080/0142159x.2017.1245856]
34. Graham A, Lattie EG, Mohr DC. Experimental therapeutics for digital mental health. JAMA Psychiatry 2019 Dec 01;76(12):1223-1224 [FREE Full text] [doi: 10.1001/jamapsychiatry.2019.2075] [Medline: 31433448]
35. Perski O, Blandford A, West R, Michie S. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. Transl Behav Med 2017 Jun;7(2):254-267 [FREE Full text] [doi: 10.1007/s13142-016-0453-1] [Medline: 27966189]
36. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018 Aug;21(3):116-119 [FREE Full text] [doi: 10.1136/eb-2018-102891] [Medline: 29871870]
37. Nwosu A, Boardman S, Husain MM, Doraiswamy PM. Digital therapeutics for mental health: is attrition the Achilles heel? Front Psychiatry 2022;13:900615 [FREE Full text] [doi: 10.3389/fpsyt.2022.900615] [Medline: 35982936]
38. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021 Oct;20(3):318-335 [FREE Full text] [doi: 10.1002/wps.20883] [Medline: 34505369]
39. Ng MM, Firth J, Minen M, Torous J. User engagement in mental health apps: a review of measurement, reporting, and validity. Psychiatr Serv 2019 Jul 01;70(7):538-544 [FREE Full text] [doi: 10.1176/appi.ps.201800519] [Medline: 30914003]
40. Merchant R, Torous J, Rodriguez-Villa E, Naslund JA. Digital technology for management of severe mental disorders in low-income and middle-income countries. Curr Opin Psychiatry 2020 Sep;33(5):501-507 [FREE Full text] [doi: 10.1097/YCO.0000000000000626] [Medline: 32520747]
41. Aref-Adib G, McCloud T, Ross J, O'Hanlon P, Appleton V, Rowe S, et al. Factors affecting implementation of digital health interventions for people with psychosis or bipolar disorder, and their family and friends: a systematic review. Lancet Psychiatry 2019 Mar;6(3):257-266 [FREE Full text] [doi: 10.1016/s2215-0366(18)30302-x]
42. Singh SR, Sagar R. Time to have effective regulation of the mental health apps market: maximize gains and minimize harms. Indian J Psychol Med 2022 Jul;44(4):399-404 [FREE Full text] [doi: 10.1177/02537176221082902] [Medline: 35949633]
43. Solmi M, Radua J, Olivola M, Croce E, Soardo L, Salazar de Pablo G, et al. Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. Mol Psychiatry 2022 Jan 02;27(1):281-295 [FREE Full text] [doi: 10.1038/s41380-021-01161-7] [Medline: 34079068]
44. Uhlhaas PJ, Davey CG, Mehta UM, Shah J, Torous J, Allen NB, et al. Towards a youth mental health paradigm: a perspective and roadmap. Mol Psychiatry 2023 Aug;28(8):3171-3181 [FREE Full text] [doi: 10.1038/s41380-023-02202-z] [Medline: 37580524]
45. Teens, social media and technology. Pew Research Center. 2023. URL: https://www.pewresearch.org/internet/2023/12/11/teens-social-media-and-technology-2023/ [accessed 2024-04-15]
46. Wong CA, Madanay F, Ozer EM, Harris SK, Moore M, Master SO, et al. Digital health technology to enhance adolescent and young adult clinical preventive services: affordances and challenges. J Adolesc Health 2020 Aug;67(2S):S24-S33 [FREE Full text] [doi: 10.1016/j.jadohealth.2019.10.018] [Medline: 32718511]
47. Bergin AD, Vallejos EP, Davies EB, Daley D, Ford T, Harold G, et al. Preventive digital mental health interventions for children and young people: a review of the design and reporting of research. NPJ Digit Med 2020;3:133 [FREE Full text] [doi: 10.1038/s41746-020-00339-7] [Medline: 33083568]
48. Doughty B. What is co-production? MQ Mental Health Research. 2021. URL: https://www.mqmentalhealth.org/co-production/ [accessed 2024-04-05]
49. Richmond L, Caton N, Downs J, Newton A, Devereux-Fitzgerald A, Brooks H. Creating positive experiences of involvement in mental health research. Lancet Psychiatry 2023 Mar;10(3):220-227. [doi: 10.1016/s2215-0366(22)00430-8]

50. Landscaping international longitudinal datasets: theory of change process and a deeper dive into lived experience input. MQ Mental Health Research. URL: https://www.mqmentalhealth.org/wp-content/uploads/MQ-Lived-Experience-Report.pdf [accessed 2024-04-15]
51. Guidance on co-producing a research project. National Institutes for Health and Care Research. URL: https://www.learningforinvolvement.org.uk/content/resource/nihr-guidance-on-co-producing-a-research-project/ [accessed 2024-04-15]
52. Sartor C. Mental health and lived experience: the value of lived experience expertise in global mental health. Glob Ment Health (Camb) 2023;10:e38-e35 [FREE Full text] [doi: 10.1017/gmh.2023.24] [Medline: 37854403]
53. Riggare S. E-patients hold key to the future of healthcare. BMJ 2018 Feb 26;360:k846 [FREE Full text] [doi: 10.1136/bmj.k846] [Medline: 29483151]
54. Price A, Clarke M, Staniszewska S, Chu L, Tembo D, Kirkpatrick M, et al. Patient and public involvement in research: a journey to co-production. Patient Educ Couns 2022 Apr;105(4):1041-1047 [FREE Full text] [doi: 10.1016/j.pec.2021.07.021] [Medline: 34334264]
55. Byrne L, Wykes T. A role for lived experience mental health leadership in the age of COVID-19. J Ment Health 2020 Jun;29(3):243-246 [FREE Full text] [doi: 10.1080/09638237.2020.1766002] [Medline: 32449392]
56. Holmes EA, O'Connor RC, Perry VH, Tracey I, Wessely S, Arseneault L, et al. Multidisciplinary research priorities for the COVID-19 pandemic: a call for action for mental health science. Lancet Psychiatry 2020 Jun;7(6):547-560 [FREE Full text] [doi: 10.1016/S2215-0366(20)30168-1] [Medline: 32304649]
57. Vaidyam A, Halamka J, Torous J. Enabling research and clinical use of patient-generated health data (the mindLAMP platform): digital phenotyping study. JMIR Mhealth Uhealth 2022 Jan 07;10(1):e30557 [FREE Full text] [doi: 10.2196/30557] [Medline: 34994710]
58. Rodriguez-Villa E, Rozatkar AR, Kumar M, Patel V, Bondre A, Naik SS, et al. Cross cultural and global uses of a digital mental health app: results of focus groups with clinicians, patients and family members in India and the United States. Glob Ment Health (Camb) 2021;8:e30 [FREE Full text] [doi: 10.1017/gmh.2021.28] [Medline: 34512999]
59. Ward T, Hardy A, Holm R, Collett N, Rus-Calafell M, Sacadura C, et al. SlowMo therapy, a new digital blended therapy for fear of harm from others: an account of therapy personalisation within a targeted intervention. Psychol Psychother 2022 Jun 12;95(2):423-446 [FREE Full text] [doi: 10.1111/papt.12377] [Medline: 35019210]
60. Garety P, Waller H, Emsley R, Jolley S, Kuipers E, Bebbington P, et al. Cognitive mechanisms of change in delusions: an experimental investigation targeting reasoning to effect change in paranoia. Schizophr Bull 2015 Mar;41(2):400-410 [FREE Full text] [doi: 10.1093/schbul/sbu103] [Medline: 25053650]
61. Ross K, Freeman D, Dunn G, Garety P. A randomized experimental investigation of reasoning training for people with delusions. Schizophr Bull 2011 Mar;37(2):324-333 [FREE Full text] [doi: 10.1093/schbul/sbn165] [Medline: 19520745]
62. Waller H, Emsley R, Freeman D, Bebbington P, Dunn G, Fowler D, et al. Thinking well: a randomised controlled feasibility study of a new CBT therapy targeting reasoning biases in people with distressing persecutory delusional beliefs. J Behav Ther Exp Psychiatry 2015 Sep;48:82-89 [FREE Full text] [doi: 10.1016/j.jbtep.2015.02.007] [Medline: 25770671]
63. Waller H, Freeman D, Jolley S, Dunn G, Garety P. Targeting reasoning biases in delusions: a pilot study of the Maudsley review training programme for individuals with persistent, high conviction delusions. J Behav Ther Exp Psychiatry 2011 Sep;42(3):414-421 [FREE Full text] [doi: 10.1016/j.jbtep.2011.03.001] [Medline: 21481815]
64. Morris RM, Sellwood W, Edge D, Colling C, Stewart R, Cupitt C, et al. Ethnicity and impact on the receipt of cognitive-behavioural therapy in people with psychosis or bipolar disorder: an English cohort study. BMJ Open 2020 Dec 15;10(12):e034913 [FREE Full text] [doi: 10.1136/bmjopen-2019-034913] [Medline: 33323425]
65. Turner DT, Burger S, Smit F, Valmaggia LR, van der Gaag M. What constitutes sufficient evidence for case formulation-driven CBT for psychosis? Cumulative meta-analysis of the effect on hallucinations and delusions. Schizophr Bull 2020 Sep 21;46(5):1072-1085 [FREE Full text] [doi: 10.1093/schbul/sbaa045] [Medline: 32221536]
66. Ward T, Garety PA. Fast and slow thinking in distressing delusions: a review of the literature and implications for targeted therapy. Schizophr Res 2019 Jan;203:80-87 [FREE Full text] [doi: 10.1016/j.schres.2017.08.045] [Medline: 28927863]
67. Greenwood KE, Gurnani M, Ward T, Vogel E, Vella C, McGourty A, SlowMo Patient, Public Involvement (PPI) team. The service user experience of SlowMo therapy: a co-produced thematic analysis of service users' subjective experience. Psychol Psychother 2022 Sep 20;95(3):680-700 [FREE Full text] [doi: 10.1111/papt.12393] [Medline: 35445520]
68. Greenwood K, Robertson S, Vogel E, Vella C, Ward T, McGourty A, SlowMo PatientPublic Involvement (PPI) team. The impact of patient and public involvement in the SlowMo study: reflections on peer innovation. Health Expect 2022 Feb;25(1):191-202 [FREE Full text] [doi: 10.1111/hex.13362] [Medline: 34585482]
69. Guidance assessment consultation document for GID-HTE10020 Digital health technologies to help manage symptoms of psychosis and prevent relapse: early value assessment. National Institute for Health and Care Excellence. 2024 Jan 18. URL: https://www.nice.org.uk/guidance/GID-HTE10020/documents/draft-guidance [accessed 2024-04-05]
70. Freeman D, Taylor KM, Molodynski A, Waite F. Treatable clinical intervention targets for patients with schizophrenia. Schizophr Res 2019 Sep;211:44-50 [FREE Full text] [doi: 10.1016/j.schres.2019.07.016] [Medline: 31326234]
71. Lambe S, Knight I, Kabir T, West J, Patel R, Lister R, et al. Developing an automated VR cognitive treatment for psychosis: gameChange VR therapy. J Behav Cogn Ther 2020 Apr;30(1):33-40 [FREE Full text] [doi: 10.1016/j.jbct.2019.12.001]

72. Freeman D, Lambe S, Kabir T, Petit A, Rosebrock LM, Yu L, et al. Automated virtual reality therapy to treat agoraphobic avoidance and distress in patients with psychosis (gameChange): a multicentre, parallel-group, single-blind, randomised, controlled trial in England with mediation and moderation analyses. Lancet Psychiatry 2022 May;9(5):375-388 [FREE Full text] [doi: 10.1016/s2215-0366(22)00060-8]
73. Altunkaya J, Craven M, Lambe S, Beckley A, Rosebrock L, Dudley R, et al. Estimating the economic value of automated virtual reality cognitive therapy for treating agoraphobic avoidance in patients with psychosis: findings from the gameChange randomized controlled clinical trial. J Med Internet Res 2022 Nov 18;24(11):e39248 [FREE Full text] [doi: 10.2196/39248] [Medline: 36399379]
74. Virtual reality technologies for treating agoraphobia or agoraphobic avoidance: early value assessment. National Institute for Health and Care Excellence. 2023 Nov 15. URL: https://www.nice.org.uk/guidance/hte15 [accessed 2024-04-05]
75. Tomlinson A, Furukawa TA, Efthimiou O, Salanti G, De Crescenzo F, Singh I, et al. Personalise antidepressant treatment for unipolar depression combining individual choices, risks and big data (PETRUSHKA): rationale and protocol. Evid Based Ment Health 2020 May;23(2):52-56 [FREE Full text] [doi: 10.1136/ebmental-2019-300118] [Medline: 31645364]
76. Sadeh Y, Denejkina A, Karyotaki E, Lenferink LI, Kassam-Adams N. Erratum: opportunities for improving data sharing and FAIR data practices to advance global mental health-corrigendum. Glob Ment Health (Camb) 2023;10:e20 [FREE Full text] [doi: 10.1017/gmh.2023.14] [Medline: 37854415]
77. Fusar-Poli P. Integrated mental health services for the developmental period (0 to 25 years): a critical review of the evidence. Front Psychiatry 2019 Jun 7;10:355 [FREE Full text] [doi: 10.3389/fpsyt.2019.00355] [Medline: 31231250]
78. Birnbaum ML, Rizvi AF, Faber K, Addington J, Correll CU, Gerber C, et al. Digital trajectories to care in first-episode psychosis. Psychiatr Serv 2018 Dec 01;69(12):1259-1263 [FREE Full text] [doi: 10.1176/appi.ps.201800180] [Medline: 30256181]
79. McDonald M, Christoforidou E, Van Rijsbergen N, Gajwani R, Gross J, Gumley AI, et al. Using online screening in the general population to detect participants at clinical high-risk for psychosis. Schizophr Bull 2019 Apr 25;45(3):600-609 [FREE Full text] [doi: 10.1093/schbul/sby069] [Medline: 29889271]
80. Reilly T, Mechelli A, McGuire P, Fusar-Poli P, Uhlhaas PJ. E-clinical high risk for psychosis: viewpoint on potential of digital innovations for preventive psychiatry. JMIR Ment Health 2019 Oct 03;6(10):e14581 [FREE Full text] [doi: 10.2196/14581] [Medline: 31584006]
81. Tiwari M. Young people can help to ease the global mental health crisis. Nat Hum Behav 2023 Jun;7(6):831-832 [FREE Full text] [doi: 10.1038/s41562-023-01595-9] [Medline: 37081100]
82. Torous J, Rodriguez J, Powell A. The new digital divide for digital biomarkers. Digit Biomark 2017 Sep;1(1):87-91 [FREE Full text] [doi: 10.1159/000477382] [Medline: 29104959]
83. Framework for the use of digital health technologies in drug and biological product development. US Food and Drug Administration. URL: https://www.fda.gov/media/166396/download [accessed 2024-04-15]
84. Kumar A, Ross JS, Patel NA, Rathi V, Redberg RF, Dhruva SS. Studies of prescription digital therapeutics often lack rigor and inclusivity. Health Aff (Millwood) 2023 Nov;42(11):1559-1567 [FREE Full text] [doi: 10.1377/hlthaff.2023.00384] [Medline: 37931187]
85. Jenkins DA, Martin GP, Sperrin M, Riley RD, Debray TP, Collins GS, et al. Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? Diagn Progn Res 2021 Jan 11;5(1):1 [FREE Full text] [doi: 10.1186/s41512-020-00090-3] [Medline: 33431065]
86. Brown P, Prest B, Miles P, Rossi V. The development of national safety and quality digital mental health standards. Australas Psychiatry 2022 Apr;30(2):154-157 [FREE Full text] [doi: 10.1177/10398562211042361] [Medline: 34569321]
87. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 2020 Mar 20;368:l6927 [FREE Full text] [doi: 10.1136/bmj.l6927] [Medline: 32198138]
88. Radua J, Koutsouleris N. Ten simple rules for using machine learning in mental health research. Biol Psychiatry (Forthcoming) 2023 Nov 18 [FREE Full text] [doi: 10.1016/j.biopsych.2023.11.012] [Medline: 37981177]
89. Bent B, Wang K, Grzesiak E, Jiang C, Qi Y, Jiang Y, et al. The digital biomarker discovery pipeline: an open-source software platform for the development of digital biomarkers using mHealth and wearables data. J Clin Transl Sci 2020 Jul 14;5(1):e19 [FREE Full text] [doi: 10.1017/cts.2020.511] [Medline: 33948242]
90. Accelerating medicines Partnership® program for schizophrenia (AMP® SCZ). National Institute of Mental Health. 2023. URL: https://www.nimh.nih.gov/research/research-funded-by-nimh/research-initiatives/accelerating-medicines-partnershipr-program-schizophrenia-ampr-scz [accessed 2024-04-15]
91. Constantinides C, Han LK, Alloza C, Antonucci LA, Arango C, Ayesa-Arriola R, ENIGMA Schizophrenia Consortium, et al. Brain ageing in schizophrenia: evidence from 26 international cohorts via the ENIGMA Schizophrenia consortium. Mol Psychiatry 2023 Mar;28(3):1201-1209 [FREE Full text] [doi: 10.1038/s41380-022-01897-w] [Medline: 36494461]
92. Ben-Zeev D, Brian R, Wang R, Wang W, Campbell AT, Aung MS, et al. CrossCheck: integrating self-report, behavioral sensing, and smartphone use to identify digital indicators of psychotic relapse. Psychiatr Rehabil J 2017 Sep;40(3):266-275 [FREE Full text] [doi: 10.1037/prj0000243] [Medline: 28368138]

93. Part 1: the connection between substance use disorders and mental illness. National Institute on Drug Abuse. 2022. URL: https://nida.nih.gov/publications/research-reports/common-comorbidities-substance-use-disorders/part-1-connection-between-substance-use-disorders-mental-illness [accessed 2024-04-05]
94. Torous J, Blease C. Return of results in digital phenotyping: ethical considerations for real-world use cases. Am J Bioeth 2024 Feb 31;24(2):91-93. [doi: 10.1080/15265161.2024.2298146] [Medline: 38295252]
95. Sieck CJ, Sheon A, Ancker JS, Castek J, Callahan B, Siefer A. Digital inclusion as a social determinant of health. NPJ Digit Med 2021 Mar 17;4(1):52-53 [FREE Full text] [doi: 10.1038/s41746-021-00413-8] [Medline: 33731887]
96. Report on mental health research priorities. The White House. 2023. URL: https://www.whitehouse.gov/wp-content/uploads/2023/02/White-House-Report-on-Mental-Health-Research-Priorities.pdf [accessed 2024-04-05]
97. Bucci S, Richardson C, Allan S. iCHARTS. Schizophrenia International Research Society. URL: https://schizophreniaresearchsociety.org/sandra-bucci-cara-richardson-stephanie-allan-icharts/ [accessed 2024-04-15]
98. Cashin AG, Richards GC, DeVito NJ, Mellor DT, Lee H. Registration of health and medical research. BMJ Evid Based Med 2023 Feb;28(1):68-72 [FREE Full text] [doi: 10.1136/bmjebm-2021-111836] [Medline: 34933926]
99. Carvalho AF, Köhler CA, Fernandes BS, Quevedo J, Miskowiak KW, Brunoni AR, et al. Bias in emerging biomarkers for bipolar disorder. Psychol Med 2016 May 19;46(11):2287-2297 [FREE Full text] [doi: 10.1017/s0033291716000957]
100. La Sala L, Teh Z, Lamblin M, Rajaram G, Rice S, Hill N, et al. Can a social media intervention improve online communication about suicide? A feasibility study examining the acceptability and potential impact of the #chatsafe campaign. PLoS One 2021;16(6):e0253278 [FREE Full text] [doi: 10.1371/journal.pone.0253278] [Medline: 34129610]
101. Online safety act 2023. UK Government. URL: https://www.legislation.gov.uk/ukpga/2023/50/enacted [accessed 2024-04-05]
102. Kirtley OJ, O'Connor RC. Suicide prevention is everyone's business: challenges and opportunities for Google. Soc Sci Med 2020 Oct;262:112691 [FREE Full text] [doi: 10.1016/j.socscimed.2019.112691] [Medline: 32247626]
103. Self-harm and the internet: a user-led critique of the proposed online safety bill. National Survivor User Network. 2023. URL: https://www.nsun.org.uk/self-harm-and-the-internet-a-user-led-critique-of-the-proposed-online-safety-bill/ [accessed 2024-04-05]
104. Ribba B, Peck R, Hutchinson L, Bousnina I, Motti D. Digital therapeutics as a new therapeutic modality: a review from the perspective of clinical pharmacology. Clin Pharmacol Ther 2023 Sep 10;114(3):578-590 [FREE Full text] [doi: 10.1002/cpt.2989] [Medline: 37392464]
105. Titov N, Hadjistavropoulos HD, Nielssen O, Mohr DC, Andersson G, Dear BF. From research to practice: ten lessons in delivering digital mental health services. J Clin Med 2019 Aug 17;8(8):1239 [FREE Full text] [doi: 10.3390/jcm8081239] [Medline: 31426460]
106. Huckvale K, Torous J, Larsen ME. Assessment of the data sharing and privacy practices of smartphone apps for depression and smoking cessation. JAMA Netw Open 2019 Apr 05;2(4):e192542 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.2542] [Medline: 31002321]
107. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. The White House. 2023 Oct. URL: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ [accessed 2024-04-05]
108. Artificial intelligence act: council and parliament strike a deal on the first rules for AI in the world. Council of the European Union. 2023. URL: https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/ [accessed 2023-12-11]
109. Blease C, Kaptchuk TJ, Bernstein MH, Mandl KD, Halamka JD, DesRoches CM. Artificial intelligence and the future of primary care: exploratory qualitative study of UK general practitioners' views. J Med Internet Res 2019 Mar 20;21(3):e12802 [FREE Full text] [doi: 10.2196/12802] [Medline: 30892270]
110. Blease C, Kharko A, Bernstein M, Bradley C, Houston M, Walsh I, et al. Machine learning in medical education: a survey of the experiences and opinions of medical students in Ireland. BMJ Health Care Inform 2022 Feb;29(1):e100480 [FREE Full text] [doi: 10.1136/bmjhci-2021-100480] [Medline: 35105606]
111. Blease C, Torous J. ChatGPT and mental healthcare: balancing benefits with risks of harms. BMJ Ment Health 2023 Nov;26(1):e300884 [FREE Full text] [doi: 10.1136/bmjment-2023-300884] [Medline: 37949485]
112. Hswen Y, Abbasi J. AI will-and should-change medical school, says Harvard's dean for medical education. JAMA 2023 Nov 21;330(19):1820-1823 [FREE Full text] [doi: 10.1001/jama.2023.19295] [Medline: 37878288]
113. Torous J, Blease C. Generative artificial intelligence in mental health care: potential benefits and current challenges. World Psychiatry 2024 Feb;23(1):1-2 [FREE Full text] [doi: 10.1002/wps.21148] [Medline: 38214643]
114. Expert comment: leading AI nations convene for day one of the UK AI summit. University of Oxford. 2023. URL: https://www.ox.ac.uk/news/2023-11-01-expert-comment-leading-ai-nations-convene-day-one-uk-ai-summit [accessed 2024-04-05]
115. Expert comment: Oxford AI experts comment on the outcomes of the UK AI safety summit. University of Oxford. 2023. URL: https://www.ox.ac.uk/news/2023-11-03-expert-comment-oxford-ai-experts-comment-outcomes-uk-ai-safety-summit [accessed 2024-04-05]
116. Powell A, Torous J. A patient-centered framework for measuring the economic value of the clinical benefits of digital health apps: theoretical modeling. JMIR Ment Health 2020 Oct 30;7(10):e18812 [FREE Full text] [doi: 10.2196/18812] [Medline: 33124995]

117. Ioannidis J. More time for research: fund people not projects. Nature 2011 Sep 28;477(7366):529-531 [FREE Full text] [doi: 10.1038/477529a] [Medline: 21956312]
118. Perret S, Alon N, Carpenter-Song E, Myrick K, Thompson K, Li S, et al. Standardising the role of a digital navigator in behavioural health: a systematic review. Lancet Digit Health 2023 Dec;5(12):e925-e932 [FREE Full text] [doi: 10.1016/s2589-7500(23)00152-8]
119. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]
120. Fleming T, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018 Jun 06;20(6):e199 [FREE Full text] [doi: 10.2196/jmir.9275] [Medline: 29875089]
121. Noori S, Jordan A, Bromage W, Fineberg S, Cahill J, Mathis WS. Navigating the digital divide: providing services to people with serious mental illness in a community setting during COVID-19. SN Soc Sci 2022;2(8):160-161 [FREE Full text] [doi: 10.1007/s43545-022-00470-0] [Medline: 35971454]
122. Hardy A, Wojdecka A, West J, Matthews ED, Golby C, Ward T, et al. How inclusive, user-centered design research can improve psychological therapies for psychosis: development of SlowMo. JMIR Ment Health 2018 Dec 05;5(4):e11222 [FREE Full text] [doi: 10.2196/11222] [Medline: 30518514]
123. Allan S, Beedie S, McLeod HJ, Farhall J, Gleeson J, Bradstreet S, et al. Using EMPOWER in daily life: a qualitative investigation of implementation experiences. BMC Psychiatry 2023 Aug 17;23(1):597 [FREE Full text] [doi: 10.1186/s12888-023-05096-x] [Medline: 37592231]
124. Leigh S, Flatt S. App-based psychological interventions: friend or foe? Evid Based Ment Health 2015 Nov;18(4):97-99. [doi: 10.1136/eb-2015-102203] [Medline: 26459466]
125. Larsen ME, Nicholas J, Christensen H. A systematic assessment of smartphone tools for suicide prevention. PLoS One 2016;11(4):e0152285 [FREE Full text] [doi: 10.1371/journal.pone.0152285] [Medline: 27073900]
126. Wilks CR, Chu C, Sim D, Lovell J, Gutierrez P, Joiner T, et al. User engagement and usability of suicide prevention apps: systematic search in app stores and content analysis. JMIR Form Res 2021 Jul 14;5(7):e27018 [FREE Full text] [doi: 10.2196/27018] [Medline: 34259163]
127. Smith K, Torous J, Cipriani A. Teaching telepsychiatry skills: building on the lessons of the COVID-19 pandemic to enhance mental health care in the future. JMIR Ment Health 2022 Oct 14;9(10):e37939 [FREE Full text] [doi: 10.2196/37939] [Medline: 35358948]
128. Camacho E, Torous J. Introducing an implementation framework for augmenting care with digital technology for early psychosis patients: theory and motivation. J Ment Health 2022 Dec;31(6):816-824 [FREE Full text] [doi: 10.1080/09638237.2021.1922634] [Medline: 34057008]
129. Evans-Lacko S, Corker E, Williams P, Henderson C, Thornicroft G. Effect of the time to change anti-stigma campaign on trends in mental-illness-related public stigma among the English population in 2003–13: an analysis of survey data. Lancet Psychiatry 2014 Jul;1(2):121-128 [FREE Full text] [doi: 10.1016/s2215-0366(14)70243-3]
130. Lee J, Lee CM, Park NK. Application of sensor network system to prevent suicide from the bridge. Multimed Tools Appl 2015 Dec 16;75(22):14557-14568 [FREE Full text] [doi: 10.1007/s11042-015-3134-z]
131. Looi M. The man helping prevent suicide with Google adverts. BMJ 2023 Dec 13;383:2847 [FREE Full text] [doi: 10.1136/bmj.p2847] [Medline: 38164638]
132. Mukherjee A, Ghosh B. An automated approach to prevent suicide in metro stations. In: Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications. 2016 Presented at: FICTA '16; September 16-17, 2016; Bhubaneswar, India p. 753-765 URL: https://doi.org/10.1007/978-981-10-3153-3_75 [doi: 10.1007/978-981-10-3153-3_75]
133. Lagan S, Sandler L, Torous J. Evaluating evaluation frameworks: a scoping review of frameworks for assessing health apps. BMJ Open 2021 Mar 19;11(3):e047001 [FREE Full text] [doi: 10.1136/bmjopen-2020-047001] [Medline: 33741674]
134. Torous J, Smith KA, Hardy A, Vinnikova A, Blease C, Milligan L, et al. Digital health interventions for schizophrenia: setting standards for mental health. Schizophr Res 2024 Apr 18;267:392-395. [doi: 10.1016/j.schres.2024.04.013] [Medline: 38640849]
135. Kozelka EE, Acquilano SC, Al-Abdulmunem M, Guarino S, Elwyn G, Drake RE, et al. Digital mental health and its discontents: assumptions about technology that create barriers to equitable access. Psychiatr Serv 2024 Mar 01;75(3):299-302 [FREE Full text] [doi: 10.1176/appi.ps.20230238] [Medline: 38050440]
136. Fricker M. Epistemic Injustice: Power and the Ethics of Knowing. Oxfordshire, UK: Oxford University Press; 2007.
137. Hawke LD, Sheikhan NY, Jones N, Slade M, Soklaridis S, Wells S, et al. Embedding lived experience into mental health academic research organizations: critical reflections. Health Expect 2022 Oct;25(5):2299-2305 [FREE Full text] [doi: 10.1111/hex.13586] [Medline: 35999670]
138. PARTNERS2 writing collective. Exploring patient and public involvement (PPI) and co-production approaches in mental health research: learning from the PARTNERS2 research programme. Res Involv Engagem 2020;6:56 [FREE Full text] [doi: 10.1186/s40900-020-00224-3] [Medline: 32974052]

139. Smith KA, Ostinelli EG, Ede R, Allard L, Thomson M, Hewitt K, et al. Assessing the impact of evidence-based mental health guidance during the COVID-19 pandemic: systematic review and qualitative evaluation. JMIR Ment Health 2023 Dec 22;10:e52901 [FREE Full text] [doi: 10.2196/52901] [Medline: 38133912]

Abbreviations

CDC: consensus development conference
CHR-P: clinical high-risk state for psychosis
FDA: Food and Drug Administration
FEP: first-episode psychosis
GALENOS: Global Alliance for Living Evidence on Anxiety, Depression, and Psychosis
LMIC: low- and middle-income country
PPI: patient and public involvement
PPIE: patient and public involvement and engagement
SMI: severe mental illness

Edited by T de Azevedo Cardoso; submitted 07.02.24; peer-reviewed by K Ludlow, S Stroup, N Josifovski; comments to author 23.02.24; revised version received 08.03.24; accepted 21.03.24; published 08.05.24.

Please cite as:
Smith KA, Hardy A, Vinnikova A, Blease C, Milligan L, Hidalgo-Mazzei D, Lambe S, Marzano L, Uhlhaas PJ, Ostinelli EG, Anmella G, Zangani C, Aronica R, Dwyer B, Torous J, Cipriani A
Digital Mental Health for Schizophrenia and Other Severe Mental Illnesses: An International Consensus on Current Challenges and Potential Solutions
JMIR Ment Health 2024;11:e57155
URL: https://mental.jmir.org/2024/1/e57155
doi:10.2196/57155
PMID:38717799

Original Paper

Understanding the Impacts of Online Mental Health Peer Support Forums: Realist Synthesis

Paul Marshall[1], PhD; Millissa Booth[1], MSc; Matthew Coole[2], PhD; Lauren Fothergill[3], PhD; Zoe Glossop[1], BSc; Jade Haines[4], BSc; Andrew Harding[3], PhD; Rose Johnston[1], BSc; Steven Jones[1], PhD; Christopher Lodge[1], MSc; Karen Machin[5], MPhil; Rachel Meacock[6], PhD; Kristi Nielson[3], MPH; Jo-Anne Puddephatt[7], PhD; Tamara Rakic[1], PhD; Paul Rayson[2], PhD; Heather Robinson[1], PhD; Jo Rycroft-Malone[8], PhD; Nick Shryane[9], PhD; Zoe Swithenbank[3], MSc; Sara Wise[4], RN; Fiona Lobban[1], PhD

[1]Spectrum Centre for Mental Health Research, Division of Health Research, Lancaster University, Lancaster, United Kingdom
[2]School of Computing and Communications, Lancaster University, Lancaster, United Kingdom
[3]Division of Health Research, Lancaster University, Lancaster, United Kingdom
[4]IT Corporate Services, Berkshire Healthcare NHS Foundation Trust, Berkshire, United Kingdom
[5]Survivor Research Network, London, United Kingdom
[6]Division of Population Health, Health Services Research & Primary Care, University of Manchester, Manchester, United Kingdom
[7]Department of Psychology, Edge Hill University, Ormskirk, United Kingdom
[8]Faculty of Health and Medicine, Lancaster University, Lancaster, United Kingdom
[9]Social Statistics, University of Manchester, Manchester, United Kingdom

Corresponding Author:
Paul Marshall, PhD
Spectrum Centre for Mental Health Research
Division of Health Research
Lancaster University
Health Innovation One
Lancaster, LA1 4YW
United Kingdom
Phone: 44 01524 522187
Email: p.marshall4@lancaster.ac.uk

Abstract

Background: Online forums are widely used for mental health peer support. However, evidence of their safety and effectiveness is mixed. Further research focused on articulating the contexts in which positive and negative impacts emerge from forum use is required to inform innovations in implementation.

Objective: This study aimed to develop a realist program theory to explain the impacts of online mental health peer support forums on users.

Methods: We conducted a realist synthesis of literature published between 2019 and 2023 and 18 stakeholder interviews with forum staff.

Results: Synthesis of 102 evidence sources and 18 interviews produced an overarching program theory comprising 22 context-mechanism-outcome configurations. Findings indicate that users' perceptions of psychological safety and the personal relevance of forum content are foundational to ongoing engagement. Safe and active forums that provide convenient access to information and advice can lead to improvements in mental health self-efficacy. Within the context of welcoming and nonjudgmental communities, users may benefit from the opportunity to explore personal difficulties with peers, experience reduced isolation and normalization of mental health experiences, and engage in mutual encouragement. The program theory highlights the vital role of moderators in creating facilitative online spaces, stimulating community engagement, and limiting access to distressing content. A key challenge for organizations that host mental health forums lies in balancing forum openness and anonymity with the need to enforce rules, such as restrictions on what users can discuss, to promote community safety.

Conclusions: This is the first realist synthesis of online mental health peer support forums. The novel program theory highlights how successful implementation depends on establishing protocols for enhancing safety and strategies for maintaining user engagement to promote forum sustainability.

Trial Registration: PROSPERO CRD42022352528; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=352528

(JMIR Ment Health 2024;11:e55750) doi:10.2196/55750

KEYWORDS

digital mental health; peer-to-peer support; social networking; moderation; systematic review

Introduction

Background

The World Health Organization recently identified poor supply of services as a primary barrier to mental health care worldwide [1]. Evidence-based digital approaches may alleviate demands on existing services and help meet the rising need for accessible models of psychosocial support [2]. Online mental health peer support forums may represent one such approach. Forums allow users to engage in asynchronous, text-based communication with those who share similar experiences. Common functions include the ability to start discussions, or threads, and post messages within these threads in response to others' comments [3]. This creates opportunities to exchange information, advice, and emotional support, often within moderated online environments. Forums may focus on certain topics, such as specific mental health diagnoses, or on mental health and well-being more broadly [4].

Given that forum websites typically remain available 24 hours per day, they hold the potential to deliver accessible mental health support at scale. However, evidence of their effectiveness is mixed. While some trials of digital interventions, including forums, have shown positive effects, such as improved mood, mindfulness, and compassion [5,6], others have shown no significant impact on user well-being [7,8]. This is consistent with a review of online peer support for young people with mental health difficulties in which just 2 of 6 identified studies reporting positive effects on anxiety or smoking behaviors [4]. Factors impacting these findings include challenges with retention and engagement in online interventions [8,9], which may not reflect the way users engage with forums outside the context of intervention studies. Many such interventions include forums along with other components, such as psychoeducational materials, confounding attempts to identify specific impacts of forum use. Therefore, at present, the reasons why some online forums are more conducive to positive user experiences than others remain unclear.

Processes underpinning observed impacts of online mental health forums are likely to be multifaceted. There is qualitative evidence suggesting that some forum users derive benefit from the social connection offered by online forums [10-12]. Indeed, a conceptual model of online peer support for severe mental illness emphasized the value of online interactions for stigma reduction and increasing participants' willingness to engage with in-person support [13]. The option to participate anonymously, a feature of many online forums, may reduce fear of judgment and promote personal disclosure related to mental health difficulties [14,15]. However, some users report that reading about mental health on the web can be distressing, and concerns have been raised about the potential of forums to proliferate harmful content [16,17]. Furthermore, it is currently unclear whether the impacts of online forums are influenced by differences in forum implementation and moderation across online contexts. Some forums are established and moderated by people with personal experience of mental health difficulties [18]. Others are delivered by mental health care providers and staffed by formally trained moderators who, depending on the service delivery model, may be health professionals or volunteers [19,20].

Objectives

There is evidence that the use of online mental health forums is growing. For example, the open Reddit discussion board "r/depression" expanded from 314,000 users at the end of 2017 to approximately 1 million in late 2023 [21]. Increased support seeking in mental health forums in response to the COVID-19 pandemic [22] also emphasizes the need for further research on the contemporary use, safety, and effectiveness of these services. Previous reviews have identified and described mental health forum user experiences. For example, it is clear that online forum–based interventions for mental health are feasible and viewed as acceptable by most users and meet some users' needs for informational and social support [23-25]. However, there have been recent calls for mechanistic research to better understand the processes that underpin the effects of online forums to inform innovations in forum design and implementation [26]. This study aimed to address this gap by applying realist synthesis to generate an explanatory model, or program theory, explaining the impacts of online mental health forums on users.

Methods

This study was preregistered on PROSPERO (registration CRD42022352528) and is reported with reference to the Realist and Meta-Narrative Evidence Syntheses: Evolving Standards guidelines [27]. The synthesis progressed through 5 stages informed by realist methodological guidance [28].

Stage 1: Define the Scope of the Synthesis

We sought evidence of the use of online mental health forums as per the definitions in Textbox 1. Instead of searching for evidence published from 1993 onward as per our research protocol, we restricted our evidence searches to between 2019 and 2023. This increased the likelihood that the included studies would reflect a similar context to the current program delivery environment, including online harms, technological functionality, and contemporary health service environments, which may serve to promote the transferability of our findings to current and future stakeholders. Our decision to search from 2019 onward was informed by a recent comprehensive review of digital health interventions that identified both a shift in the online context of health service delivery as a result of the COVID-19 pandemic and a notable increase in relevant literature in the same year [29]. It is also likely that pre-2019 articles are captured in related systematic reviews [23,24]. The scope of this synthesis was also informed by engagement with a public and patient involvement group comprising forum staff and users. The group participated in a stakeholder prioritization workshop that involved ranking and discussing the importance of research areas related to online forums. This influenced the subsequent analysis, as described in stage 5.

Textbox 1. Definitions of key terms.

Key terms and definitions used in this synthesis

- *Mental health peer support forum*—forums were defined as on the web, primarily asynchronous text-based discussion platforms. All forum types were eligible, including those hosted on widely used social media platforms such as Reddit and those delivered as part of specifically designed interventions. To be included in this synthesis, articles must have studied a mental health peer support forum. Mental health forum was broadly defined to include any forum primarily intended to support people experiencing psychological distress, including those with specific mental health difficulties, experiencing isolation or substance misuse or addiction, or caregiving for someone with a mental health difficulty. To be eligible for inclusion, evidence sources must have described forums focused on facilitating peer-to-peer discussion.
- *User*—forum users are people who access online mental health forums to seek support for their own psychological well-being or in the capacity of an informal supporter of someone experiencing mental health difficulties. Activities in which users typically engage in online forums include reading posts, starting their own discussion topics, and responding to other users.
- *Moderator*—moderators are staff or volunteers with oversight responsibilities within online forums. While these roles vary across contexts, they typically include facilitating discussions within forums, moderating content, enforcing rules, and providing support directly to users by responding to their posts. Any moderation type was eligible, including moderation by volunteers, health professionals, or community members with personal experience of the mental health difficulties discussed on the forum. Forums with no peer-to-peer discussion, such as those in which service users interact solely with mental health professionals, were not eligible for inclusion.
- *Program*—the intervention or service under investigation, which in realist research is typically a health or social care program [28]. In this synthesis, "program" refers to online mental health forums.
- *Context*—a factor that determines the activation or triggering, or the strength of activation or triggering, of a given mechanism. Contexts are varied and can include factors at psychological, social, economic, and institutional levels [28].
- *Mechanism*—the hidden force that causes a program to work, defined as the ways in which a participant responds to a program [28]. We used the distinction between mechanism resource (what is offered by a program) and reasoning (how users respond to what is offered) to further elaborate our analysis [30].
- *Outcome*—the expected and unexpected results of a program [28].
- *Context-mechanism-outcome configuration* (CMO)—a CMO is a heuristic used in realist research to articulate causal insights regarding how a particular program generates outcomes, with reference to the operation of mechanisms in specific contexts [27].
- *Program theory*—an explanation of how the program under investigation works. In this synthesis, CMOs represent individual program theories. We also outline an "overarching program theory," which refers to the integration of individual program theories to create an explanatory model of how the entire program under investigation operates.

Stage 2: Develop Initial Program Theories

Initial program theories are preliminary accounts of how a program is expected to work, which are subsequently refined through the synthesis process [28]. To support initial program theory development, we held a multistakeholder workshop comprising forum users, staff, and researchers with expertise in digital mental health. Attendees completed a group activity focused on designing a hypothetical forum and explored its potential impacts, mechanisms, and challenges. The research team used these discussions to generate a list of initial ideas of how forums work, including a series of "if, then" statements used to support causal reasoning in the early stages of realist research [31]. An overview of the ideas and initial program theories generated is presented in Multimedia Appendix 1.

Stage 3: Search for Evidence

Searches were conducted on the following health and social science databases from January 2019 to May 2023: PsycINFO, MEDLINE, CINAHL, Academic Search Ultimate, Embase, Scopus, and Web of Science. The search strategy was developed in collaboration with an information specialist at Lancaster University and was informed by a sensitivity analysis. This involved checking whether searches returned key articles previously identified by the study team as relevant to the research question. Database searches were supplemented with gray literature searches on Google, the TRIP medical database, Overton, the International Clinical Trials Registry, the National Grey Literature Collection, ProQuest, and the National Health Service Knowledge and Library Hub. The full search strategy is available in Multimedia Appendix 2 [5,13,16,32-38].

Stage 4: Selection and Appraisal

Documents were assessed for eligibility against the following criteria: (1) documents referred to a peer online mental health forum as per the definition in Textbox 1; (2) documents referring to users who were adults or young people, defined as >50% of participants being aged ≥13 years; (3) full texts available in English; and (4) any document type or study design.

Documents were ineligible for inclusion if they met any of the following criteria: (1) documents focused on a online platform that did not primarily support asynchronous text-based group discussion. Examples include interventions whose primary functions were direct instant messaging, live chat, or image sharing; (2) documents focused on an intervention principally aimed to support the self-management of a physical health difficulty and did not have an explicit focus on psychological distress; and (3) documents focused on an intervention aimed at supporting the practice of mental health professionals.

Titles and abstracts were screened in the web-based systematic review platform Rayyan (Rayyan Systems Inc) [39]. Screening was completed by a team of 9 researchers who initially independently screened 100 articles and then met to identify discrepancies and refine the screening procedure. Each team member screened a separate batch of articles. The team met weekly and made decisions on articles collaboratively where ambiguity existed regarding their eligibility. Articles that passed title and abstract screening were reviewed in full by a second reviewer to confirm their eligibility against the inclusion criteria, during which articles were appraised with reference to realist-informed principles of rigor and richness [40], where evidence sources are judged with respect to their relevance to theory development. We applied an inclusive assessment of rigor based on the "good enough" test [41]. Therefore, articles were included if they were sufficiently transparent to allow the reader to understand how the data had been generated and if they were credible given the methodology used. That is, the methods used in each evidence source were congruent with the results and conclusions drawn from them. To judge richness, we adapted the "traffic-light" system [42] for judging the usefulness of evidence sources for their potential contribution to program theory development, with screeners making judgments of low, moderate, or high usefulness. Highly useful evidence sources were eligible for inclusion in the synthesis. Judgments regarding rigor and usefulness were double-checked by a second reviewer before data extraction. The full-text screening instructions are presented in Multimedia Appendix 3 [40,43].

Stage 5: Data Extraction and Synthesis

Data were extracted describing key study characteristics, including year of publication, title, authors, and a description of the forum and moderation. We then extracted data relevant to program theory development. To do this, researchers copied data segments from eligible documents into an Excel (Microsoft Corp) spreadsheet and added an analytical code. Analytical codes articulated how the data segment contributed to program theory development, with a specific focus on identifying contexts, mechanisms, and outcomes as per the definitions in Textbox 1. Data analysis ran in parallel. This initially involved developing candidate context-mechanism-outcome configurations (CMOs) informed by concepts identified in the initial stakeholder workshops. CMOs were then iteratively refined by integrating insights from data segments and analytical codes obtained from the data extraction process. To promote analytical rigor, a core team of analysts including the lead interviewer (PM) and researchers with expertise in online mental health forum research (FL and HR) and realist methods (AH) met regularly to review extracted data and collaboratively develop each individual CMO in group discussion. The CMOs were reviewed by the wider research team, comprising academics, clinicians, forum moderators, and lived experience researchers with broad experience in digital mental health and peer support.

Data from the existing literature identified in the database searches predominantly focus on the experience of forum users, with many articles using qualitative methods, including interviews. Therefore, we supplemented this literature with exploratory interviews with other key stakeholders likely to provide insights into the positive and negative impacts of online mental health forums (the topic guide and participant demographics are available in Multimedia Appendix 4). Participants were recruited from UK institutions involved in the delivery of online mental health forums as part of a funded project to investigate mental health forum use in the United Kingdom [44]. Participants were purposively sampled via email advertisements circulated to forums partnered with the project. They included "forum hosts" (n=5), who were clinicians with oversight roles within organizations that host mental health forums and clinical academics who have designed and delivered online forum–based interventions, and forum moderators (n=13). The sample size was informed by both the research team's judgment of data sufficiency in answering the research question and a pragmatic consideration regarding the resources available for this study The interviews focused on articulating participants' views on the impacts of online forum use (outcomes), the processes underlying these impacts (mechanisms), and the factors influencing the presence of those processes or the extent to which they occurred (contexts). The interviews were conducted in parallel to data extraction from eligible studies and used the same analytic procedure. That is, transcribed data were copied into the data extraction document with a code identifying how the data informed CMO development.

Individual CMOs were then grouped thematically into 5 theory areas reflecting the research priorities set in the stakeholder prioritization workshop. These were how learning in forums benefits user mental health (theory area 1—mental health self-efficacy), factors impacting safety (theory area 2—psychological safety), factors impacting the use of forums and other services (theory area 3—service use), the role of forum moderators (theory area 4—moderation), and how connecting with others online impacts user well-being (theory area 5—social connection). Following CMO refinement, figures were generated indicating links between program theories to provide overarching explanations for how forums generate intended and unintended impacts. Finally, 3 members of a public and patient involvement group were invited to individually review and provide feedback on an interim version of the analysis before the final write-up. This group is facilitated by the research team and comprises forum users, staff, and people with lived experience of mental health difficulties. Therefore, group members had a previous relationship with the research team and knowledge of the aims of this study.

Ethical Considerations

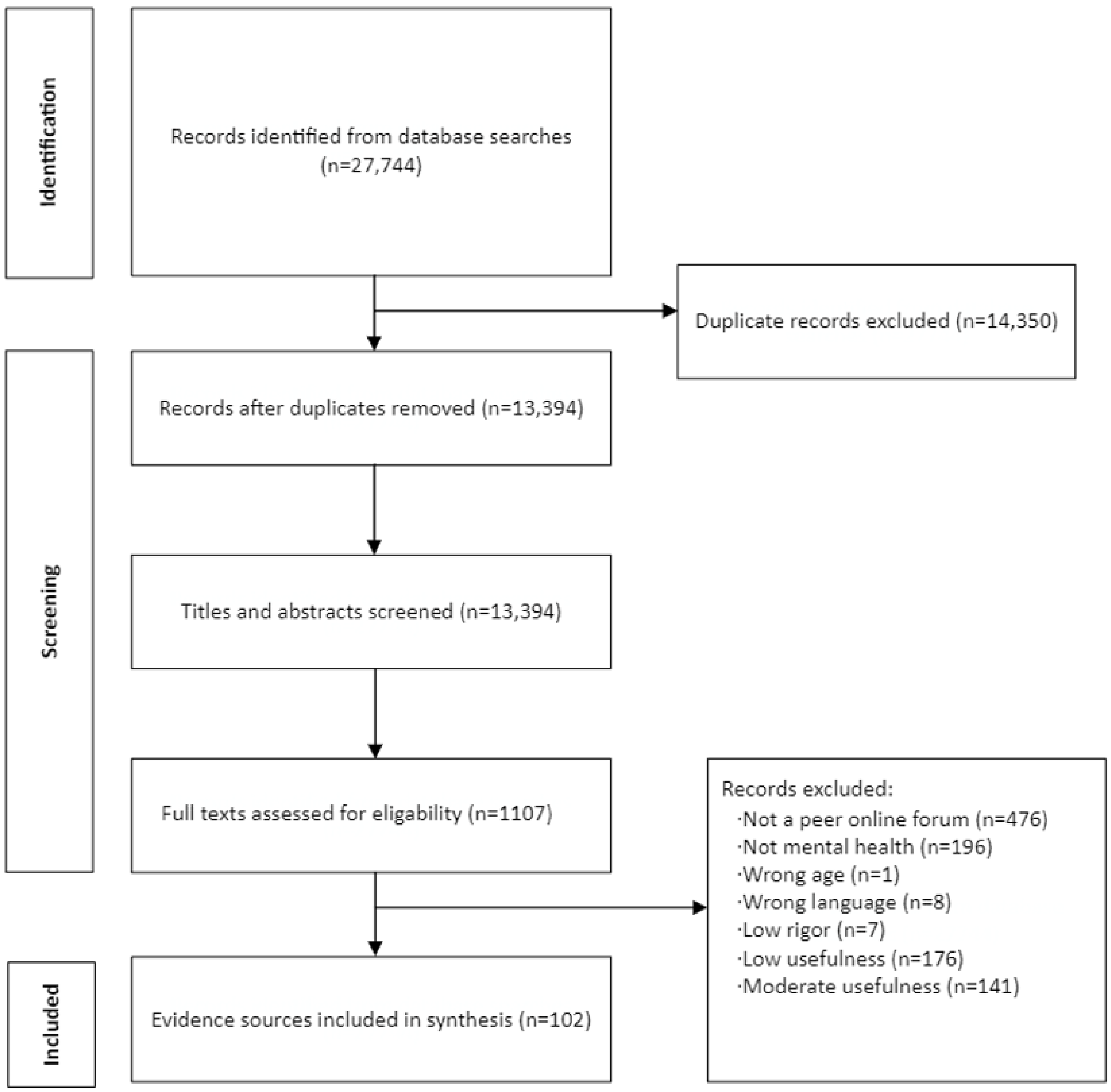
This study received ethical approval from Solihull Research Ethics Committee on 20 June 2022 (IRAS314029).

Results

Included Evidence Sources

The process of evidence identification (Figure 1) led to the inclusion of 102 documents. The characteristics of the included evidence sources are presented in Multimedia Appendix 5 [18,25,32,33,44-81].

Figure 1. Process of evidence identification.



Program Theories

Data analysis produced 22 CMOs (Textbox 2) in 5 theory areas. Figures 2 and 3 present overarching program theories representing intended and unintended impacts, respectively. A narrative summary of the program theories is presented in this section with reference to illustrative evidence. Quotes with citations are from evidence sources identified in the literature search. Quotes from interviews are indicated by the participant's role and number, for example, forum moderator 1.

Textbox 2. Context-mechanism-outcome configurations (CMOs) by theory area.

Theory area 1: mental health self-efficacy

- CMO 1: in well-populated and active forums that are clearly organized (context) to allow users to find posts and receive responses that are personally relevant (mechanism—resource), users will be more likely to identify credible and actionable information that they can use to better manage their mental health (mechanism—reasoning), promoting mental health self-efficacy (outcome).
- CMO 2: when users feel safe to share their mental health experiences (context) with others whom they perceive to be nonjudgmental and as having relevant experiences (mechanism—resource), they will use the forum to reflect on their circumstances and integrate others' perspectives into their own (mechanism—reasoning), resulting in novel and more hopeful ways of making sense of their mental health experiences (outcome).

Theory area 2: psychological safety

- CMO 3: posts detailing personal experiences of potentially harmful behaviors (eg, self-injury and restrictive eating; context) that frame them as helpful (mechanism—resource) may normalize and reinforce their use (mechanism—reasoning), increasing the likelihood of users adopting these behaviors (outcome).
- CMO 4: when seeking support for issues that others may find distressing (context), users are more likely to post in forums that have ways to flag the potentially distressing nature of their experiences (eg, tags, trigger warnings, or a separate subforum; mechanism—resource). This provides reassurance that posts will not inadvertently cause harm to others (mechanism—reasoning), increasing the likelihood that users will use the forum to seek support (outcome). Other users are less likely to be exposed to distressing content (mechanism—reasoning), reducing potential distress in the wider community (outcome).
- CMO 5: for users making an original post (context), the absence of a response or responses that are unrelated to the original post (mechanism—resource) will prompt feelings of being ignored or misunderstood (mechanism—reasoning). This leads to increased isolation (outcome) and reduces forum engagement (outcome).
- CMO 6: those yet to post to forums may be concerned about feeling exposed or receiving negative responses if they share their experiences (context). Observers who see others receiving constructive and respectful responses (mechanism—resource) will be reassured of the safety of posting to the forum (mechanism—reasoning), increasing the likelihood that they will actively participate in discussions (outcome).
- CMO 7.1: negative social consequences of discussing mental health difficulties, including shame and stigma (context), are overcome by forum anonymity (mechanism—resource), which disinhibits (mechanism—reasoning) users discussing their experiences, leading to greater self-disclosure (outcome).
- CMO 7.2: because users' personal identities are hidden (context), they are insulated from the negative social consequences of rule breaking (mechanism—resource). This can have a disinhibiting effect on some users (mechanism—reasoning), making them more likely to engage in antisocial behavior such as bullying (outcome), reducing safety for other users (outcome).
- CMO 8: open online forums with no ways to flag distressing content, poor moderation, or lenient rules (context) are more likely to expose users to posts detailing users' highly distressing circumstances, misinformation, and "toxic" discussions (mechanism—resource), which can contribute to distress (mechanism—reasoning) and disengagement from the forum (outcome).

Theory area 3: service use

- CMO 9: when users experience barriers to in-person mental health care, such as stigma, poor service availability, or living in a rural area (context), accessible online forums (mechanism—resource) are seen as convenient sources of support (mechanism—reasoning), increasing forum use in those experiencing these barriers (outcome).
- CMO 10: in cases in which forums are populated with people who have positive experiences of mental health services (context) and who share these experiences with a view to encouraging other users to seek help (mechanism—resource), readers will feel more confident in approaching those services (mechanism—reasoning), increasing use of other forms of mental health support (outcome).
- CMO 11: design features (mechanism—resource) that inhibit the autonomous and competent use of forum technology (mechanism—reasoning) decrease users' motivation for engagement (outcome), particularly in cases in which that technology is novel for individual users (context).
- CMO 12: when users who are yet to seek alternative (non–forum-related) mental health support (context) are exposed to negative comments about those sources of support, such as mental health services (mechanism—resource), they will be more skeptical of the potential value of those services (mechanism—reasoning) and, therefore, less likely to approach them (outcome).
- CMO 13: in cases in which a supportive online community (context) provides the emotional and informational support a user requires (mechanism—resource), that user will feel that their needs are met sufficiently by that community, reducing the perceived need for alternative support (mechanism—reasoning) and, therefore, the use of other mental health services (outcome).

Theory area 4: forum moderation

- CMO 14: when initially accessing an online forum (context), friendly support with how to use the site and the presence of moderators who are seen to promote positive engagement (mechanism—resource) generates confidence in using the forum (mechanism—reasoning), increasing subsequent engagement (outcome).
- CMO 15: in cases in which forums are moderated and users post to the forum (context), moderator responses that are timely, show empathy and understanding, are personalized to the content of users' original posts, and invite further discussion (mechanism—resource) will lead users to feel heard and supported (mechanism—reasoning), prompting further engagement with the forum (outcome).
-

CMO 16: when forum moderators intervene in forum discussions to restrict or delete users' rule-breaking posts (context), doing so in a way that demonstrates consistency and makes site rules clear (mechanism—resource) will mean that users view their actions as fair and unobtrusive (mechanism—reasoning), promoting trust and safety among the wider user base (outcome).

- CMO 17: when forums have low tolerance for discussions of potentially distressing issues (context), moderators are more likely to delete comments referencing related topics such as self-injury (mechanism—resource). While this may promote a sense of safety (mechanism—reasoning) and engagement (outcome) for some users, those whose posts are deleted may feel that this action infringes on their autonomy and ability to seek support (mechanism—reasoning), prompting attempts to avoid moderation (eg, by tangential references to banned material; outcome) or seek support in less restrictive forums (outcome).

Theory area 5: social connection

- CMO 18: when forums bring together people with similar personal experiences (context), users have access to posts that resonate with their circumstances (mechanism—resource). This normalizes their mental health experiences and validates their own reactions to similar situations (mechanism—reasoning). This can reduce self-stigma (outcome) and provide a sense of belonging (outcome).
- CMO 19: when forums provide users with a reliable source of support (context), the ability to interact with the community when needed (mechanism—resource) decreases users' reliance on in-person informal support (mechanism—reasoning), reducing perceived burdensomeness on friends and relatives (outcome).
- CMO 20: users who share their personal experiences on the web (context) derive satisfaction (outcome) from the knowledge that their posts help others (mechanism—reasoning), particularly when others express gratitude (mechanism—resource).
- CMO 21: when forum users post messages (context) and receive timely, constructive, and empathetic responses from other users (mechanism—resource), they will feel recognized and understood (mechanism—reasoning). This will contribute to a sense of connection (outcome) with the online community (outcome), increasing forum engagement (outcome).
- CMO 22: when users disclose lived experience (context), other users are more likely to view the user as authentic (mechanism—reasoning). This makes users more likely to share their own experiences in response, generating reciprocal and mutually supportive conversations and relationships (outcome).

Figure 2. Overarching program theory—intended impacts of online mental health peer support forums.

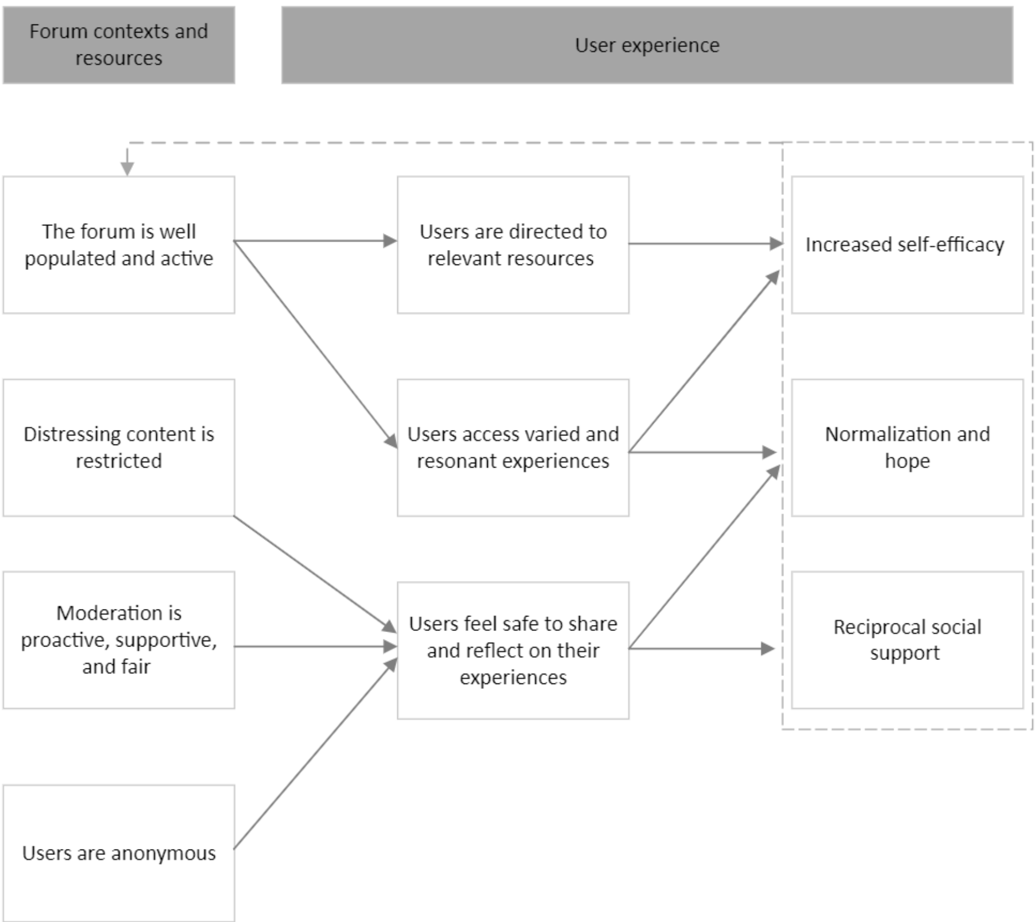
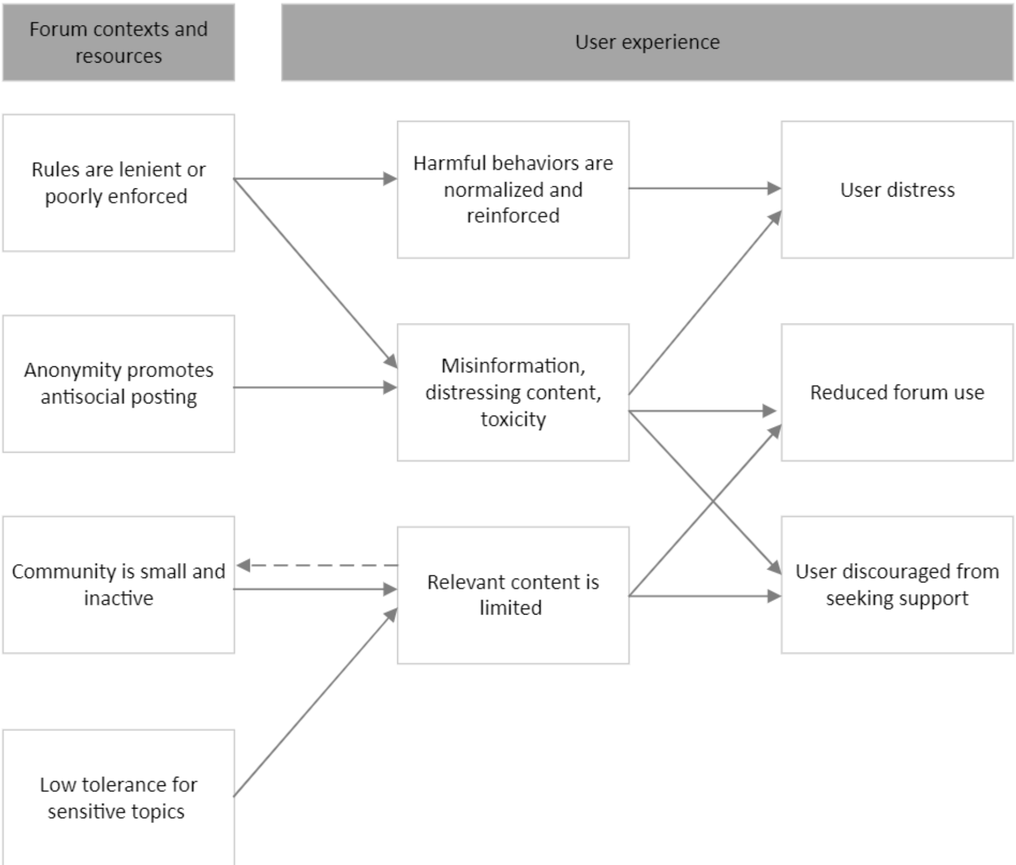


Figure 3. Overarching program theory—unintended impacts of online mental health peer support forums.



Theory Area 1: Mental Health Self-Efficacy

Promoting mental health self-efficacy, a belief in one's own capacity to manage mental health, is a common goal of forum-based interventions [45,82]. This is enabled by providing users with access to credible and actionable information of personal relevance (CMO 1). Examples include lived experience expertise regarding how to manage certain situations, such as caring for a relative in distress, and techniques users can apply to support their own recovery [46,47]:

> *...when they share information about coping strategies, I "cherry-pick" the things that suit me most.*

Users may identify this information in the repository of existing threads or post their own requests for information on specific issues. Finding relevant information and receiving helpful responses relies on an active community that generates new content:

> *...we did also run the project in small Trusts [regional health services] and we tried to have forums in the Trust, and it didn't work at all and it's basically because you need a big enough population that there's always activity otherwise people won't go to it.* [Forum host 1]

As one forum host noted, grouping users into forums with a narrow focus can limit the breadth of information generated:

> *...you want a 24/7 digital, vibrant community where any concern that someone brings will be mirrored by somebody else or at some stage in life...to avoid running people into dead ends, very, very narrow, let's say, identifications, we try to keep the groupings as large as possible because that's where you get that diversity of experience and perspective.* [Forum host 2]

CMO 2 highlights the value of collaborative sense making within forum conversations. In the context of a safe and nonjudgmental online space, peer discussion facilitates an exploration of personal difficulties through the lens of others' experiences [32,48,83]. Such discussions allow for a broadening of users' understandings of mental health, which can serve to reframe their own [49] by, for example, reconceptualizing a shared difficulty as something that can be managed [32]:

> *I remember there was one guy that used to say, it does get better. It gets a lot better, and I know that this is an awful feeling now, but it does get better and that was really big. That was life-changing.*

Therefore, forums offer opportunities to reflect on and experiment with new personal narratives, a process that can lead to greater clarity and self-awareness [47]:

> *...sharing experiences you get different viewpoint and more insight, this makes you think more seriously, like, "ah, that could be the same for me, or maybe that's a pitfall for me too."*

Exposure to others' perspectives, particularly those describing progress in managing mental health, could lead users to adopt more positive expectations [46]:

> *...posters glimpsed the possibility of recovery for the first time; this prompted the initial act of hoping as they came to believe that this transformation was possible for themselves.*

Theory Area 2: Psychological Safety

Several articles identified a sense of safety as foundational to positive forum experiences [32,50,83]. Aspects of psychological safety may include the confidence to use the platform without being exposed to distressing content or negative judgment. Evidence sources also identified factors that could undermine this sense of safety. For example, some communities allow users to share details of potentially harmful and distressing experiences, including self-injury and restrictive eating (CMO 3) [51-53]. Descriptions of these behaviors may frame them as helpful and understandable ways to manage emotional distress, creating a culture that normalizes and encourages their use. This has been described as a "functional paradox" [53], where those in suicidal distress seek a form of social support that increases the risk of adverse events or where those stigmatized by issues such as eating disorders look for understanding peer communities that perpetuate the issues for which users require support [52].

While some users derive benefit from simply reading online forums [54], others may desire interaction but not feel sufficiently comfortable to engage in conversation directly. For example, some users are reluctant to post about their own distressing circumstances for fear of causing harm to others in the community (CMO 4) [55]:

> *I guess I didn't even post things, like thoughts that I might be having...life's not worth living, I feel so hopeless, I feel like I don't have a future. I wouldn't say stuff like that, because I thought it would be too triggering for other people.*

This may be mitigated by the inclusion of subsections or content notices to allow potential readers to avoid harmful content and promote community safety [84]. Moreover, receiving no reply can cause frustration and undermine the sense of connection that draws people to forums (CMO 5) [55,56]. Seeing others receive timely and encouraging responses can promote confidence in the forum and help those considering posting become active forum participants (CMO 6) [57].

There is broad recognition that anonymity facilitates dialogue within online forums (CMO 7.1) [47,53,58-60]. Online anonymity alleviates concerns linked to in-person help seeking [82]:

> *The desire for anonymity was often a result of fear and stigma and could be more prominent in certain communities and cultures....*"*You are anonymous. And you can leave when you want, whereas if you go to a group you tend to be there for at least a polite amount of time.*"

Therefore, forums provide an outlet for users to express their "true self" [61] and broach topics that may be difficult or seemingly impossible to address in person:

> *One lady disclosed domestic abuse. She hadn't disclosed it to her midwife, GP, health visitor who'd she'd all seen in the flesh etcetera, but she felt safe enough to disclose it to us because it was anonymous, she was anonymous.* [Forum moderator 1]

However, this same feature can lead to different user responses. Therefore, CMO 7.2 represents a rival theory [31], stating that anonymity may render interactions impersonal and limit consequences for "toxic" or "trolling" behaviors [61], including hostile comments and harassment [18,62]:

> *I also volunteer on a different online community with anonymity and I find there's more conflict of people thinking you can literally say whatever and not so supportive.* [Forum moderator 1]

Harmful content and antisocial behaviors may be more likely to occur in communities that have little or no moderation (CMO 8). Clear and visible rules for what cannot be discussed on the forum, proactive monitoring for rule-breaking content, and the restriction of users' access to forums where rules are broken are important for promoting a supportive culture and mitigating risks to the wider user base [53]. Such risks may include being exposed to stigmatizing messages, misinformation, or highly distressing content [63]. Cultivating a community characterized by supportive communication that also provides the opportunity for users to openly express their difficulties has been described as a "balancing act" for moderators [64]. Rules need to be sufficiently stringent to limit harmful posting but not to the extent that they are a deterrent to engagement:

> *I think guidelines are important. I think they need to be easy to understand and not make it feel like, not make someone feel scared for breaking a rule sort of thing.* [Forum host 1]

A related concern for promoting forum safety relates to the potential cumulative effect of being exposed to a large volume of particularly distressing posts, which can be detrimental for users [47]:

> *Well, you can feel overwhelmed by it all. I had mixed feelings. On the one hand, I felt relief because I could share my experiences. But on the other hand...all the new posts—I thought, "This is not good. I'm too preoccupied with the forum and worry too much about others right now."*

Theory Area 3: Service Use

Motivations for using online peer support forums provide insights into what works for whom (CMO 9). Motivations include a lack of understanding from friends or family members and associated challenges with in-person help seeking [32,61,65], such as fear of judgment from social contacts or health services. Reflecting on an eating disorder support forum, one participant noted the following [52]:

> *...we need a place where we can feel accepted, appreciated, and safe. And this is it.*

Indeed, limited access to in-person sources of support can lead some users to engage with online communities [32,66]. Alternatively, users who are yet to seek in-person support can be empowered to do so through interactions with peers (CMO 10). Therefore, forums may represent a "stepping stone" [47] to further support, which can result from direct encouragement, the sharing of specific advice regarding how to access support, and a growing confidence with discussing mental health with others [67-69]. Reflecting on their own forum use, a moderator noted the following:

> *I would never have got into services had it not been from peer support online with people encouraging me to tell my parents what was going on, tell my teachers, go to the GP, helping me to even write the letter that I wrote to my GP expressing what was going on.* [Forum moderator 2]

Forum design that supports competent self-directed use, such as easy-to-use interfaces, is an important determinant of ongoing service engagement [70] (CMO 11). Features users may find beneficial include easy navigation, app integration, and the inclusion of capabilities such as emojis [58]. Conversely, frustrating technological issues can undermine perceived convenience:

> *...having a community that's easy to use is important so I think some feedback we've had...when you first join our community, we've got so many subforums that it looks quite busy and overwhelming so something we want to do is condense it a little bit.* [Forum host 3]

There are also contexts in which forum use may lead users to be less likely to seek in-person support. The stakeholder workshop highlighted the possibility that forum use may alleviate pressures on health services by meeting users' needs without them requiring other support (CMO 12). However, in some circumstances, users may be exposed to posts describing unsuccessful attempts to access services or negative experiences in health care that could discourage readers from seeking support (CMO 13) [62,71], an issue noted by one interviewee:

> *...it could be quite triggering if someone's having an experience that's very similar to yours and it could be quite depressing I suppose just thinking, "god, I'm struggling but these people are struggling even more..." No one's got an answer to this. Quite hopeless. I think it could lead to a bit of hopelessness about how bad the system is...* [Forum host 1]

Theory Area 4: Forum Moderation

Moderators play important roles in users' early forum experiences, where their guidance helps users engage with the community and sets expectations for what the forum can provide (CMO 14) [25,70]. Being consistently visible can offer reassurance that interactions will remain positive and supportive, as one moderator highlighted:

> *I think also what's important is moderator presence, so they know that we're around and looking after the community and kind of replying to reports quite quickly I think yeah it just helps people feel safe, able to reach out.* [Forum moderator 3]

The style and content of moderators' online posts are likely to influence users' satisfaction and ongoing engagement with the forum (CMO 15). As identified by one forum host, effective moderators achieve a balance between sharing relevant information and offering understanding:

> *...it was a very skilled approach, so it was thoughtful, it was deep—extensive. It was informed and it was sort of tapping into both the empathic side of it and the need for information, so it wasn't just providing information and it wasn't, "Oh I'm sorry you feel like that," I think there was a really good balance between empathy and information.* [Forum host 2]

Users also valued timely responses, which both helped address current difficulties [48,82] and mitigated against forum users disengaging from the community:

> *I think there is data as well that says like 50% of people don't come back if they haven't had a reply within 24 hours.* [Forum host 3]

When removing content that contravenes forum rules, moderators must balance a desire to facilitate open peer discussion with a need to ensure the appropriateness of the content shared on the platform [32]. Making decisions transparent and consistent helps establish boundaries for what users can expect to do and see within a forum and may mitigate the risk that users feel unfairly treated (CMO 16) [55]. However, moderators may face challenges when implementing rules related to potentially harmful behaviors, such as self-injury or restrictive eating (CMO 17). Users affected by these issues may feel that their need for support is undermined by strict content policies, particularly in cases in which online forums represent one of the few safe spaces to seek help. Regarding an eating disorder forum banned by a host website, a former user recalled the following:

> *...someone had posted what to do if you feel like you're going to binge, what to do if you feel like you can't eat today. I would go and read that actively, like, "Oh, here's some reminders for myself," and now it's gone, and I can't access that.*

Restricting community discussion in this way can lead to frustration [55]; isolation [72,84]; and attempts to navigate these boundaries, for example, limiting how open a user is about their difficulties [32] or using novel terminology to overcome restrictions on what can be discussed [51]. While such moderator decisions are often made in the interest of the wider community, individual support may be deprioritized:

> *I've worked for a few other online communities...they kind of banned people right away if they were in crisis which I found really difficult to do because it didn't feel very fair that they were reaching out for support and they get banned instantly.* [Forum host 3]

This illustrates how organizations and moderators negotiate an often sensitive and challenging responsibility to balance the interests and safety of individuals against impacts on the wider community.

Theory Area 5: Social Connection

The social connection offered by online mental health forums facilitates several distinct positive impacts. Others' accounts of similar circumstances validate users' own thoughts and behaviors [73], contributing to a recognition that they are not isolated in their mental health experiences (CMO 18) [62]:

> *...knowing you aren't alone, that you're not crazy or lazy, that other people go through the same thing every day, is a strangely comforting thing to experience.*

This normalizing experience can reduce self-stigma and blame [46]:

> *I've always felt that my responses to what had happened seemed abnormal and crazy, so I feel reassured reading that it's okay to react the way I did. I can now work on finding healthier solutions to my issues.*

Other elements of user responses likely to promote perceived social support include them being offered in a timely way when users most need help [47,68], constructive suggestions for problem-solving [45,74], and empathetic communication [32,75] (CMO 19). For example, in one forum, this support was represented by statements such as "I'm glad you're here," "So much love to you," or "This whole thing is a nightmare." Emotional support was also provided relatively prominently in the form of encouragement, which would frequently occur in short interjections such as "you can do it, mama" or "There IS light at the end of the tunnel!!" [76]. Community members may derive benefit from offering such support (CMO 20) [50,55,77], particularly in cases in which users are motivated to support others through situations that they have personally experienced [62]:

> *I find great joy in helping others find resources and helping them to learn about this condition and its comorbid conditions, as well as relating my personal experiences to theirs so that they, in turn, don't feel so alone.*

Within forum threads, the presence of personal narratives promotes the authenticity of users' requests for support and prompts reciprocal and mutually beneficial sharing in response (CMO 21) [78,79]. In this way, relationships develop within the community, improving users' perceptions of being socially supported [47]:

> *...I was trying to focus on solving my own problems until I saw that users were helping each other. I realized I could also benefit from their support. I began typing up my personal story. I got positive replies and then also started to respond to others.*

This process not only occurs through conversations about mental health, but also, as with in-person relationships, connections develop through exchanging updates on daily life, venting frustrations, and discussing personal interests [50,83]. Providing an example of ongoing community support, one moderator recalled the following:

> *...they [users] come back to us and they say, "yeah, the problem I had with my friend, it's all sorted now. It's great," and all the young people, all their peers are like, "Ah that's so amazing. I remember your post and it sounded so awful. I'm so proud of you for sorting that out," so they get really positive validation from their peers.* [Forum moderator 4]

In cases in which online mental health forums provide users with a regular source of social support, their perception of being a burden on in-person social contacts may diminish (CMO 22) [32,47,80].

Discussion

Principal Findings

This paper presents a novel program theory highlighting the potential for safe and active online peer support forums to promote mental health self-efficacy through access to actionable information and the opportunity to explore personal difficulties with nonjudgmental peers (theory area 1). It points to the importance of psychological safety (theory area 2) in facilitating positive experiences and identifies barriers to safety, including exposure to distressing content and concerns about posting. Motivations for forum use include stigma and difficulties accessing in-person services, whereas the nature of forum experiences may shape users' perceptions of those wider services (theory area 3). Proactive and supportive forum moderation is important for creating a space for dialogue (theory area 4) where users can engage in mutual and reciprocal social support, which can lead to reduced isolation and a sense of connection (theory area 5).

Comparison With Prior Work

The findings reported in this paper are consistent with but also extend those of the conceptual model of online peer support by Naslund et al [13] for people with severe mental illness. Their model highlights stigma, isolation, and fear of judgment as precursors to forum use. Positive forum experiences are proposed to result in reduced stigma; increased help seeking; and participant activation, defined as learning from and acting upon others' experiential knowledge. By drawing on a diverse evidence base, this synthesis highlights how similar processes underlie positive forum user experiences in a range of contemporary mental health–related contexts, including for family carers [56] and among people with physical health conditions seeking psychological support [62]. This implies that, as with in-person peer support, key processes underlying effective online peer support, including experiential knowledge sharing and reciprocally supportive relationships [81], may represent transdiagnostic mechanisms that are present across service delivery modalities. However, the results of this synthesis emphasize that positive impacts are context dependent. Outcomes including mental health self-efficacy and social connection occur within the context of supportive and vibrant forum cultures proactively managed to minimize rule-breaking content and behaviors. Therefore, when successfully implemented, forums become places for what has been termed "infomotional support" [25], which reflects the combination of simultaneous empathetic emotional support and practical information highly valued by forum users.

This synthesis highlights the importance of the ways in which potentially distressing topics such as suicidal thoughts and behaviors are permitted and managed within peer online forums. Research into online forums specifically focused on suicide prevention provides insights into challenges associated with implementing these services. Consistent with the findings reported in this paper, clear expectation setting was a key component of a social networking intervention for youth suicidal ideation [80]. On agreeing to terms of use, users were informed of forum rules, including restrictions on suicide-related discussion, and the intermittent nature of moderation. The platform used automatic keyword detection to block posts about suicide and prioritized safety over complete anonymity by collecting personal and clinical details used to raise risk concerns with health services. Qualitative research with the users of this forum suggests that, while this proactive approach to risk management did contribute to a safe and supportive environment, there are tensions inherent in restricting discussions of suicidal experiences [55]. Participants noted the value of this policy, particularly for limiting access to distressing content, yet others experienced frustration at not having a space to express their experiences linked to suicidal behavior and noted that such restrictions could perpetuate stigma. Moreover, a suicide prevention forum in the Netherlands implemented proactive moderation to remove descriptions of self-injury and provide ongoing signposting to crisis support services [53]. However, survey data indicated that, while 35% of the participants felt better after use, 12% felt worse and 13% used the forum to find information about suicide methods, with the authors calling into question the service's benefit-to-harm ratio. This evidence highlights both the importance of initial forum design that accounts for potential harms and also the necessity of ongoing evaluation to understand how the content generated on forums impacts user experience and well-being over time.

Our findings emphasize the importance of forum activity for creating the conditions in which users find relevant information and social support. As indicated in the program theory presented in Figure 2, positive forum experiences can serve to create a feedback loop, sustaining online communities via a "network effect" of accelerating online connections [85]. Conversely, inactive forums provide little incentive for users to return, and if limited use is prolonged, there is likely to be a critical point at which forums cease to operate as intended. Therefore, online forum hosts may wish to consider both how to attract new users and how to promote engagement in current forum participants. Regarding the former, forum designers could be guided by previous intervention research, which indicates that a combination of online methods and offline strategies is required to optimize participant recruitment, including social media promotion and endorsement by third-sector and health service providers [33,86]. Regarding sustaining activity, this synthesis suggests that moderators are key to within-forum engagement. This mirrors previous review findings highlighting the facilitative role of forum moderators, who may take on activity-promoting tasks including inducting users into the platform and ensuring that the forum features up-to-date content [25]. A related direction for further research that could extend these findings relates to differences in moderator roles and users' perceptions of moderation across different forum contexts. For example, it is currently uncertain to what extent the status of the moderator as a volunteer, health professional, or peer with lived experience influences factors such as forum activity, user disclosure, and forum culture.

The findings of this synthesis raise several important implications for future research and practice. The evidence included in this synthesis highlights the breadth of settings in which forums have been used to support different mental health problems. Despite differences in delivery contexts, the findings of this study indicate key factors underpinning forum safety and effectiveness, including rule enforcement, proactive and interpersonally sensitive moderation, and the importance of sustaining user engagement to facilitate peer interaction. This suggests that a core set of design features and implementation steps may improve the use and helpfulness of online forums aimed at supporting mental health. As previously stated [44], the findings of this synthesis will inform "best practice" design guidance that aims to advance standards for forum development and evaluation. Relatedly, the findings reported in this paper highlight a range of potential psychosocial processes that could inform future empirical work. For example, future research may seek to investigate the extent to which impacts of forum use on mental health are explained by improvements in perceived social support and mental health self-efficacy. The specific program theories reported in this synthesis are being assessed in a mixed methods realist evaluation with forum users from several UK mental health communities [44].

Limitations

This study has some limitations. The stakeholders interviewed for this synthesis were recruited from UK-based mental health organizations. Participant experiences may not reflect those in other settings and locations, such as user-led communities hosted worldwide. Furthermore, the evidence sources and the subsequent analysis focused primarily on positive and negative experiences of those using forums, with little attention paid to why people who are offered access to forums decline to use them. Better understanding the reasons for nonuse is an important goal of further research with the potential to address barriers to engagement.

Conclusions

Online mental health forums are becoming increasingly prominent resources for people seeking support. This synthesis of recent evidence and stakeholder interviews provides a program theory to explain how positive impacts, such as an improved ability to manage mental health and fulfilling social connection, are more likely to occur in the context of well-organized, regulated, and active forums that provide a supported space for open discussion. Forum design and implementation should consider the limits of what specific forums can and should be used for and how potentially distressing content that falls beyond these limits can be managed in ways that mitigate risks to individuals and broader forum communities.

Acknowledgments

This study is funded by the National Institute for Health and Care Research Health and Social Care Delivery Research Programme (NIHR134035). The views expressed are those of the author or authors and not necessarily those of the National Institute for Health and Care Research or the Department of Health and Social Care. The host of this research was Berkshire Healthcare National Health Service Foundation Trust, and the sponsor was Lancaster University. The research team would like to acknowledge the support of the public and patient involvement group in the development of this manuscript.

Data Availability

The data sets generated during and analyzed during this study are not publicly available. This is to protect participant confidentiality as the qualitative interview data used in this study cannot be fully anonymized. Data are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Initial program theories.
[DOCX File , 22 KB - mental_v11i1e55750_app1.docx]

Multimedia Appendix 2
Search strategy.
[DOCX File , 25 KB - mental_v11i1e55750_app2.docx]

Multimedia Appendix 3
Full-text screening instructions.
[DOCX File , 32 KB - mental_v11i1e55750_app3.docx]

Multimedia Appendix 4
Interview participant demographics and topic guide.
[DOCX File , 21 KB - mental_v11i1e55750_app4.docx]

Multimedia Appendix 5
Characteristics of the included studies.
[XLSX File (Microsoft Excel File), 172 KB - mental_v11i1e55750_app5.xlsx]

References

1. World mental health report: transforming mental health for all. World Health Organization. 2022 Jun 16. URL: https://www.who.int/publications/i/item/9789240049338 [accessed 2024-03-26]
2. Lattie EG, Stiles-Shields C, Graham AK. An overview of and recommendations for more accessible digital mental health services. Nat Rev Psychol 2022 Feb 26;1(2):87-100 [FREE Full text] [doi: 10.1038/s44159-021-00003-1] [Medline: 38515434]
3. O'Grady L, Bender J, Urowitz S, Wiljer D, Jadad AR. Promoting and participating in online health forums: a guide to facilitation and evaluation for health professionals. J Commun Healthc 2013 Jul 18;3(3-4):246-257. [doi: 10.1179/175380710x12870623776478]
4. Ali K, Farrer L, Gulliver A, Griffiths KM. Online peer-to-peer support for young people with mental health problems: a systematic review. JMIR Ment Health 2015 May 19;2(2):e19 [FREE Full text] [doi: 10.2196/mental.4418] [Medline: 26543923]
5. Griffiths KM, Mackinnon AJ, Crisp DA, Christensen H, Bennett K, Farrer L. The effectiveness of an online support group for members of the community with depression: a randomised controlled trial. PLoS One 2012;7(12):e53244 [FREE Full text] [doi: 10.1371/journal.pone.0053244] [Medline: 23285271]
6. Yeo G, Loo G, Oon M, Pang R, Ho D. A digital peer support platform to translate online peer support for emerging adult mental well-being: randomized controlled trial. JMIR Ment Health 2023 Apr 18;10:e43956 [FREE Full text] [doi: 10.2196/43956] [Medline: 36756843]

7. Gleeson JF, Koval P, Zyphur M, Lederman R, Herrman H, Eleftheriadis D, et al. A randomized controlled trial of moderated online social therapy for family carers of first-episode psychosis patients in a specialist treatment setting. Schizophr Res 2023 May;255:203-212. [doi: 10.1016/j.schres.2023.03.019] [Medline: 37004332]
8. Sin J, Henderson C, Elkes J, Cornelius V, Woodham LA, Batchelor R, et al. Effect of digital psychoeducation and peer support on the mental health of family carers supporting individuals with psychosis in England (COPe-support): a randomised clinical trial. Lancet Digit Health 2022 May;4(5):e320-e329 [FREE Full text] [doi: 10.1016/S2589-7500(22)00031-0] [Medline: 35379593]
9. Lobban F, Akers N, Appelbe D, Chapman L, Collinge L, Dodd S, et al. Clinical effectiveness of a web-based peer-supported self-management intervention for relatives of people with psychosis or bipolar (REACT): online, observer-blind, randomised controlled superiority trial. BMC Psychiatry 2020 Apr 14;20(1):160 [FREE Full text] [doi: 10.1186/s12888-020-02545-9] [Medline: 32290827]
10. Widemalm M, Hjärthag F. The forum as a friend: parental mental illness and communication on open internet forums. Soc Psychiatry Psychiatr Epidemiol 2015 Oct;50(10):1601-1607. [doi: 10.1007/s00127-015-1036-z] [Medline: 25702167]
11. Poole R, Smith D, Simpson S. How patients contribute to an online psychoeducation forum for bipolar disorder: a virtual participant observation study. JMIR Ment Health 2015;2(3):e21 [FREE Full text] [doi: 10.2196/mental.4123] [Medline: 26543925]
12. Evans M, Donelle L, Hume-Loveland L. Social support and online postpartum depression discussion groups: a content analysis. Patient Educ Couns 2012 Jun;87(3):405-410. [doi: 10.1016/j.pec.2011.09.011] [Medline: 22019021]
13. Naslund JA, Aschbrenner KA, Marsch LA, Bartels SJ. The future of mental health care: peer-to-peer support and social media. Epidemiol Psychiatr Sci 2016 Apr;25(2):113-122 [FREE Full text] [doi: 10.1017/S2045796015001067] [Medline: 26744309]
14. De Choudhury M, De S. Mental health discourse on Reddit: self-disclosure, social support, and anonymity. Proc Int AAAI Conf Web Soc Media 2014;8(1):71-80. [doi: 10.1609/icwsm.v8i1.14526]
15. Kendal S, Kirk S, Elvey R, Catchpole R, Pryjmachuk S. How a moderated online discussion forum facilitates support for young people with eating disorders. Health Expect 2017 Feb 03;20(1):98-111 [FREE Full text] [doi: 10.1111/hex.12439] [Medline: 26725547]
16. Easton K, Diggle J, Ruethi-Davis M, Holmes M, Byron-Parker D, Nuttall J, et al. Qualitative exploration of the potential for adverse events when using an online peer support network for mental health: cross-sectional survey. JMIR Ment Health 2017 Oct 30;4(4):e49 [FREE Full text] [doi: 10.2196/mental.8168] [Medline: 29084710]
17. Lavis A, Winter R. #Online harms or benefits? An ethnographic analysis of the positives and negatives of peer-support around self-harm on social media. J Child Psychol Psychiatry 2020 Aug 27;61(8):842-854. [doi: 10.1111/jcpp.13245] [Medline: 32459004]
18. Saha K, Ernala SK, Dutta S, Sharma E, De Choudhury M. Understanding moderation in online mental health communities. In: Proceedings of the Social Computing and Social Media. Participation, User Experience, Consumer Experience, and Applications of Social Computing. 2020 Presented at: HCII 2020; July 19-24, 2020; Copenhagen, Denmark. [doi: 10.1007/978-3-030-49576-3_7]
19. Stevens M, Cartagena Farías J, Mindel C, D'Amico F, Evans-Lacko S. Pilot evaluation to assess the effectiveness of youth peer community support via the Kooth online mental wellbeing website. BMC Public Health 2022 Oct 12;22(1):1903 [FREE Full text] [doi: 10.1186/s12889-022-14223-4] [Medline: 36224546]
20. Marinova N, Rogers T, MacBeth A. Predictors of adolescent engagement and outcomes - A cross-sectional study using the togetherall (formerly Big White Wall) digital mental health platform. J Affect Disord 2022 Aug 15;311:284-293 [FREE Full text] [doi: 10.1016/j.jad.2022.05.058] [Medline: 35588912]
21. r/depression, because nobody should be alone in a dark place. Reddit. URL: https://subredditstats.com/r/depression [accessed 2024-03-27]
22. Sorkin DH, Janio EA, Eikey EV, Schneider M, Davis K, Schueller SM, et al. Rise in use of digital mental health tools and technologies in the United States during the COVID-19 pandemic: survey study. J Med Internet Res 2021 Apr 16;23(4):e26994 [FREE Full text] [doi: 10.2196/26994] [Medline: 33822737]
23. Storman D, Jemioło P, Swierz MJ, Sawiec Z, Antonowicz E, Prokop-Dorner A, et al. Meeting the unmet needs of individuals with mental disorders: scoping review on peer-to-peer web-based interactions. JMIR Ment Health 2022 Dec 05;9(12):e36056 [FREE Full text] [doi: 10.2196/36056] [Medline: 36469366]
24. Fortuna KL, Naslund JA, LaCroix JM, Bianco CL, Brooks JM, Zisman-Ilani Y, et al. Digital peer support mental health interventions for people with a lived experience of a serious mental illness: systematic review. JMIR Ment Health 2020 Apr 03;7(4):e16460 [FREE Full text] [doi: 10.2196/16460] [Medline: 32243256]
25. Hanley T, Prescott J, Gomez KU. A systematic review exploring how young people use online forums for support around mental health issues. J Ment Health 2019 Oct 03;28(5):566-576. [doi: 10.1080/09638237.2019.1630725] [Medline: 31267813]
26. Rayland A, Andrews J. From social network to peer support network: opportunities to explore mechanisms of online peer support for mental health. JMIR Ment Health 2023 Feb 28;10:e41855 [FREE Full text] [doi: 10.2196/41855] [Medline: 36853738]

27. Wong G, Greenhalgh T, Westhorp G, Buckingham J, Pawson R. RAMESES publication standards: realist syntheses. BMC Med 2013 Jan 29;11(1):21 [FREE Full text] [doi: 10.1186/1741-7015-11-21] [Medline: 23360677]
28. Hunter R, Gorely T, Beattie M, Harris K. Realist review. Int Rev Sport Exerc Psychol 2022 Jan 18;15(1):242-265. [doi: 10.1080/1750984x.2021.1969674]
29. Philippe TJ, Sikder N, Jackson A, Koblanski ME, Liow E, Pilarinos A, et al. Digital health interventions for delivery of mental health care: systematic and comprehensive meta-review. JMIR Ment Health 2022 May 12;9(5):e35159 [FREE Full text] [doi: 10.2196/35159] [Medline: 35551058]
30. Dalkin SM, Greenhalgh J, Jones D, Cunningham B, Lhussier M. What's in a mechanism? Development of a key concept in realist evaluation. Implement Sci 2015 Apr 16;10(1):49 [FREE Full text] [doi: 10.1186/s13012-015-0237-x] [Medline: 25885787]
31. Jagosh J, Stott H, Halls S, Thomas R, Liddiard C, Cupples M, et al. Benefits of realist evaluation for rapidly changing health service delivery. BMJ Open 2022 Jul 27;12(7):e060347 [FREE Full text] [doi: 10.1136/bmjopen-2021-060347] [Medline: 35896288]
32. Smith-Merry J, Goggin G, Campbell A, McKenzie K, Ridout B, Baylosis C. Social connection and online engagement: insights from interviews with users of a mental health online forum. JMIR Ment Health 2019 Mar 26;6(3):e11084 [FREE Full text] [doi: 10.2196/11084] [Medline: 30912760]
33. Morriss R, Kaylor-Hughes C, Rawsthorne M, Coulson N, Simpson S, Guo B, et al. A direct-to-public peer support program (big white wall) versus web-based information to aid the self-management of depression and anxiety: results and challenges of an automated randomized controlled trial. J Med Internet Res 2021 Apr 23;23(4):e23487 [FREE Full text] [doi: 10.2196/23487] [Medline: 33890858]
34. Griffiths KM, Reynolds J, Vassallo S. An online, moderated peer-to-peer support bulletin board for depression: user-perceived advantages and disadvantages. JMIR Ment Health 2015;2(2):e14 [FREE Full text] [doi: 10.2196/mental.4266] [Medline: 26543919]
35. Lewis SP, Michal NJ. Start, stop, and continue: preliminary insight into the appeal of self-injury e-communities. J Health Psychol 2016 Feb;21(2):250-260. [doi: 10.1177/1359105314527140] [Medline: 24707032]
36. Park A, Conway M, Chen AT. Examining thematic similarity, difference, and membership in three online mental health communities from Reddit: a text mining and visualization approach. Comput Human Behav 2018 Jan;78:98-112 [FREE Full text] [doi: 10.1016/j.chb.2017.09.001] [Medline: 29456286]
37. Pruksachatkun Y, Pendse SR, Sharma A. Moments of change: analyzing peer-based cognitive support in online mental health forums. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI '19; May 4-9, 2019; Glasgow, Scotland.
38. Tucker IM, Goodings L. Digital atmospheres: affective practices of care in Elefriends. Sociol Health Illn 2017 May;39(4):629-642. [doi: 10.1111/1467-9566.12545] [Medline: 28326556]
39. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
40. Dada S, Dalkin S, Gilmore B, Hunter R, Mukumbang F. Applying and reporting relevance, richness and rigour in realist evidence appraisals: advancing key concepts in realist reviews. Res Synth Methods 2023 May;14(3):504-514. [doi: 10.1002/jrsm.1630] [Medline: 36872619]
41. Williams L, Rycroft-Malone J, Burton CR, Edwards S, Fisher D, Hall B, et al. Improving skills and care standards in the support workforce for older people: a realist synthesis of workforce development interventions. BMJ Open 2016 Aug 26;6(8):e011964 [FREE Full text] [doi: 10.1136/bmjopen-2016-011964] [Medline: 27566640]
42. Morton T, Wong G, Atkinson T, Brooker D. Sustaining community-based interventions for people affected by dementia long term: the SCI-Dem realist review. BMJ Open 2021 Jul 07;11(7):e047789 [FREE Full text] [doi: 10.1136/bmjopen-2020-047789] [Medline: 34233990]
43. Pawson R. Digging for nuggets: how ‘bad’ research can yield ‘good’ evidence. Int J Soc Res Methodol 2006 Apr;9(2):127-142. [doi: 10.1080/13645570600595314]
44. Lobban F, Coole M, Donaldson E, Glossop Z, Haines J, Johnston R, et al. Improving Peer Online Forums (iPOF): protocol for a realist evaluation of peer online mental health forums to inform practice and policy. BMJ Open 2023 Jul 30;13(7):e075142 [FREE Full text] [doi: 10.1136/bmjopen-2023-075142] [Medline: 37518092]
45. Liu J, Wang J. Users' intention to continue using online mental health communities: empowerment theory perspective. Int J Environ Res Public Health 2021 Sep 07;18(18):9427 [FREE Full text] [doi: 10.3390/ijerph18189427] [Medline: 34574361]
46. Liao Siling J, Truss K, Philips L, Eastwood O, Bendall S. Young people's journeys of recovery from trauma: a qualitative study of narratives from internet forums. Psychol Trauma 2021 Mar;13(3):293-301. [doi: 10.1037/tra0000990] [Medline: 33661674]
47. Smit D, Vrijsen JN, Groeneweg B, Vellinga-Dings A, Peelen J, Spijker J. A newly developed online peer support community for depression (depression connect): qualitative study. J Med Internet Res 2021 Jul 12;23(7):e25917 [FREE Full text] [doi: 10.2196/25917] [Medline: 34255659]
48. Browne NL, Carragher NO, O’Toole A, Pimm J, Ryder J, Thew GR. Evaluating user experiences of SHaRON: an online CBT-based peer support platform. Cogn Behav Ther 2022 Apr 11;15:e18. [doi: 10.1017/s1754470x22000150]

49. Moore D, Drey N, Ayers S. A meta-synthesis of women's experiences of online forums for maternal mental illness and stigma. Arch Womens Ment Health 2020 Aug 24;23(4):507-515. [doi: 10.1007/s00737-019-01002-1] [Medline: 31646392]
50. Gibson K, Trnka S. Young people's priorities for support on social media: “it takes trust to talk about these issues”. Comput Hum Behav 2020 Jan;102:238-247. [doi: 10.1016/j.chb.2019.08.030]
51. Guccini F, McKinley G. "How deep do I have to cut?": non-suicidal self-injury and imagined communities of practice on Tumblr. Soc Sci Med 2022 Mar;296:114760 [FREE Full text] [doi: 10.1016/j.socscimed.2022.114760] [Medline: 35124545]
52. Quiniones C, Oster C. Embracing or resisting masculinity: male participation in the proeating disorders (proana) online community. Psychol Men Masculinities 2019 Jul;20(3):368-378. [doi: 10.1037/men0000169]
53. Mokkenstorm JK, Mérelle SY, Smit JH, Beekman AT, Kerkhof AJ, Huisman A, et al. Exploration of benefits and potential harmful effects of an online forum for visitors to the suicide prevention platform in the Netherlands. Crisis 2020 May;41(3):205-213. [doi: 10.1027/0227-5910/a000627] [Medline: 31657643]
54. Bradley A, James RJ. Defining the key issues discussed by problematic gamblers on web-based forums: a data-driven approach. Int Gambl Stud 2020 Jul 30;21(1):59-73. [doi: 10.1080/14459795.2020.1801793]
55. Bailey E, Robinson J, Alvarez-Jimenez M, Nedeljkovic M, Valentine L, Bendall S, et al. Moderated online social therapy for young people with active suicidal ideation: qualitative study. J Med Internet Res 2021 Apr 05;23(4):e24260 [FREE Full text] [doi: 10.2196/24260] [Medline: 33818392]
56. Barraclough J. Caregivers’ engagement with online support. Lancaster University. 2021. URL: https://www.proquest.com/openview/260892e536412e1ceb4007364bc12dcf/1?pq-origsite=gscholar&cbl=2026366&diss=y [accessed 2024-03-27]
57. Heyes K. Socialization or social isolation?: mental health community support in the digital age. In: El Morr C, editor. Novel Applications of Virtual Communities in Healthcare Settings. Hershey, PA: IGI Global; 2018.
58. Campbell A, Ridout B, Amon K, Navarro P, Collyer B, Dalgleish J. A customized social network platform (kids helpline circles) for delivering group counseling to young people experiencing family discord that impacts their well-being: exploratory study. J Med Internet Res 2019 Dec 20;21(12):e16176 [FREE Full text] [doi: 10.2196/16176] [Medline: 31859671]
59. Sindoni M. ‘#YouCanTalk’: a multimodal discourse analysis of suicide prevention and peer support in the Australian BeyondBlue platform. Discourse Commun 2019 Dec 26;14(2):202-221. [doi: 10.1177/1750481319890386]
60. Abou Seif N, John-Baptiste Bastien R, Wang B, Davies J, Isaken M, Ball E, et al. Effectiveness, acceptability and potential harms of peer support for self-harm in non-clinical settings: systematic review. BJPsych Open 2022 Jan 17;8(1):e28 [FREE Full text] [doi: 10.1192/bjo.2021.1081] [Medline: 35034666]
61. Strand M, Eng LS, Gammon D. Combining online and offline peer support groups in community mental health care settings: a qualitative study of service users' experiences. Int J Ment Health Syst 2020;14:39 [FREE Full text] [doi: 10.1186/s13033-020-00370-x] [Medline: 32514303]
62. Morehouse S, Schaible K, Williams O, Herlache-Pretzer E, Webster S. Impacts of online support groups on quality of life, and perceived anxiety and depression in those with ME/CFS: a survey. Fatigue Biomed Health Behav 2021 Jul 07;9(2):113-122. [doi: 10.1080/21641846.2021.1950406]
63. Heinsch M, Geddes J, Sampson D, Brosnan C, Hunt S, Wells H, et al. Disclosure of suicidal thoughts during an e-mental health intervention: relational ethics meets actor-network theory. Ethics Behav 2019 Dec 01;31(3):151-170. [doi: 10.1080/10508422.2019.1691003]
64. Perry A, Lamont-Mills A, du Preez J, du Plessis C. "I want to be stepping in more" - professional online forum moderators' experiences of supporting individuals in a suicide crisis. Front Psychiatry 2022 Jun 13;13:863509 [FREE Full text] [doi: 10.3389/fpsyt.2022.863509] [Medline: 35774095]
65. Kruzan KP, Bazarova NN, Whitlock J. Investigating self-injury support solicitations and responses on a mobile peer support application. Proc ACM Hum Comput Interact 2021 Oct 18;5(CSCW2):1-23. [doi: 10.1145/3479498]
66. Ekström V, Johansson M. Choosing internet-based treatment for problematic alcohol use-why, when and how? Users' experiences of treatment online. Addict Sci Clin Pract 2020 Jun 29;15(1):22 [FREE Full text] [doi: 10.1186/s13722-020-00196-5] [Medline: 32600456]
67. Croucher L, Mertan E, Shafran R, Bennett SD. The use of Mumsnet by parents of young people with mental health needs: qualitative investigation. JMIR Ment Health 2020 Sep 03;7(9):e18271 [FREE Full text] [doi: 10.2196/18271] [Medline: 32880583]
68. Prescott J, Hanley T, Ujhelyi Gomez K. Why do young people use online forums for mental health and emotional support? Benefits and challenges. Brit J Guid Counsel 2019 May 20;47(3):317-327. [doi: 10.1080/03069885.2019.1619169]
69. Shi J, Dai Y. Promoting favorable attitudes toward seeking counseling among people with depressive symptomatology: a masspersonal communication approach. Health Commun 2022 Feb;37(2):242-254. [doi: 10.1080/10410236.2020.1834209] [Medline: 33074028]
70. Lederman R, Gleeson J, Wadley G, D’alfonso S, Rice S, Santesteban-Echarri O, et al. Support for carers of young people with mental illness. ACM Trans Comput Hum Interact 2019 Feb 21;26(1):1-33. [doi: 10.1145/3301421]
71. Tan YT, Rehm IC, Stevenson JL, De Foe A. Social media peer support groups for obsessive-compulsive and related disorders: understanding the predictors of negative experiences. J Affect Disord 2021 Feb 15;281:661-672. [doi: 10.1016/j.jad.2020.11.094] [Medline: 33234279]

72. Feuston JL. Caring together: a digital ethnography of how people with mental illness participate online. Northwestern University . 2020 Dec. URL: https://www.proquest.com/openview/e4beca5a1ecc74809d2b03e31e49d1e6/1?pq-origsite=gscholar&cbl=18750&diss=y [accessed 2024-03-27]
73. Figueras Bates C. Cognitive and affective dimensions of mitigation in advice. Corpus Pragmatics 2019 Oct 26;4(1):31-57. [doi: 10.1007/s41701-019-00064-x]
74. Trail T, Friedman E, Rutter CM, Tanielian T. The relationship between engagement in online support groups and social isolation among military caregivers: longitudinal questionnaire study. J Med Internet Res 2020 Apr 23;22(4):e16423 [FREE Full text] [doi: 10.2196/16423] [Medline: 32324141]
75. Yip JW. Evaluating the communication of online social support: a mixed-methods analysis of structure and content. Health Commun 2020 Sep;35(10):1210-1218. [doi: 10.1080/10410236.2019.1623643] [Medline: 31154856]
76. Stana A, Miller AR. "Being a mom = having all the feels": social support in a postpartum depression online support group. Atlantic J Commun 2019 Jun 12;27(5):297-310. [doi: 10.1080/15456870.2019.1616736]
77. Liu Y, Kornfield R, Shaw BR, Shah DV, McTavish F, Gustafson DH. Giving and receiving social support in online substance use disorder forums: how self-efficacy moderates effects on relapse. Patient Educ Couns 2020 Jun;103(6):1125-1133 [FREE Full text] [doi: 10.1016/j.pec.2019.12.015] [Medline: 31901364]
78. Drioli-Phillips PG, Oxlad M, LeCouteur A, Feo R, Scholz B. Men's talk about anxiety online: constructing an authentically anxious identity allows help-seeking. Psychol Men Masculinities 2021 Jan;22(1):77-87. [doi: 10.1037/men0000268]
79. Lu Y, Luo S, Liu X. Development of social support networks by patients with depression through online health communities: social network analysis. JMIR Med Inform 2021 Jan 07;9(1):e24618 [FREE Full text] [doi: 10.2196/24618] [Medline: 33279878]
80. Bailey E, Alvarez-Jimenez M, Robinson J, D'Alfonso S, Nedeljkovic M, Davey CG, et al. An enhanced social networking intervention for young people with active suicidal ideation: safety, feasibility and acceptability outcomes. Int J Environ Res Public Health 2020 Apr 03;17(7):2435 [FREE Full text] [doi: 10.3390/ijerph17072435] [Medline: 32260111]
81. Gidugu V, Rogers ES, Harrington S, Maru M, Johnson G, Cohee J, et al. Individual peer support: a qualitative study of mechanisms of its effectiveness. Community Ment Health J 2015 May 23;51(4):445-452. [doi: 10.1007/s10597-014-9801-0] [Medline: 25535049]
82. Lobban F, Akers N, Appelbe D, Iraci Capuccinello R, Chapman L, Collinge L, et al. A web-based, peer-supported self-management intervention to reduce distress in relatives of people with psychosis or bipolar disorder: the REACT RCT. Health Technol Assess 2020 Jun;24(32):1-142 [FREE Full text] [doi: 10.3310/hta24320] [Medline: 32608353]
83. Reupert A, Maybery D, Bartholomew C, Cuff R, Matar J, Pettenuzzo L, et al. An online intervention for vulnerable young adults: identifying mechanisms of change using a grounded theory approach. Soc Psychiatry Psychiatr Epidemiol 2022 Feb;57(2):293-303. [doi: 10.1007/s00127-021-02082-0] [Medline: 33864095]
84. Feuston JL, Taylor AS, Piper AM. Conformity of eating disorders through content moderation. Proc ACM Hum Comput Interact 2020 May 29;4(CSCW1):1-28. [doi: 10.1145/3392845]
85. Gopalsamy R, Semenov A, Pasiliao E, McIntosh S, Nikolaev A. Engagement as a driver of growth of online health forums: observational study. J Med Internet Res 2017 Aug 29;19(8):e304 [FREE Full text] [doi: 10.2196/jmir.7249] [Medline: 28851677]
86. Robinson H, Appelbe D, Dodd S, Flowers S, Johnson S, Jones SH, et al. Methodological challenges in web-based trials: update and insights from the relatives education and coping toolkit trial. JMIR Ment Health 2020 Jul 17;7(7):e15878 [FREE Full text] [doi: 10.2196/15878] [Medline: 32497018]

Abbreviations

CMO: context-mechanism-outcome configuration

Edited by J Torous; submitted 22.12.23; peer-reviewed by A Radovic-Stakic, M Evans; comments to author 14.02.24; revised version received 27.02.24; accepted 01.03.24; published 09.05.24.

Please cite as:
Marshall P, Booth M, Coole M, Fothergill L, Glossop Z, Haines J, Harding A, Johnston R, Jones S, Lodge C, Machin K, Meacock R, Nielson K, Puddephatt JA, Rakic T, Rayson P, Robinson H, Rycroft-Malone J, Shryane N, Swithenbank Z, Wise S, Lobban F
Understanding the Impacts of Online Mental Health Peer Support Forums: Realist Synthesis
JMIR Ment Health 2024;11:e55750
URL: https://mental.jmir.org/2024/1/e55750
doi:10.2196/55750
PMID:38722680

Review

Insights Derived From Text-Based Digital Media, in Relation to Mental Health and Suicide Prevention, Using Data Analysis and Machine Learning: Systematic Review

Colm Sweeney[1*], MSc; Edel Ennis[1*], PhD; Maurice D Mulvenna[2], Prof Dr; Raymond Bond[2], Prof Dr; Siobhan O'Neill[1], Prof Dr

[1]Department of Psychlogy, Ulster University, Coleraine, United Kingdom

[2]School of Computing, Ulster University, Belfast, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Colm Sweeney, MSc
Department of Psychlogy
Ulster University
Cromore Rd
Coleraine, BT52 1SA
United Kingdom
Phone: 44 02870 123 456
Email: Sweeney-C23@ulster.ac.uk

Abstract

Background: Text-based digital media platforms have revolutionized communication and information sharing, providing valuable access to knowledge and understanding in the fields of mental health and suicide prevention.

Objective: This systematic review aimed to determine how machine learning and data analysis can be applied to text-based digital media data to understand mental health and aid suicide prevention.

Methods: A systematic review of research papers from the following major electronic databases was conducted: Web of Science, MEDLINE, Embase (via MEDLINE), and PsycINFO (via MEDLINE). The database search was supplemented by a hand search using Google Scholar.

Results: Overall, 19 studies were included, with five major themes as to how data analysis and machine learning techniques could be applied: (1) as predictors of personal mental health, (2) to understand how personal mental health and suicidal behavior are communicated, (3) to detect mental disorders and suicidal risk, (4) to identify help seeking for mental health difficulties, and (5) to determine the efficacy of interventions to support mental well-being.

Conclusions: Our findings show that data analysis and machine learning can be used to gain valuable insights, such as the following: web-based conversations relating to depression vary among different ethnic groups, teenagers engage in a web-based conversation about suicide more often than adults, and people seeking support in web-based mental health communities feel better after receiving online support. Digital tools and mental health apps are being used successfully to manage mental health, particularly through the COVID-19 epidemic, during which analysis has revealed that there was increased anxiety and depression, and web-based communities played a part in reducing isolation during the pandemic. Predictive analytics were also shown to have potential, and virtual reality shows promising results in the delivery of preventive or curative care. Future research efforts could center on optimizing algorithms to enhance the potential of text-based digital media analysis in mental health and suicide prevention. In addressing depression, a crucial step involves identifying the factors that contribute to happiness and using machine learning to forecast these sources of *happiness*. This could extend to understanding how various activities result in improved happiness across different socioeconomic groups. Using insights gathered from such data analysis and machine learning, there is an opportunity to craft digital interventions, such as chatbots, designed to provide support and address mental health challenges and suicide prevention.

(JMIR Ment Health 2024;11:e55747) doi:10.2196/55747

KEYWORDS
mental health; machine learning; text analysis; digital intervention

Introduction

Background

Text-based digital media platforms have revolutionized communication and information sharing, offering valuable opportunities to gain insights into various domains, including mental health and suicide prevention.

Social media platforms have become significant sources of data for studying mental health and suicide prevention, where researchers have explored the potential of using platforms such as X (X Corp), formerly known as Twitter (Twitter, Inc) and Facebook (Meta Platforms, Inc) to gain insights into individuals' mental well-being, detect mental health concerns, and identify suicide risk factors. For example, Coppersmith et al [1] developed a machine learning model to detect signals related to depression in user posts on Twitter, achieving promising results. In addition, De Choudhury et al [2] analyzed Facebook posts to identify individuals at risk of depression, demonstrating the feasibility of using social media data for mental health monitoring. Research methods involve various techniques, including sentiment analysis, topic modeling, and natural language processing (NLP), to analyze large volumes of data and identify patterns and trends. For instance, Park et al [3] applied sentiment analysis to examine suicide-related tweets and identified specific linguistic features associated with suicidal ideation. Sik et al [4] used topic modeling to identify mental health–related topics in web-based forums, facilitating targeted interventions and support. In addition, Burnap et al [5] used NLP techniques to analyze web-based content and identify individuals expressing suicidal ideation, which could enable timely interventions.

Data analysis and machine learning techniques have been used for detecting mental health issues and identifying individuals at risk of suicide, where these sophisticated techniques could enhance clinical decision-making in relation to suicide [6]. Some researchers have explored the use of predictive models to assess suicide risk factors and facilitate early intervention. For example, O'Dea et al [7] developed a predictive model using machine learning algorithms to identify suicide attempt risk among social media users, highlighting the potential for targeted prevention strategies. Data analysis can also be used to provide a valued understanding of factors associated with suicide and mental health, which are not easily identifiable. These insights can then be used to develop strategies for prevention and intervention. For example, data analysis can identify potential underlying causes and risk factors associated with suicide, which can then lead to the development of interventions for susceptible groups. Finally, data analysis can also be used to analyze the effectiveness of current prevention efforts to improve targeted interventions and strategies.

Objectives

With the rise in the use of smartphones, digital interventions have been able to offer a solution to address the increasing demand for mental health services [8] and to relieve certain barriers in mental health provision, such as the stigma around accessing psychological health services and geographic isolation [9]. This paper presents a systematic review of the research on the application of machine learning and data analysis to text-based digital media data in relation to mental health and suicide prevention to help answer the following research question: How can machine learning and data analysis be applied to text-based digital media data to understand mental health and aid suicide prevention?

Methods

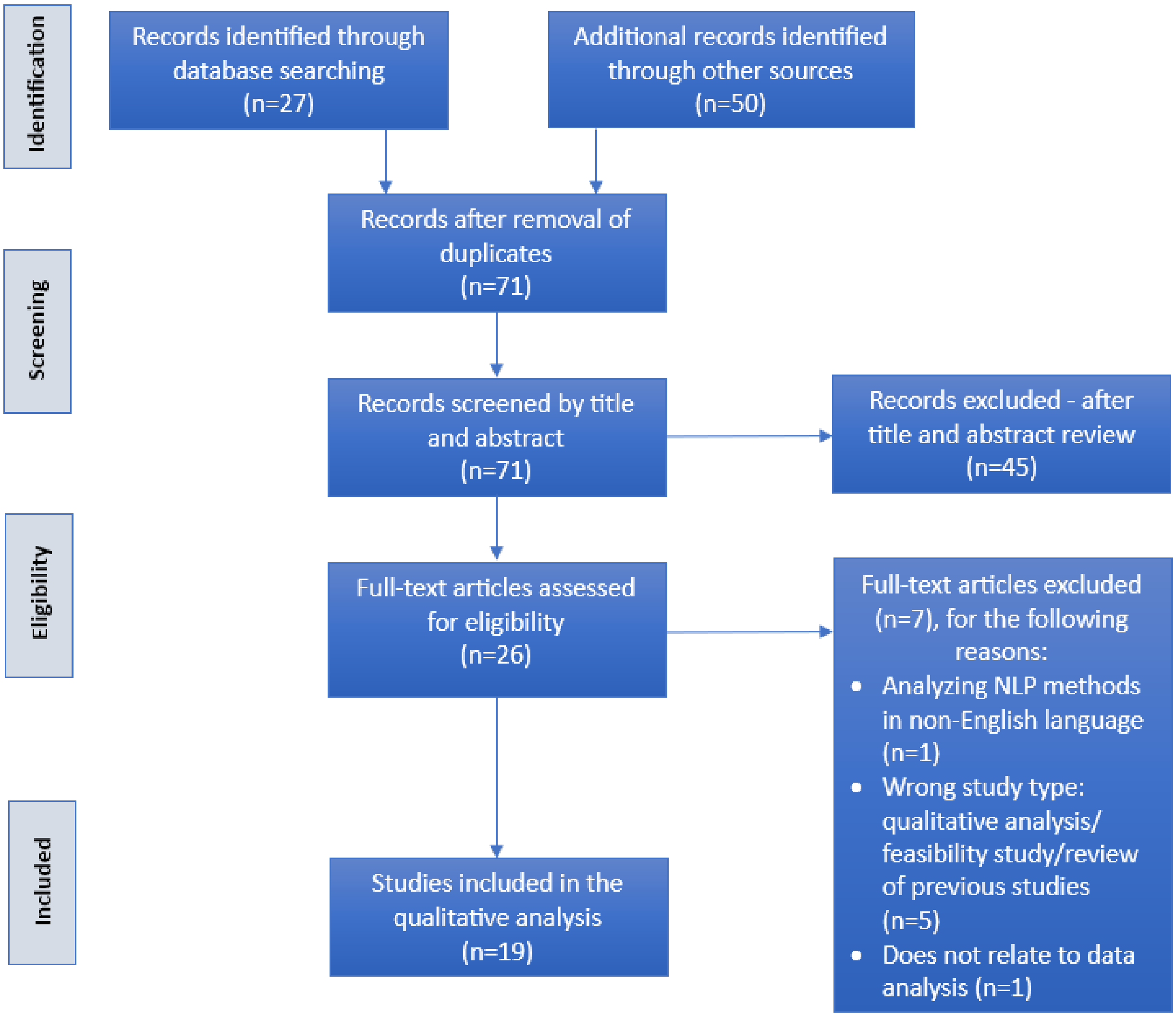
Search Strategy: Electronic Database Search

A systematic literature search was performed for articles published from January 1, 2013, to July 10, 2023, and was conducted using 4 databases, namely Web of Science, MEDLINE, Embase (via MEDLINE), and PsycINFO (via MEDLINE), using the following search terms, which were adapted for each database: (mental health OR depression OR suicide) AND (machine learning OR deep learning OR artificial intelligence) AND (text analysis OR text mining OR data analysis) AND (digital intervention OR digital mental health). Retrospective searches were conducted (using the same criteria) using both PubMed and Scopus databases to extend the research to bigger databases. However, no new relevant papers were detected. The complete search strings are included in Multimedia Appendix 1. CS performed the literature search. EE, MDM, and RB discussed and verified the inclusion or exclusion criteria. The *Study Selection* section identifies how articles were included in or excluded from this review. These database searches were supplemented by hand-search techniques. An additional manual search was run using advanced search within Google Scholar (date: July 10, 2023). The first 5 pages of search results (n=50 records) were screened based on title, as per PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [10].

Study Selection

A total of 27 records were identified according to the search methods explained in the *Search Strategy* section. An additional 50 records were identified by searching Google Scholar articles. Of the 71 unique articles, 45 (63%) were excluded after abstract screening. A full-text review was performed for the remaining 26 (37%) articles according to study inclusion criteria, after which 19 (73%) of these articles were included (Figure 1; Multimedia Appendix 2 [10]). A total of 7 reports failed to meet the stated inclusion criteria. These included papers (1/7, 14%) analyzing NLP methods in a non-English language; papers (5/7, 71%) with a wrong study type, such as qualitative analysis of the use of social media in mental health and teaching mental health intervention in schools or feasibility study or review of previous studies; and papers (1/7, 14%) that did not relate to data analysis. Figure 1 shows a flowchart of the study inclusion process.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram. NLP: natural language processing.



Quality Assessment

An assessment for bias risk was performed using the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) guidelines [11]. Multimedia Appendix 3 provides more details relating to how the TRIPOD checklist was used and the TRIPOD ratio calculated for the articles relating to prediction and classification (refer to Table S1 in Multimedia Appendix 3 for risk bias results).

Results

Types of Analyses to Assess Text-Based Digital Data and Outcomes

This review aimed to determine how machine learning and data analysis can be used to assess text-based digital media data in relation to mental health and suicide prevention. Regarding the type of analysis and outcome measures used within the publications reviewed in this study, machine learning and text-based data analysis were used in 4 (21%) of the 19 studies [12-15]. A total of 3 (16%) studies performed some sort of analysis on survey or questionnaire data [16-19], and 3 (16%) papers analyzed the value of text-based digital media [20-22]. The analysis of digital interventions was the main type of analysis used by Onyeaka et al [23], Vermetten et al [24], and Van Gemert-Pijnen et al [25]. The remaining types of investigations included the analysis of forum or discussion data [26] and longitudinal analysis [27]. Where machine learning was used for prediction within the studies, the outcome metrics were also listed in the table. These include the study by Roy et al [28], who investigated how machine learning approaches could be used to predict suicidal ideation from social media data. They trained a random forest model using neural networks to predict suicide ideation status with an area under the curve of 0.88. Gu et al [29] used convolutional neural network for text for classifier training and classification, which produced the following scores: precision=0.84, recall=0.84, and F_1-score=0.84. Oyebode et al [30] used 5 different machine learning methods to evaluate mental health apps based on user reviews. The 5 models produced similar scores, with the stochastic gradient descent showing the best performance of the 5 classifiers (Table 1).

Table 1. Review of themes, showing study and year, title, population, data volume, and theme.

Study and year	Title	Population	Data	Volume	Type of analysis	Outcome metrics	Theme
Aitken et al [16], 2021	How Much of the Effect of Disability Acquisition on Mental Health is Mediated Through Employment and Income? A Causal Mediation Analysis Quantifying Interventional Indirect Effects Using Data From Four Waves of an Australian Cohort Study	Australian households	Data from the HILDA[a] survey	10,450 respondents who provided data on disability in the survey	Analysis of survey data using causal mediation analysis	• Total causal effect of disability acquisition on mental health was estimated to be a 4.8-point decline in mental health.	Predictors of personal mental health
Khattar et al [19], 2020	Effects of the Disastrous Pandemic COVID-19 on Learning Styles, Activities and Mental Health of Young Indian Students-a Machine Learning Approach	Young Indian students	Web-based survey and questionnaire results	583 students' responses	Analysis of survey and questionnaire data using association rule mining	• Rule 1 produced the following scores: support=0.286, confidence=0.671, and lift=1.454	Predictors of personal mental health
Valdez et al [27], 2020	Social Media Insights Into US Mental Health During the COVID-19 Pandemic: Longitudinal Analysis of Twitter Data	Users of the Twitter (Twitter, Inc) social media platform	Publicly available Twitter data	86,581 tweets	Longitudinal analysis using the VADER[b] sentiment analysis tool	• Negative trajectory in sentiment scores for the user-timeline data	Predictors of personal mental health
Xiao et al [17], 2020	Mental Health of Chinese Online Networkers Under COVID-19: A Sociological Analysis of Survey Data	Users of the Chinese WeChat (Tencent Holdings Limited) network platform	Results of completed questionnaires	Across 3491 participants, 2015 questionnaires were valid	Analysis of survey data using OLS[c] regression	• With 1-unit (SD) increase in SES[d], depression decreased by a margin of –0.52 (*P*<.001).	Predictors of personal mental health
Roy et al [28], 2020	A Machine Learning Approach Predicts Future Risk to Suicidal Ideation From Social Media Data	Users of the Twitter social media platform	Publicly available Twitter data	512,526 tweets	Machine learning for prediction: RF[e] model using NN[f] outputs	• RF and NN: • AUC[g]: 0.88 (95% CI 0.86-0.90)	Detection of mental disorders and suicidal risk
Simms et al [12], 2017	Detecting Cognitive Distortions Through Machine Learning Text Analytics	Users of the Tumblr (Automattic) microblogging and social networking site	Personal blogs from Tumblr	493 posts	Machine learning and text-based data analysis using logistic regression	• Accuracy of the logistic model was 73%.	Detection of mental disorders and suicidal risk
Golz et al [26], 2022	Mental Health-Related Communication in a Virtual Community: Text Mining Analysis of a Digital Exchange Platform During the COVID-19 Pandemic	Almost 700 users of inCLOUsiv, a web-based community platform for mental health	Data from the forums and live discussions were stored in the MySQL database	Data set consisted of 31,764 words	Sentiment analysis of forum and discussion data	• 72% of the identified sentiments were positive.	Understanding how personal mental health and suicidal behavior are communicated

Study and year	Title	Population	Data	Volume	Type of analysis	Outcome metrics	Theme
Castilla-Puentes et al [13], 2021	Digital Conversations About Depression Among Hispanics and non-Hispanics in the US: a Big‐Data, Machine Learning Analysis Identifies Specific Characteristics of Depression Narratives in Hispanics	Hispanic and non-Hispanic population in the United States	Open-source sites, message boards, social networks, and blogs	441,000 unique open-source conversations about depression	Machine learning and text-based data analysis: content analysis	• Content analysis shows that 66% of conversations among Hispanic population portray a negative tone as compared to 39% among non-Hispanic population.	Understanding how personal mental health and suicidal behavior are communicated
Liu, and Kong [20], 2021	Why Do Users of Online Mental Health Communities Get Likes and Reposts: a Combination of Text Mining and Empirical Analysis	Users on a super topic community relating to depression	Text was obtained from the super topic community relating to depression	Text data and user data for 49,047 posts in the super topic community relating to depression	Analyzing the value of text-based digital media using the LDA[h] topic model	• Social experience in posts (coefficient=0.368), emotional expression (coefficient=0.353), and the sentiment contained in the text (coefficient=0.002) all had significant positive relationships with the number of likes and reposts	Understanding how personal mental health and suicidal behavior are communicated
Falcone et al [14], 2020	Digital Conversations About Suicide Among Teenagers and Adults With Epilepsy: a Big‐Data, Machine Learning Analysis	Teenagers (aged 13 to 19 years) and adults (aged ≥20 years)	Open-source digital conversations across topical sites, blogs, social network, and message boards	222,000 unique conversations about epilepsy, including 9000 (4%) conversations related to suicide	Machine learning and text-based data analysis: thematic analysis	• Higher percentage of adults show a defeatist (given up) attitude compared to teenagers (42% vs 4%).	Understanding how personal mental health and suicidal behavior are communicated
Feuston and Piper [21], 2018	Beyond the Coded Gaze: Analysing Expression of Mental Health and Illness on Instagram	Users of the Instagram (Meta Platforms, Inc) social media platform	Posts relating to mental health from Instagram	>3000 posts	Analyzing the value of text-based digital media using constructivist grounded theory approach	• Semistructured interviews with 14 adults	Understanding how personal mental health and suicidal behavior are communicated
Waddell et al [18], 2023	Families' Experiences of Supporting Australian Veterans to Seek Help for a Mental Health Problem: a Linked Data Analysis of National Surveys with Families and Veterans	Veterans and family members	FWS[i] with linked MHWTS[j] data	1217 FWS respondents linked with 1123 MHWTS respondents	Analysis of survey data	• 53% of the respondents thought seeking help would negatively affect their career, and 63% were afraid to ask for help.	Help seeking for mental health difficulties
Gu et al [29], 2023	An Analysis of Cognitive Change in Online Mental Health Communities: a Textual Data Analysis Based on Post Replies of Support Seekers	Users of web-based mental health communities	Reply data from web-based mental health communities	31,935 replies to comments	Machine learning for prediction using TextCNN[k]	• TextCNN • Precision: 0.84 • Recall: 0.84 • *F*1-score: 0.84	Efficacy of interventions to support mental well-being
Onyeaka et al [23], 2021	Use of Smartphones, Mobile Apps and Wearables for Health Promotion by People With Anxiety or Depression: an Analysis of a Nationally Representative Survey Data	Respondents to HINTS[l] 5	Data from HINTS 5	5438 respondents	Analysis of digital interventions using chi-square tests		Efficacy of interventions to support mental well-being

Study and year	Title	Population	Data	Volume	Type of analysis	Outcome metrics	Theme
						• Respondents with anxiety or depression were generally more likely than people without anxiety or depression, to report that their smart device had helped them in their discussions with their health care providers (42.7% vs 35.3%; P=.03).	
Chikersal et al [22], 2020	Understanding Client Support Strategies to Improve Clinical Outcomes in an Online Mental Health Intervention	Supporters (of clients) on the Silver Cloud (iCBT[m]) platform	Supporter messages to clients	234,735 supporter messages	Analyzing the value of text-based digital media using association rule mining	• Lower word count is more salient in more successful messages.	Efficacy of interventions to support mental well-being
Goldberg et al [15], 2020	Machine Learning and Natural Language Processing in Psychotherapy Research: Alliance as Example Use Case	Therapists and clients attending counseling sessions	Recordings from sessions with clients and therapists	Recordings from 1235 sessions	Machine learning and text-based data analysis using MSE[n] and Spearman rank correlation	• The model that used therapist text and extracted features using TF-IDF[o] performed the best overall, with an MSE of 0.67 and Spearman rank correlation coefficient of 0.15 (P<.001).	Efficacy of interventions to support mental well-being
Oyebode et al [30], 2020	Using Machine Learning and Thematic Analysis Methods to Evaluate Mental Health Apps Based on User Reviews	Users of mental health apps	Reviews by users of mental health apps	88,125 user reviews	Machine learning for prediction using SVM[p], LR[q], MNB[r], SGD[s], and RF	• SVM • Precision: 0.8940 • Recall: 0.8939 • F_1-score=0.8939 • LR • Precision: 0.8938 • Recall: 0.8937 • F_1-score=0.8937 • MNB • Precision: 0.8908 • Recall: 0.8908 • F_1-score=0.8907 • SGD • Precision: 0.8945 • Recall: 0.8943 • F_1-score=0.8942 • RF • Precision: 0.8769 • Recall: 0.8770 • F_1-score=0.8769	Efficacy of interventions to support mental well-being
Vermetten et al [24], 2020	Using VR-Based Interventions, Wearable Technology, and Text Mining to Improve Military and Veteran Mental Health	Military members and veterans	Self-narratives collected online	300 self-narratives collected online	Analysis of digital interventions	• The variable of the word "family" was found to be the most significant predictor in LIWC[t].	Efficacy of interventions to support mental well-being
Van Gemert-Pijnen et al [25], 2014	Understanding the Usage of Content in a Mental Health Intervention for Depression: an Analysis of Log Data	Users of the web-based intervention "Living to the Full"	Log data from the web-based intervention system	206 participants	Analysis of digital interventions using linear regression	• Linear regression yielded a significant model with log-in quartile as a significant predictor (explained variance was 2.7%).	Efficacy of interventions to support mental well-being

[a]HILDA: Household, Income, and Labor Dynamics in Australia.
[b]VADER: Valence Aware Dictionary and Sentiment Reasoner.
[c]OLS: ordinary least squares.
[d]SES: socioeconomic status.
[e]RF: random forest.
[f]NN: neural network.
[g]AUC: area under the curve.
[h]LDA: latent Dirichlet allocation.
[i]FWS: Family Well-Being Study.
[j]MHWTS: Mental Health Well-Being Transition Study.
[k]TextCNN: convolutional neural network for text.
[l]HINTS: Health Information National Trends Survey.
[m]iCBT: internet-based cognitive behavioral therapy.
[n]MSE: mean squared error.
[o]TF-IDF: term frequency–inverse document frequency.
[p]SVM: support vector machine.
[q]LR: logistic regression.
[r]MNB: multinomial naive Bayes.
[s]SGD: stochastic gradient descent.
[t]LIWC: Linguistic Inquiry and Word Count.

Another study [12] used logistic regression, with a 73% accuracy of the logistic model in detecting cognitive distortions. Linear regression was another method used in predicting depressive symptoms and yielded a significant model as a significant predictor of depression [25]. Machine learning was also used in a psychotherapy research study, where the model that used therapist text and extracted features using term frequency–inverse document frequency performed the best overall, with a mean squared error of 0.67 and Spearman rank correlation coefficient of 0.15 (P<.001) [15]. Association rule mining was used in analyzing survey data [19], where the top rule identified an association between strong disappointment with missing events and missing friends in person (support=0.286, confidence=0.671, and lift=1.454) due to the COVID-19 pandemic.

Sentiment was measured for various studies; it was measured as positive for a web-based community platform for mental health [26], and text had a positive score, which correlated with the number of likes [20] of the posts. Another survey [23] found that respondents with anxiety or depression were generally more likely to report that their smart device had helped them in their discussions with their health care providers, compared to respondents that did not have anxiety or depression (42.7% vs 35.3%; P=.03). Furthermore, a negative tone was observed in 66% of conversations among Hispanic populations compared to 39% of conversations among non-Hispanic populations [13], and the total causal effect of disability acquisition on mental health was estimated to be a 4.8-point decline in mental health [16]. Moreover, there was a negative trajectory in sentiment scores from a longitudinal analysis of Twitter data during the COVID-19 pandemic [27]. Another study [14] reported a higher percentage of adults with epilepsy showing a defeatist attitude compared to teenagers with epilepsy (42% vs 4%). In a family well-being study, 53% of respondents thought seeking help would negatively affect their career, and 63% were afraid to ask for help [18]. The results of a questionnaire to establish the mental health of Chinese web-based networkers found that with an increase in socioeconomic status, depression decreased by a margin of –0.52 (P<.001) [17].

Having identified the 19 papers for further analysis, we attempted to identify any themes within these papers. This involved an initial in-depth review to become familiarized with the text, and using simple coding to highlight sections of the texts that best describe the content, we were able to identify shorthand labels or codes, for example, prediction and detection of mental disorders and suicide risk. From the coding, we were then able to identify 5 themes as to how machine learning and data analysis techniques could be applied. The themes are outlined with the number of papers per theme in Table 2.

Table 2. Themes and number of papers per theme (N=19).

Theme	Name	Papers, n (%)
1	As predictors of personal mental health	4 (21)
2	To detect mental disorders and suicidal risk	2 (11)
3	To understand how personal mental health and suicidal behavior are communicated	5 (26)
4	To identify help seeking for mental health difficulties	1 (5)
5	To determine the efficacy of interventions to support mental well-being	7 (37)

Details of the 19 papers that were reviewed, including the author, year, title, population studied, data volume, and main themes, are provided in Table 1. The themes are further expanded in the subsequent sections.

Predictors of Personal Mental Health

Personal mental health can be influenced by various factors, such as employment status and income, and various analytical tools have been used to determine sentiment or other predictors of personal mental health. Research by Aitken et al [16] sought to determine the extent to which alterations in employment and income impact mental health. They used logistic regression models specifically for employment and income, considering their conditional relationship with disability acquisition. The analysis technique focused on evaluating the significance of text-based digital media; their findings indicated that 10.6% of the effect of disability acquisition on mental health was explained by changes in individuals' employment status, but no similar effect was observed through changes in income. This underscores the importance of measures for addressing disability-related mental health disparities, specifically the equalization of employment rates between individuals with and individuals without disabilities to reduce disability-related mental health inequalities.

Research by Xiao et al [17] sought to examine survey data to measure the prevalence of depression symptoms and their correlation with an individual's socioeconomic status and lifestyle during the COVID-19 pandemic in China. The methodology involved statistical analyses using SPSS (IBM Corp) to evaluate survey data. The findings revealed a noteworthy impact of the pandemic, indicating that respondents experienced more severe mental symptoms when their residential communities were more exposed to SARS-CoV-2. The implications drawn from these findings suggest that mental health conditions among survey respondents varied based on the level of the COVID-19 pandemic severity. Notably, residents in communities with a high severity of the pandemic exhibited more pronounced symptoms of depression and anxiety.

Khattar et al [19] conducted a web-based survey study with the goal of understanding the day-to-day experiences and mental well-being of young students in India during the COVID-19 pandemic. They analyzed survey responses using R (The R Foundation) and Python (Python Software Foundation) to evaluate the mental health of diverse populations during the ongoing COVID-19 pandemic. Their findings revealed that approximately 19.2% of the students expressed weariness with phone use, while 42.9% reported feeling a mix of frustration, profound boredom, anxiety, overwork, and depression. Conversely, 37.9% indicated experiencing emotions such as relaxation, peace, optimism, calmness, hopefulness, and love. This suggests a crucial role for teachers and mentors in providing emotional support to students. They also used association rule mining to analyze the survey data, where the top rule identified an association between strong disappointment with missing events and missing meeting friends in person (support=0.286, confidence=0.671, and lift=1.454) due to the pandemic.

Valdez et al [27] investigated the extent of social media use at the onset of the COVID-19 pandemic to uncover emerging themes from tweets related to COVID-19 and to examine whether sentiments changed in response to the COVID-19 crisis. They used the latent Dirichlet allocation method for topic modeling and the Valence Aware Dictionary and Sentiment Reasoner for sentiment analysis. Their findings indicated that sentiment scores were initially high and stable but exhibited a significant decrease over time, indicating reduced sentiment over the long term.

Various data analysis techniques have been applied as predictors of personal mental health, where the effect of disability acquisition on mental health, for example, was explained by changes to people's employment but not by changes to income [16]. In relation to the COVID-19 pandemic, the overall emotional state of students during lockdown showed a mix of various moods, with feelings ranging from frustration to boredom to anxiety to depression [17]. In addition, themes emerged from tweets about COVID-19 to highlight the extent to which social media use increased during the onset of the COVID-19 pandemic [19] and how the sentiment changed in response to the pandemic [27]. The pandemic has had a significant impact on mental health, where respondents had more serious mental symptoms when their residential communities exhibited a greater exposure to the spread of SARS-CoV-2 [17].

Detection of Mental Disorders and Suicidal Risk

Machine learning can be used in the detection of cognitive distortions, which may fuel anxiety, and in the detection of those at risk of suicide. Roy et al [28] developed a model capable of predicting individuals at risk and assessing the likelihood of experiencing suicidal thoughts within a specific time frame. This involved using a random forest model that used output from neural networks to predict binary suicidal ideation status when there is a match with at least one of the word patterns in the ordered word screening, for example, "feeling suicidal." This study found that the neural network models successfully predicted suicidal ideation even before individuals articulated explicit thoughts of suicide. These findings suggest that there may be potential for predicting suicidal ideation before individuals explicitly express such thoughts, offering opportunities for early intervention and support.

Simms et al [12] demonstrated that machine learning could also be applied to detecting cognitive distortions (eg, the user would be thinking negatively and discounting the positive) from personal blogs. Through the use of the Linguistic Inquiry and Word Count software, this study found that it is feasible to automatically detect cognitive distortions from personal blogs with a relatively high accuracy of 73%. The implications drawn from these findings underscore the potential benefits of continued work in this area for mental health care and psychotherapy. This progress has the potential to lead to lower costs, earlier detection, and more efficient use of counseling time.

These findings show that it is possible to detect cognitive distortions automatically from personal blogs with an accuracy of 73% [12], and this could lead to an earlier detection of anxiety and possible intervention at an earlier stage. Neural network models, which are powerful machine learning tools, have been shown to successfully detect mental disorders and suicidal risk, where certain models were shown to predict suicide ideation even before suicidal thoughts were articulated [28].

Understanding How Personal Mental Health and Suicidal Behavior Are Communicated

When attempting to understand how personal mental health and suicidal behavior are communicated, machine learning has been used to explore big data from open-source digital conversations with regard to suicidality. The aim of the research by Castilla-Puentes et al [13] was to delve into big data derived from open-source digital conversations among Hispanic populations to determine attitudes toward depression, comparing Hispanic and non-Hispanic populations. The methodology involved the analysis of tone, topic, and attitude relating to depression using machine learning and NLP. This study revealed a notable disparity in attitudes, beliefs, and treatment-seeking behavior between the 2 groups, providing insights into the mindset and attitudes toward depression from a previously unexplored vantage point.

Falcone et al [14] investigated big data derived from open-source digital conversations among teenagers and adults with epilepsy with regard to suicidality. They used NLP and text analytics to reveal that a higher percentage of teenagers, compared to adults, expressed a fear of "the unknown" due to seizures (63% vs 12%), concern about the social consequences of seizures (30% vs 21%), and desire for emotional support (29% vs 19%). In contrast, a significantly higher percentage of adults exhibited a defeatist ("given up") attitude compared to teenagers (42% vs 4%). The implications of this study suggest that teenagers engage more frequently in web-based conversations about suicide than adults and that there are notable differences in attitudes and concerns between the 2 groups. These distinctions may have implications for the treatment of younger patients with epilepsy.

Liu and Kong [20] sought to identify the factors influencing the number of likes and reposts within a web-based community dedicated to depression. This involved using a combination of text mining and empirical analysis to delve into the factors affecting user engagement, specifically the number of likes and reposts. They found that users within web-based mental health communities exhibit a higher level of attention to topics related to social experiences and emotional expressions. These findings emphasize that understanding the factors influencing the number of likes and reposts in web-based mental health communities can be advantageous for users, facilitating greater support and providing a sense of relief and comfort within the community.

Feuston and Piper [21] integrated manual data collection with digital ethnography (study of human interaction through the internet technologies used) and semistructured interviews to explore how various modes of expression (eg, visual, textual, and oral) contribute to the overall understanding of mental health. By evaluating the value of text-based digital media, they found that individuals adopt a diverse range of practices and use Instagram (Meta Platforms, Inc) features to render their experiences with mental health and illness visible to others. This would have implications for the analysis of user interactions, suggesting an information flow from one person to the next.

Golz et al [26] used the inCLOUsiv platform to identify and interpret the communication patterns and verbal expressions of the users of the platform during the initial lockdown in 2020. The methodology involved analyzing discussions in forums and live chats using text mining, frequency analysis, correlation analysis, n-gram analysis, and sentiment analysis. Their analysis found that the communication behavior of users on the inCLOUsiv platform was characterized by generosity and support, with 72% of the identified sentiments being positive. Users actively engaged with topics such as *corona*, *anxiety*, and *crisis*, sharing coping strategies, which suggest that positive and supportive interactions within mental health–related virtual communities, emphasizing the potential impact of such interactions on the well-being of community members.

When it comes to understanding how personal mental health and suicidal behavior are communicated, it was found that teenagers engage more frequently in web-based conversations about suicide than adults [14] and that the communication behavior of users on a digital exchange platform was supportive and sentiments were mostly positive [20]. Data analysis was also shown to reveal that individuals use a variety of practices and features of social media to make experiences with mental health and illness visible to others [21] and that users of web-based mental health communities were found to be more attentive to the topics of social experience and emotional expressions [20]. Furthermore, help seeking was shown to vary between different populations, where the attitudes, beliefs, and treatment-seeking behavior toward depression showed great disparity between Hispanic and non-Hispanic populations [13]. Finally, in relation to a specific illness, epilepsy, a higher percentage of teenagers were fearful of "the unknown" due to seizures and concerned about the social consequences of seizures, while a significantly higher percentage of adults showed a defeatist ("given up") attitude compared to teenagers [14].

Help Seeking for Mental Health Difficulties

An analysis of survey data has been shown to identify help seeking for mental health difficulties. Research by Waddell et al [18] sought to examine survey data to gain insights into the dynamics of help-seeking relationships within veteran families. The findings of the study brought to light that family members of veterans play a significant role in both the initial and ongoing processes of seeking help. However, the study also revealed substantial barriers to help seeking, primarily linked to the military culture. These barriers included the belief that mental health concerns could be self-managed (if recognized), highlighting concerns about potential impacts on careers and the fear of judgment by others. Educating families about identifying early signs of mental health problems is crucial to inform families about the potential mental health risks associated with military careers. This knowledge can then contribute to fostering a supportive environment and breaking down barriers to help seeking within veteran families.

Efficacy of Interventions to Support Mental Well-Being

The effectiveness of interventions to support mental well-being has also been analyzed using machine learning. Gu et al [29] used NLP technology to identify psychological cognitive changes. Using an emotion dictionary along with Word2vec semantic training, a model was trained to transform labeled text into a vector matrix, and the convolutional neural network for text was used for classifying the labeled text. The findings of the study indicated that posts signaling cognitive change tended to have longer word lengths. In addition, support seekers who had not undergone cognitive change tended to express themselves more in web-based replies. This highlights the potential for supporting individuals with mental health problems, promoting the development of web-based mental health communities, and constructing web-based psychological chatbots.

Research by Goldberg et al [15] used NLP and machine learning techniques to predict one of the most studied process variables in psychotherapy: therapeutic alliance. The methodology involved using Sent2vec to map sentences to vectors of real numbers, and linear regression was then used as the prediction model. The findings of the study revealed that across the 1235 alliance ratings, the mean rating was 5.47 (SD 0.83), indicating a negative slant often found in the assessment of therapeutic alliance. The implications drawn from these findings suggest that machine learning holds promise for predicting observable linguistic behaviors, these models could be trained using human coding as the gold standard, and thorough testing should be conducted using large data sets.

Oyebode et al [30] used sentiment analysis and other machine learning approaches to evaluate 104 mental health apps available on Google Play (Google LLC) and App Store (Apple Inc). By integrating NLP and the term frequency–inverse document frequency weighting technique to vectorize the reviews, supervised machine learning classifiers were used to predict sentiment. The study revealed that the majority of the reviews were positive, indicating that most users found mental health apps to be useful and helpful, emphasizing the importance of ensuring that mental health apps are not only usable and of high quality but also supportive, secure, and noninvasive.

Research by Chikersal et al [22] provided a deeper understanding of how supporter behaviors impact the use of web-based therapy programs. The methodology involved the application of unsupervised machine learning, along with statistical and data mining methods, to analyze complex, large-scale supporter-client interactions. They found that concrete, positive, and supportive feedback from supporters, particularly those referencing social behaviors, were strongly associated with better outcomes. This suggests the importance of identifying effective context-specific support strategies using data for personalized mental health support. This knowledge can contribute to improving the design and implementation of personalized human support in internet-based cognitive behavioral therapy and enhance our understanding of big data in digital health interventions.

Onyeaka et al [23] investigated the use and perceived benefits of digital health tools, identifying the association between the use of digital interventions and the adoption of healthy lifestyle behaviors, and the sociodemographic factors linked to the use of digital tools among individuals with anxiety or depression. Basic descriptive statistics and chi-square tests were used, identifying a notable prevalence of digital interest among individuals with anxiety or depression, with up to 84.7%, 60.6%, and 57.7% of the individuals reporting ownership of smartphones, tablets, and health apps, respectively. These results suggest that digital tools may offer promise for a subset of individuals with mental illness who prefer engaging in technology-based strategies for managing their health.

Vermetten et al [24] investigated the potential use of virtual reality (VR)–based interventions, wearable technology, and text mining to enhance the mental health of military personnel and veterans. Using text mining and the statistical technique of item response theory, they demonstrated that there was a high agreement of 82% with the diagnoses provided by psychiatrists and suggested that the combination of text mining and VR-based interventions holds promise as a valuable tool for psychological and psychiatric assessments in the future.

Van Gemert-Pijnen et al [25] demonstrated how log data could be used to comprehend the adoption of web-based interventions and provide value in improving the incorporation of content in such interventions. By performing a statistical analysis using SPSS, this study showed that pattern recognition could be used to customize the interventions based on use patterns from earlier lessons and act as an aid in supporting the adoption of content essential for therapy. Understanding how participants can derive greater benefits from the intervention and identifying the most effective combination of features can lead to enhancing the effectiveness of web-based interventions.

There are many ways in which data analysis can be used to support mental well-being; for example, textual data analysis can be used to signal cognitive change, where it has been found that the average word length within text is longer for posts that indicate a cognitive or emotional change [29]. Other analysis results indicate a high prevalence of digital interest among people with anxiety or depression [23], and when NLP and machine learning were used to predict therapeutic alliance, the mean rating showed a typical negative skew found in the assessment of the alliance [15]. VR-based interventions, wearable technology, and text mining are expected to be promising tools in psychiatric assessments in the future [24]. Regarding the use of log data to improve the uptake of a web-based intervention, user pattern recognition from earlier lessons can be applied to tailor the intervention and support the uptake of content essential for therapy [25]. For web-based and non–web-based mental health apps, the majority of the reviews from a study of mental health apps available on Google Play and the App Store were positive, showing that most users found mental health apps useful and helpful [30].

Discussion

Principal Findings

When attempting to discover useful insights from text-based digital media in relation to mental health and depression, machine learning and data analysis techniques can be applied in many different ways. They can be used as predictors of personal mental health, for example, to measure how an individual's socioeconomic status can relate to depression. With the increasing prevalence of mental health issues since the COVID-19 pandemic [31] and the need for effective suicide prevention strategies, using data analysis and machine learning techniques in textual digital media data research has demonstrated that the COVID-19 pandemic and its associated restrictions have resulted in increased depression, anxiety, and feelings of loneliness [32], but this sentiment improved following the news of vaccine rollout to defend against the virus [33]. The pandemic has made a big impact on research in this area, where findings show that students' overall emotional well-being reflected a combination of diverse moods, encompassing feelings of frustration, boredom, anxiety, and being overworked, and experiencing depression during the pandemic. Further themes that emerged from tweets related to the COVID-19 pandemic showed that social media use increased during the onset of the pandemic and that participants of a survey exhibited more pronounced mental health symptoms if their residential communities faced heightened exposure to the spread of SARS-CoV-2.

Machine learning and data analysis techniques can also be used to detect mental ill health and suicidal risk, where neural network models can be used to predict suicide ideation before suicidal thoughts are articulated and to generate models capable of predicting individuals who would be at risk of suicidal thoughts. These tools can also be used to comprehend help seeking for mental health difficulties. Survey data were analyzed to understand help seeking in relation to mental health, identifying that the role of the family is important in encouraging help seeking for war veterans and revealing substantial barriers to help seeking, particularly in relation to the military culture, such as the belief that mental health concerns can be self-managed (if recognized) and a fear of being judged by others.

When attempting to understand how we communicate personal mental health and suicidal behavior, machine learning techniques can be used in diverse ways, such as to explore digital conversations with regard to suicidality and to identify factors influencing the number of likes in a web-based community for depression. Users were shown to exhibit both benevolent and supportive communication behaviors, with predominantly positive sentiments, on a digital exchange platform. When examining a specific illness, epilepsy, it was revealed that a higher percentage of teenagers expressed a fear of the unknown associated with seizures and concern about the social consequences of seizures, and a higher percentage of adults demonstrated a defeatist attitude compared to teenagers. When Instagram was used to better understand how we can communicate personal mental health, it was disclosed that individuals use various practices and features on the platform to make their experiences with mental health and illness visible to others. Finally, seeking assistance was found to differ across different populations, with significant differences in attitudes, beliefs, and the propensity to seek treatment for depression observed between Hispanic and non-Hispanic populations.

Insights from data analysis and machine learning can be used to assist in the development of digital interventions, and the effectiveness of these interventions can be shown to provide support to people living with depression and improve mental well-being. Through textual data analysis, it was determined, for example, that posts signaling cognitive change exhibit longer word lengths and that support seekers who have not undergone cognitive change tend to express themselves more in web-based replies. Similarly, it was found that there was a heightened prevalence of digital interest among individuals with anxiety or depression. NLP and machine learning can also be used to predict therapeutic alliance between the patient and therapist.

When exploring the potential of VR-based interventions integrating wearable technology and text mining to enhance mental health, it emerged that text mining coupled with VR-based interventions is anticipated as a promising tool for psychological and psychiatric assessments in the future. The use of mental health apps was analyzed, which showed that attitudes toward them were mainly positive, indicating that a majority of users find these apps useful and helpful. In the context of understanding the uptake of web-based interventions, pattern recognition was used to tailor individual interventions based on use patterns from earlier lessons, thereby supporting the uptake of content essential for therapy.

Limitations

This study exhibits limitations in the selection of articles because it used only 4 journal databases (ie, Web of Science, MEDLINE, Embase, and PsycINFO) as well as Google Scholar. Moreover, only articles published in English and related to mental health or suicide, machine learning and data analysis, and digital interventions were included. The search for articles started in March 2023, and the collected articles were published between 2013 and 2023. As some of the researched articles identified some sort of machine learning classification or prediction, we should have considered explainable artificial intelligence to facilitate the understanding of any predictions made by the machine learning models to better understand the models' behavior. Another limitation involves how the inclusion and exclusion of papers were resolved. Even though CS, EE, MDM, and RB assessed the papers and decided what was to be included or excluded based on the applicability criteria, it was CS who made the final decision about what went into this review.

Conclusions

In conclusion, this review illustrates that the use of data analysis and machine learning techniques to extract useful insights from text-based digital media related to mental health and suicide prevention holds significant promise. Data analysis and machine learning were used to gain valuable insights; for example, findings show that engagement in web-based conversations relating to depression may vary among different ethnic groups and that teenagers engage in web-based conversations about suicide more often than adults. Another finding was that disability acquisition (which is associated with a deterioration in mental health) was shown to be affected by changes to employment but not income.

The efficacy of digital tools was also analyzed, with machine learning approaches being used to understand users' opinions regarding mental health apps. Using positive and negative sentiments, it was shown that those with mental illness are digitally connected and are incorporating these tools to manage their health. Predictive analytics was also identified to be able to detect cognitive distortions, which are associated with depression and anxiety, from personal blogs with an accuracy of 73%, while other machine learning models were able to predict the risk of suicidal ideation from social media. The use of modern technology has also been investigated, with the application of VR-based interventions showing promising contributions to the field of military and veteran mental health by developing new approaches to delivering preventive or curative care.

The recent pandemic has also had an influence on this area of research. Analysis was undertaken to try to discover to what extent social media use increased during the onset of the COVID-19 pandemic and to assess how different populations communicated regarding their mental health. It was discovered that virtual communities played an important role in mental health during the pandemic and that social media may be used as a coping mechanism to combat feelings of isolation related to long-term social distancing. Web-based communities also offer great support for people with mental disorders, where the analysis of the number of likes and reposts for posts in web-based mental health communities allowed for these users to gain more support within the community.

Future research could focus on investigating further benefits of textual digital media analysis in mental health and suicide prevention when dealing with depression and, importantly, what makes people happy. Machine learning can be used to predict what are the sources of "happiness" or even how different activities make different socioeconomic groups "happy," and these insights can then be used to assist in the development of a wide range of digital interventions, such as chatbots.

Ultimately, this systematic review underscores the importance of harnessing advanced analytical methods to derive valuable insights that can lead to improved mental health interventions and enhanced strategies for suicide prevention.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Electronic database search criteria.

[DOCX File , 17 KB - mental_v11i1e55747_app1.docx]

Multimedia Appendix 2

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.

[DOCX File , 33 KB - mental_v11i1e55747_app2.docx]

Multimedia Appendix 3

Details relating to the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) checklist.

[DOCX File , 52 KB - mental_v11i1e55747_app3.docx]

References

1. Coppersmith G, Dredze M, Harman C. Quantifying mental health signals in Twitter. In: Proceedings of the 2014 Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. 2014 Presented at: EMNLP '14; June 17, 2014; Baltimore, MA p. 51-60 URL: https://aclanthology.org/W14-3207.pdf [doi: 10.3115/v1/w14-3207]
2. De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M. Discovering shifts to suicidal ideation from mental health content in social media. Proc SIGCHI Conf Hum Factor Comput Syst 2016 May;2016:2098-2110 [FREE Full text] [doi: 10.1145/2858036.2858207] [Medline: 29082385]
3. Park M, Cha C, Cha M. Depressive moods of users portrayed in Twitter. In: Proceedings of the 2012 ACM SIGKDD Workshop on Healthcare Informatics. 2012 Presented at: HIKDD '12; August 12, 2012; Beijing, China p. 1-8 URL: https://nyuscholars.nyu.edu/ws/files/134720119/depressive_moods_kdd.pdf
4. Sik D, Németh R, Katona E. Topic modelling online depression forums: beyond narratives of self-objectification and self-blaming. J Ment Health 2023 Apr;32(2):386-395 [FREE Full text] [doi: 10.1080/09638237.2021.1979493] [Medline: 34582309]
5. Burnap P, Colombo G, Amery R, Hodorog A, Scourfield J. Multi-class machine classification of suicide-related communication on Twitter. Online Soc Netw Media 2017 Aug;2:32-44 [FREE Full text] [doi: 10.1016/j.osnem.2017.08.001] [Medline: 29278258]

6. Pigoni A, Delvecchio G, Turtulici N, Madonna D, Pietrini P, Cecchetti L, et al. Machine learning and the prediction of suicide in psychiatric populations: a systematic review. Transl Psychiatry 2024 Mar 09;14(1):140 [FREE Full text] [doi: 10.1038/s41398-024-02852-9] [Medline: 38461283]
7. O'Dea B, Wan S, Batterham P, Calear A, Paris C, Christensen H. Detecting suicidality on Twitter. Internet Interv 2015 May;2(2):183-188 [FREE Full text] [doi: 10.1016/j.invent.2015.03.005]
8. Kalucy R, Thomas L, King D. Changing demand for mental health services in the emergency department of a public hospital. Aust N Z J Psychiatry 2005;39(1-2):74-80 [FREE Full text] [doi: 10.1080/j.1440-1614.2005.01512.x] [Medline: 15660708]
9. Sweeney C, Potts C, Ennis E, Bond R, Mulvenna M, O'neill S, et al. Can chatbots help support a person's mental health? Perceptions and views from mental healthcare professionals and experts. ACM Trans Comput Healthcare 2021 Jul 15;2(3):1-15 [FREE Full text] [doi: 10.1145/3453175]
10. Rethlefsen M, Kirtley S, Waffenschmidt S, Ayala A, Moher D, Page M, PRISMA-S Group. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 2021 Jan 26;10(1):39 [FREE Full text] [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
11. Collins G, Reitsma J, Altman D, Moons K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015 Jan 06;162(1):55-63 [FREE Full text] [doi: 10.7326/m14-0697]
12. Simms T, Ramstedt C, Rich M, Richards M, Martinez T, Giraud-Carrier C. Detecting cognitive distortions through machine learning text analytics. In: Proceedings of the 2017 IEEE International Conference on Healthcare Informatics. 2017 Presented at: ICHI '17; August 23-26, 2017; Park City, UT p. 508-512 URL: https://ieeexplore.ieee.org/document/8031202 [doi: 10.1109/ichi.2017.39]
13. Castilla-Puentes R, Dagar A, Villanueva D, Jimenez-Parrado L, Valleta L, Falcone T. Digital conversations about depression among Hispanics and non-Hispanics in the US: a big-data, machine learning analysis identifies specific characteristics of depression narratives in Hispanics. Ann Gen Psychiatry 2021 Nov 29;20(1):50 [FREE Full text] [doi: 10.1186/s12991-021-00372-0] [Medline: 34844618]
14. Falcone T, Dagar A, Castilla-Puentes R, Anand A, Brethenoux C, Valleta L, et al. Digital conversations about suicide among teenagers and adults with epilepsy: a big-data, machine learning analysis. Epilepsia 2020 May;61(5):951-958 [FREE Full text] [doi: 10.1111/epi.16507] [Medline: 32383797]
15. Goldberg S, Flemotomos S, Martinez V, Tanana M, Kuo P, Pace B, et al. Machine learning and natural language processing in psychotherapy research: alliance as example use case. J Couns Psychol 2020 Jul;67(4):438-448 [FREE Full text] [doi: 10.1037/cou0000382] [Medline: 32614225]
16. Aitken Z, Simpson J, Bentley R, Kavanagh A. How much of the effect of disability acquisition on mental health is mediated through employment and income? A causal mediation analysis quantifying interventional indirect effects using data from four waves of an Australian cohort study. BMJ Open 2021 Nov 22;11(11):e055176 [FREE Full text] [doi: 10.1136/bmjopen-2021-055176] [Medline: 34810192]
17. Xiao Y, Bian Y, Zhang L. Mental health of Chinese online networkers under COVID-19: a sociological analysis of survey data. Int J Environ Res Public Health 2020 Nov 28;17(23):8843 [FREE Full text] [doi: 10.3390/ijerph17238843] [Medline: 33260696]
18. Waddell E, Rioseco P, Van Hooff M, Daraganova G, Lawrence D, Rikkers W, et al. Families' experiences of supporting Australian veterans to seek help for a mental health problem: a linked data analysis of national surveys with families and veterans. J Ment Health 2023 Oct;32(5):899-909 [FREE Full text] [doi: 10.1080/09638237.2023.2182418] [Medline: 36880331]
19. Khattar A, Jain PR, Quadri SM. Effects of the disastrous pandemic COVID 19 on learning styles, activities and mental health of young Indian students - a machine learning approach. In: Proceedings of the 4th International Conference on Intelligent Computing and Control Systems. 2020 Presented at: ICICCS '20; May 13-15, 2020; Madurai, India p. 1190-1195 URL: https://ieeexplore.ieee.org/abstract/document/9120955 [doi: 10.1109/iciccs48265.2020.9120955]
20. Liu J, Kong J. Why do users of online mental health communities get likes and reposts: a combination of text mining and empirical analysis. Healthcare (Basel) 2021 Aug 31;9(9):1133 [FREE Full text] [doi: 10.3390/healthcare9091133] [Medline: 34574907]
21. Feuston JL, Piper AM. Beyond the coded gaze: analyzing expression of mental health and illness on Instagram. Proc ACM Hum Comput Interact 2018 Nov;2(CSCW):1-21 [FREE Full text] [doi: 10.1145/3274320]
22. Chikersal P, Belgrave D, Doherty G, Enrique A, Palacios J, Richards D, et al. Understanding client support strategies to improve clinical outcomes in an online mental health intervention. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI '20; April 25-30, 2020; Honolulu, HI p. 1-16 URL: https://doi.org/10.1145/3313831.3376341 [doi: 10.1145/3313831.3376341]
23. Onyeaka H, Firth J, Kessler R, Lovell K, Torous J. Use of smartphones, mobile apps and wearables for health promotion by people with anxiety or depression: an analysis of a nationally representative survey data. Psychiatry Res 2021 Oct;304:114120 [FREE Full text] [doi: 10.1016/j.psychres.2021.114120] [Medline: 34303946]

24. Vermetten E, Tielman M, van Dort E, Binsch O, Li X, Rozendaal M, et al. Using VR-based interventions, wearable technology, and text mining to improve military and Veteran mental health. J Mil Veteran Fam Health 2020 Mar 01;6(S1):26-35 [FREE Full text] [doi: 10.3138/jmvfh.2019-0033]
25. Van Gemert-Pijnen JE, Kelders SM, Bohlmeijer ET. Understanding the usage of content in a mental health intervention for depression: an analysis of log data. J Med Internet Res 2014 Jan 31;16(1):e27 [FREE Full text] [doi: 10.2196/jmir.2991] [Medline: 24486914]
26. Golz C, Richter D, Sprecher N, Gurtner C. Mental health-related communication in a virtual community: text mining analysis of a digital exchange platform during the Covid-19 pandemic. BMC Psychiatry 2022 Jun 25;22(1):430 [FREE Full text] [doi: 10.1186/s12888-022-04080-1] [Medline: 35752758]
27. Valdez D, Ten Thij M, Bathina K, Rutter LA, Bollen J. Social media insights into US mental health during the COVID-19 pandemic: longitudinal analysis of Twitter data. J Med Internet Res 2020 Dec 14;22(12):e21418 [FREE Full text] [doi: 10.2196/21418] [Medline: 33284783]
28. Roy A, Nikolitch K, McGinn R, Jinah S, Klement W, Kaminsky ZA. A machine learning approach predicts future risk to suicidal ideation from social media data. NPJ Digit Med 2020;3:78 [FREE Full text] [doi: 10.1038/s41746-020-0287-6] [Medline: 32509975]
29. Gu D, Li M, Yang M, Gu Y, Zhao Y, Liang C, et al. An analysis of cognitive change in online mental health communities: a textual data analysis based on post replies of support seekers. Inf Process Manag 2023 Mar;60(2):103192 [FREE Full text] [doi: 10.1016/j.ipm.2022.103192]
30. Oyebode O, Alqahtani F, Orji R. Using machine learning and thematic analysis methods to evaluate mental health apps based on user reviews. IEEE Access 2020;8:111141-111158. [doi: 10.1109/access.2020.3002176]
31. Winkler P, Formanek T, Mlada K, Kagstrom A, Mohrova Z, Mohr P, et al. Increase in prevalence of current mental disorders in the context of COVID-19: analysis of repeated nationwide cross-sectional surveys. Epidemiol Psychiatr Sci 2020 Sep 29;29:e173 [FREE Full text] [doi: 10.1017/S2045796020000888] [Medline: 32988427]
32. Kelly BD. Impact of COVID-19 on mental health in Ireland: evidence to date. Ir Med J 2020;113(10):214 [FREE Full text]
33. Weger R, Lossio-Ventura JA, Rose-McCandlish M, Shaw JS, Sinclair S, Pereira F, et al. Trends in language use during the COVID-19 pandemic and relationship between language use and mental health: text analysis based on free responses from a longitudinal study. JMIR Ment Health 2023 Mar 01;10:e40899 [FREE Full text] [doi: 10.2196/40899] [Medline: 36525362]

Abbreviations

NLP: natural language processing
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
VR: virtual reality

Edited by J Torous; submitted 22.12.23; peer-reviewed by A AL-Asadi, A Pigoni; comments to author 25.03.24; revised version received 16.05.24; accepted 17.05.24; published 27.06.24.

Please cite as:
Sweeney C, Ennis E, Mulvenna MD, Bond R, O'Neill S
Insights Derived From Text-Based Digital Media, in Relation to Mental Health and Suicide Prevention, Using Data Analysis and Machine Learning: Systematic Review
JMIR Ment Health 2024;11:e55747
URL: https://mental.jmir.org/2024/1/e55747
doi:10.2196/55747
PMID:38935419

Review

Application of Positive Psychology in Digital Interventions for Children, Adolescents, and Young Adults: Systematic Review and Meta-Analysis of Controlled Trials

Sundas Saboor[1], MBBS, MPH; Adrian Medina[2], MPH, EdM; Laura Marciano[3], PhD

[1]Department of Health Behavior & Health Education, University of Michigan School of Public Health, Ann Arbor, MI, United States

[2]Deptartment of Social & Behavioral Sciences, Harvard T.H. Chan School of Public Health, Boston, MA, United States

[3]Lee Kum Sheung Center for Health and Happiness, Department of Social and Behavioral Sciences, Harvard T.H. Chan School of Public Health, Boston, MA, United States

Corresponding Author:
Laura Marciano, PhD
Lee Kum Sheung Center for Health and Happiness
Department of Social and Behavioral Sciences
Harvard T.H. Chan School of Public Health
677 Huntington Avenue
Boston, MA, 02115
United States
Phone: 1 6175828025
Email: lmarciano@hsph.harvard.edu

Abstract

Background: The rising prevalence of mental health issues in children, adolescents, and young adults has become an escalating public health issue, impacting approximately 10%-20% of young people on a global scale. Positive psychology interventions (PPIs) can act as powerful mental health promotion tools to reach wide-ranging audiences that might otherwise be challenging to access. This increased access would enable prevention of mental disorders and promotion of widespread well-being by enhancing self-efficacy, thereby supporting the achievement of tangible objectives.

Objective: We aimed to conduct a comprehensive synthesis of all randomized controlled trials and controlled trials involving children, adolescents, and young adults, encompassing both clinical and nonclinical populations, to comprehensively evaluate the effectiveness of digital PPIs in this age group.

Methods: After a literature search in 9 electronic databases until January 12, 2023, and gray literature until April 2023, we carried out a systematic review of 35 articles, of which 18 (51%) provided data for the meta-analysis. We included randomized controlled trials and controlled trials mainly based on web-based, digital, or smartphone-based interventions using a positive psychology framework as the main component. Studies included participants with a mean age of <35 years. Outcomes of PPIs were classified into indicators of well-being (compassion, life satisfaction, optimism, happiness, resilience, emotion regulation and emotion awareness, hope, mindfulness, purpose, quality of life, gratitude, empathy, forgiveness, motivation, and kindness) and ill-being (depression, anxiety, stress, loneliness, and burnout). PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines were used for the selection of studies and data extraction. Quality assessment was performed following the CONSORT (Consolidated Standards of Reporting Trials) guidelines.

Results: For well-being outcomes, meta-analytic results showed that PPIs augmented the feeling of purpose, gratitude, and hope (Hedges g=0.555), compassion (Hedges g=0.447), positive coping behaviors (Hedges g=0.421), body image–related outcomes (Hedges g=0.238), and positive mindset predisposition (Hedges g=0.304). For ill-being outcomes, PPIs reduced cognitive biases (Hedges g=–0.637), negative emotions and mood (Hedges g=–0.369), and stress levels (Hedges g=–0.342). Of note, larger effect sizes were found when a waiting list control group was considered versus a digital control group. A funnel plot showed no publication bias. Meta-regression analyses showed that PPIs tended to show a larger effect size on well-being outcomes in studies including young adults, whereas no specific effect was found for ill-being outcomes.

Conclusions: Revised evidence suggests that PPIs benefit young people’s well-being and mitigate ill-being symptoms. Digital platforms offer a unique way to address their mental health challenges, although not without limitations. Future research should explore how they work for the needs of the young population and further examine what specific PPIs or combination of interventions is most beneficial with respect to other digital control groups.

Trial Registration: PROSPERO International Prospective Register of Systematic Reviews CRD42023420092; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=420092

(JMIR Ment Health 2024;11:e56045) doi:10.2196/56045

KEYWORDS

positive psychology; digital interventions; ill-being; well-being; systematic review; meta-analysis; smartphone; mobile phone

Introduction

Background

Mental health problems among children, adolescents, and young adults are a growing public health concern, affecting 10% to 20% of young people worldwide [1]. According to World Mental Health Report 2022, 970 million people have mental disorders, with 3% to 7% of mental disorders among those aged <10 years, 13.5% to 14.7% among those aged 10 to 19 years, and 14.1% to 14.9% among those aged 20 to 49 years [2]. Globally, 1 in 7 (14%) people who are aged 10 to 19 years have different mental health conditions, and most of them remain untreated [3]. In addition, among three-quarters of adults, long-term mental health problems occurred before the age of 24 years [1]. According to the US Surgeon General's Advisory Report, from 2009 to 2019, the proportion of high school students with persistent feelings of sadness or hopelessness increased by 40% [4]. In addition, the mental health conditions of young people faced unprecedented challenges during the COVID-19 pandemic, wherein the risk of depression and anxiety doubled [4], together with feelings of loneliness [5,6].

Considering the above statistics, enhancing youth well-being is an urgent public health need and concern today [7]. To date, the field of psychiatry and psychology have primarily addressed challenges in treating mental illness, focusing on therapy access and engagement [8] only after a symptomatology has occurred. However, there has been less attention on enhancing and protecting mental well-being [8,9] before the onset of mental health issues. Interventions aiming at diminishing mental health problems by using the prevailing disease model of human functioning (ie, ill-being model) largely ignore positive psychological assets such as meaning, courage, compassion, and kindness that could not only relieve ill-being states but also prevent them [10]. Positive psychology aims to provide a more comprehensive scientific knowledge of the human experience, from positive to negative, and better integrate and complement the illness framework with concepts related to positive mental health and use them at scale to solve public health issues [10].

Advocating for a more holistic approach to mental health promotion and expanding the field's focus to include strategies for improving mental well-being are crucial to boost intervention effectiveness, prevent mental illness and relapse, and broaden our understanding of how to support individuals in flourishing and improving their overall quality of life [9]. Tomé et al [11] conducted a systematic review on 13 studies, focusing on children and adolescents aged 0 to 18 years who had been a target audience for mental health and well-being promotion interventions and suggested that preventive school-based interventions can reduce the onset and progression of clinical disorders and promote good mental health. Another systematic review and meta-analysis of 16 studies concluded that people with severe mental illness benefit from positive psychology interventions (PPIs) in terms of enhancing mental health [12]. Williams et al [13] proposed in their systematic review that social interventions to increase positive emotions for people diagnosed with mental health disorders are suitable and effective adjuncts to mental health treatment.

Hence, in this systematic review and meta-analysis, we aimed to focus on digital interventions based on positive psychology as a promising option to promote well-being and prevent mental health issues in a population (children, adolescents, and young adults) that is not at a high risk of developing such issues.

Positive Psychology Framework

According to the American Psychological Association, *positive psychology* is defined as "a field of psychological theory and research that focuses on the psychological states (e.g., contentment, joy), individual traits or character strengths (e.g., intimacy, integrity, altruism, wisdom), and social institutions that enhance subjective well-being and make life most worth living" [14]. It is the scientific study of optimal functioning that identifies skills and strengths so that an individual or a community can thrive [15]. Positive psychology complements and extends the ill-being framework: PPIs focus on the science of positive mental states and behavioral patterns to improve quality of life [16,17]. Positive psychology involves the promotion of well-being differentiated between hedonic well-being, focusing on happiness, pleasure attainment, and pain avoidance, and eudaimonic well-being, related to meaning, self-realization, and full functioning of the person [18]. By doing so, positive psychology solves problems by identifying and leveraging individual and societal strengths [19]. Also, positive psychology enhances the importance of flourishing, a construct that encompasses positive emotions and relationships, engagement, meaning, and accomplishments directly or indirectly related to different dimensions of well-being, including psychological, emotional, social, and subjective [20].

Martin EP Seligman, the father of positive psychology, introduced 5 dimensions essential for well-being known as the PERMA model: positive emotions, engagement, relationships, meaning, and accomplishment [21,22]. Positive emotions (eg, joy, interest, contentment, and love) serve as markers of flourishing and optimal well-being [23]. Engagement is the extent of use and a subjective experience characterized by interest, affect, and attention [24]. Positive relationships can be regarded as strong connections with family and friends, developing a sense of belonging [25]. Meaning is understood as coherence, purpose, and a sense of life's inherent value, making it worth living [26]. Accomplishment refers to achievement, mastery, and competence [27].

Digital PPIs

On the basis of positive psychology theories and empirical research, PPIs aim to ameliorate well-being and health outcomes by increasing positive feelings, healthier lifestyle behaviors, and better cognitive functioning [28]. PPIs promote positive well-being outcomes, especially in people dealing with stress, by fostering positive daily emotions, providing people with stress-free time, mindful attention and positive cognition, and strengthening social relationships based on the Positive Pathways to Health theoretical model [29-31]. This theoretical model posits that PPIs promote physical and psychological well-being for people dealing with stress by elevating positive emotions experienced in their daily lives [30]. PPIs rely on elements such as optimism, spirituality, hopefulness, happiness, gratitude, creativity, meaning, and purpose [32].

From a public health perspective, PPIs can serve as effective mental health promotion tools to reach large target audiences, which would be challenging to reach otherwise. PPIs can be used as preventive and easily accessible tools that can promote well-being at scale by building self-efficacy and reinforcing the effects of meeting concrete goals [33]. Health promotion strategies can address complex mental health issues, treat preclinical and underdiagnosed cases, and prevent the recurrence of health problems to sustain health networks [34]. These strategies bolster public policies such as providing employment opportunities and antidiscriminatory laws, establishing supportive environments through interventions such as parenting programs, strengthening community action through initiatives such as media campaigns and research, and improving health services such as depression screening, all aimed at enhancing health and well-being [10].

Although mental health problems are a growing public health concern among youth, research on the impact of digital PPIs on this population remains limited. Prior reviews primarily addressed conventional interventions, such as in-person therapies within clinical settings and nonclinical settings [35-38]. However, considering the ongoing digitalization of health care, web-based resources and mental health applications have emerged as new avenues for young individuals to access health care services. Surprisingly, there is a notable absence of previous reviews exclusively focusing on digital PPIs for this demographic.

In previous reviews, individual meta-analyses for interventions were assessed for behavioral interventions [39] and ecological momentary interventions [40]. Other meta-analyses were performed for well-being components individually, for example, optimism [41], anxiety [42,43], depression [44-48], well-being [46,47,49], employee or work-based well-being [50-52], happiness [49], and school-based well-being [53]. However, previous reviews and meta-analyses excluded studies that did not mention outcome measure of well-being [54]; had no restriction on age groups [47,55]; or included only adults [56,57] or clinical population (eg, cardiovascular disease, psychiatric or somatic disorder, medical patients, schizophrenia, severe mental illness, and chronic pain) [12,42,58-61]. Although these reviews provide in-depth analysis of the effects of PPIs, the effects of digital PPIs on children, adolescents, and young adults have not been consistently summarized. Moreover, previous reviews mainly included traditional interventions (eg, cognitive therapy or cognitive behavioral therapy [CBT], mindfulness CBT, face-to-face group therapies and meditation, mainly among college students, young community members, or pediatric clinical settings) [41,47,62]; however, with the digitalization process, web-based resources and mental health apps are becoming the new way for youth to access health services. However, no previous reviews focused only on digital PPIs. Finally, when it comes to included studies, few reviews and meta-analyses included design such as randomized controlled trials (RCTs) and controlled trials (CTs) [41,47,62].

Study Aim

To overcome the abovementioned limitations, our objective was to comprehensively synthesize all RCTs and CTs conducted with young population (ie, children, adolescents, and young adults), encompassing both clinical and nonclinical populations, in order to assess the global effectiveness of digital PPIs on individuals in this age group holistically, without differentiation of prevention and treatment. In particular, we aimed to carry out a systematic review and meta-analysis including both clinical and nonclinical population to determine the efficacy of digital PPIs by considering if digital PPIs maintain health (by improving well-being constructs of compassion, life satisfaction, optimism, happiness, hope, resilience, etc or by reducing ill-being constructs of depression, anxiety, stress, loneliness, burnout, etc) and if there is any difference with respect to other (digital) control conditions.

In this study, we described study characteristics, theoretical background of the PPIs, quality assessment of the studies, the diverse range of PPIs used by the studies, the well-being and ill-being outcomes of these PPIs, and the meta-analytic results for the outcomes.

Methods

Overview

This systematic review and meta-analysis followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [63].

Study Sources and Searches

In total, 9 electronic databases (Communication & Mass Media Complete, Psychology and Behavioral Sciences Collection, PsycINFO, CINAHL, ERIC: Education Resource Information Center, MEDLINE Proquest, ProQuest Sociology, Web of Science [ISI Web of Knowledge], and MEDLINE PubMed) were searched up to January 12, 2023 (Textbox 1).

All citations were imported into Zotero reference manager (Corporation for Digital Scholarship) to automatically remove any duplicates. An additional hand search was carried out by scanning the references of relevant review articles identified along with all gray literature in Google Scholar until April 2023.

Textbox 1. Keywords used in the search string combined by the Boolean operator "AND."

Keywords

- Online* OR internet* OR digital* OR smartphone* OR social media OR EMI* OR EMA* OR in-situ OR ecological momentary assess* OR ecological momentary intervention* OR ESM* OR experience sampling* OR ambulatory assess* OR trace data OR chatbot* OR artificial intelligence* OR AI* OR conversational agent* OR chatterbot* OR virtual agent*
- positive psychology* OR positive psychotherapy* OR kindness* OR optimism* OR gratitude* OR happ* OR flourish* OR satisfaction* OR optimis* OR strength* OR forgiveness* OR positive relationship* OR savoring* OR altruism* OR gift* OR meaning* OR purpose* OR hedon* OR eudaimon* OR compassion* OR hop*
- Interven* OR treatment* OR therap* OR RCT* OR random* OR trial* OR control*

Study Selection

After duplicates were removed from the initial list of extracted publications, 2 authors independently completed title and abstract screening. For title and abstract screening, we included studies that (1) were either RCTs or CTs; (2) had the intervention that was mainly based on positive psychology (eg, gratitude, hope, optimism, etc) as the main component (for interventions, we included psychotherapies, therapy, interventions, mindfulness, training, exercise, and similar); and (3) included children, adolescents, and young adults with a mean age of <35 years. Although mental health disorders and treatment vary among this diverse age range, aggregating mental health across this range is appropriate due to similarities in mental health challenges and responses to interventions observed across different developmental stages within this age range. The proportion of individuals with the onset of any mental disorders before the ages of 14, 18, and 25 years were 34.6%, 48.4%, and 62.5%, respectively, and the peak age was 14.5 years [64]. Separation anxiety disorder, specific phobia, and social phobia have their mean onset before the age of 15 years, whereas agoraphobia, obsessive-compulsive disorder, posttraumatic stress disorder, panic disorder, and generalized anxiety disorder began, on average, between 21.1 and 34.9 years [65]. The mean age of <35 years is in line with the upper age limit of the early psychosis paradigm reporting on universal interventions or selective interventions [66-68]. In addition, studies needed to (4) include from both clinical and nonclinical population and (5) be carried out web-based or digital or through smartphone-based interventions. We excluded studies with face-to-face interventions and psychotherapy only of any kind, including digital or web-based or smartphone based. We also excluded studies in which positive psychology was not the main focus of the intervention (eg, when positive psychology was an additional component of a mindfulness-based intervention or CBT or other therapies). Moreover, we removed studies with no experimental design or control group and studies where the average age of the sample was >35 years or the focus was on caregivers. We further excluded conference abstracts, theses, books, or book sections. We excluded studies that were not in the English language. If at least 1 of the 2 authors decided to retain an article during the title and abstract screening process, we included it in the full-text screening. Discrepancies after full-text screening were solved through a consensus meeting with a third author.

Data Extraction

For each included study, we extracted information about the article (first author, year of publication, journal, and title); the study (country where the study was conducted, study design, sample size of experimental group, presence of control group, type of control group, sample size of control group, type of sampling, and attrition rates); and characteristics of the sample (including clinical or general population with details, ethnicity, gender distribution, and age). For intervention, we extracted information regarding the kind of positive intervention and its details, reference theory of the intervention, the type of activity and intervention, the setting of the intervention with details, the duration of the intervention, number of follow-ups and the time of the follow-ups, and data collection survey details. Outcomes included different ill-being and well-being constructs. Finally, we collected information on intervention evaluation, statistical analyses, and results to be converted into effect sizes.

Quality Assessment

In total, 2 authors independently assessed the quality of the studies according to the CONSORT (Consolidated Standards of Reporting Trials) guidelines [69], and a sum score was created, with a higher score indicating methodological quality. CONSORT guidelines are better suited for assessing the quality of study reporting for RCTs and CTs as recommended by Altman [70] and Versluis et al [40]. As a form of quality assessment, we checked whether studies explicitly mentioned (1) title and abstract; (2) introduction (including background and objectives); (3) methods (including trial design, participants, interventions, outcomes, sample size, randomization-sequence generation, randomization-sequence allocation concealment, randomization-implementation, blinding, and statistical methods); (4) results (including participants flow diagram, participant flow, recruitment, baseline data, number analyzed, outcomes and estimations, ancillary analyses, and harms); (5) discussion (including limitations, interpretations, generalizability, and registration); and (6) other information such as funding and protocol. For each paper, we rated if each criterion of the quality assessment was 0="absent" and 1="completely met." Then, we calculated a sum score of all criteria. The maximum score obtainable for each study was 34.

Data Synthesis and Analysis

We conducted the meta-analysis using "meta" [27] packages in R statistical software (R Foundation for Statistical Computing). A standardized mean difference approach was used as a measure of effect size using the Hedges adjusted *g*, which is similar to Cohen *d*, but it includes an adjustment for small sample bias [71]. All the analyses were implemented using the inverse-variance method with a random effects model using the Hartung-Knapp-Sidik-Jonkman adjustment [55] to limit the effect of studies' diversities. According to Cohen [72], a final effect falling in the ranges of 0 to 0.2, 0.3 to 0.5, and 0.6 to 0.8 was interpreted, respectively, as small, moderate, and large. Meta-analyses were run for well-being and ill-being outcomes separately, with additional specifications of the type of outcome. In particular, we further grouped well-being and ill-being outcomes in the following dimensions: body image related, cognitive bias/flexibility, compassion, coping, mindset predisposition, mood/affect/emotions, purpose/gratitude/hope, satisfaction/quality of life, stress, and 3 funny things/3 good things. A complete list of well-being and ill-being outcomes categorized in each of the abovementioned dimensions is reported in Table 1.

The heterogeneity of the effect size was computed with the between-study variance τ^2 and the Hartung-Knapp-Sidik-Jonkman adjustment, which allows to control for errors due to diversities in the sample sizes [73]. Heterogeneity level was interpreted as low (25%), moderate (50%), and high (75%), according to Higgins et al [74]. Potential publication biases were assessed by both funnel plots and Egger regression test for funnel plot asymmetry [75,76]. In addition, influence analyses were conducted to test if a single study accounted for a significant part of the variance in the final effect. Additional subgroup analyses were conducted to test if the effect size differed depending on the control group (waiting list vs digital control) when possible ($k \geq 2$ in each subgroup). To further explore the effect of age, meta-regression analyses were performed by different age categories (ie, children, adolescents, and young adults or a combination of these categories). In contrast to what we anticipated in our study protocol registered in PROSPERO, we could not run subgroup analyses to differentiate the effect of specific interventions due to the paucity of studies using the same PPIs.

Table 1. Well-being versus ill-being outcomes for meta-analysis.

Studies and original construct	Category	Outcome
Krifa et al [77]		
Absorption	Cognitive bias/flexibility	Well-being
Emotion regulation	Coping	Well-being
Dedication	Mindset predisposition	Well-being
Optimism	Mindset predisposition	Well-being
Vigor	Mindset predisposition	Well-being
Depression	Mood/affect/emotions	Ill-being
Anxiety	Mood/affect/emotions	Ill-being
Well-being	Mood/affect/emotions	Well-being
Hope	Purpose/gratitude/hope	Well-being
Stress	Stress	Ill-being
Drabu et al [78]		
Inclination to self-injury	Cognitive bias/flexibility	Ill-being
Pain endurance	Cognitive bias/flexibility	Well-being
Explicit self-criticism	Cognitive bias/flexibility	Ill-being
Implicit affect toward self	Mindset predisposition	Well-being
Lennard et al [79]		
Fear of compassion from others	Cognitive bias/flexibility	Ill-being
Fear of self-compassion	Cognitive bias/flexibility	Ill-being
Psychological flexibility	Cognitive bias/flexibility	Well-being
Self-compassion action	Compassion	Well-being
Self-compassion engagement	Compassion	Well-being
Compassion from others' action	Compassion	Well-being
Compassion from others' engagement	Compassion	Well-being
Adjustment anxiety	Mood/affect/emotions	Ill-being
Adjustment depression	Mood/affect/emotions	Ill-being
Breastfeeding satisfaction total	Satisfaction/quality	Well-being
Adjustment stress	Stress	Ill-being
Posttraumatic stress syndrome total	Stress	Ill-being
Andersson et al [80]		
Self-compassion	Compassion	Well-being
Emotion awareness/alexithymia	Mood/affect/emotions	Ill-being
Psychological problems of clinical origin	Mood/affect/emotions	Ill-being
Perceived stress	Stress	Ill-being
Beshai et al [81]		
Self-compassion	Compassion	Well-being
Dispositional mindfulness	Mindset predisposition	Well-being
Nonattachment	Mindset predisposition	Well-being
State mindfulness	Mindset predisposition	Well-being
Anxiety	Mood/affect/emotions	Ill-being
Depression	Mood/affect/emotions	Ill-being
Stress	Stress	Ill-being

Studies and original construct	Category	Outcome
Kelman et al [82]		
Inadequate self-compassion	Cognitive bias/flexibility	Ill-being
Self-reassurance	Compassion	Well-being
Affect	Mood/affect/emotions	Well-being
Kappen et al [83]		
Partner acceptance	Mindset predisposition	Well-being
Relationship satisfaction	Satisfaction/quality	Well-being
Koydemir and Sun-Selışık [84]		
Emotional well-being	Mood/affect/emotions	Well-being
Happiness	Mood/affect/emotions	Well-being
Psychological quality of life	Satisfaction/quality	Well-being
Social quality of life	Satisfaction/quality	Well-being
Satisfaction with life	Satisfaction/quality	Well-being
Sergeant and Mongrain [85]		
Maladaptive belief	Cognitive bias/flexibility	Ill-being
Engagement	Mindset predisposition	Well-being
Depression	Mood/affect/emotions	Ill-being
Meaning	Purpose/gratitude/hope	Well-being
Pleasure	Satisfaction/quality	Well-being
Lappalainen et al [86]		
Psychological flexibility behavior	Cognitive bias/flexibility	Well-being
Psychological flexibility openness	Cognitive bias/flexibility	Well-being
Psychological flexibility value	Cognitive bias/flexibility	Well-being
Total psychological flexibility	Cognitive bias/flexibility	Well-being
Self-compassion	Compassion	Well-being
Anxiety	Mood/affect/emotions	Ill-being
Depression	Mood/affect/emotions	Ill-being
Webb et al [87]		
Body image flexibility	Body image related	Well-being
Internal body shame	Body image related	Ill-being
External body shame	Body image related	Ill-being
Functional body appreciation	Body image related	Well-being
Functional body awareness	Body image related	Well-being
Functional body satisfaction	Body image related	Well-being
Physical activity behavior	Body image related	Well-being
Physical activity cognitive	Body image related	Well-being
Body appreciation	Body image related	Well-being
Weight bias	Body image related	Ill-being
Drive for leanness	Body image related	Ill-being
Self-compassion	Compassion	Well-being
Brouzos et al [88]		
Fear of COVID-19	Cognitive bias/flexibility	Ill-being
Resilience	Coping	Well-being

Studies and original construct	Category	Outcome
Empathic concern	Mindset predisposition	Well-being
Fantasy	Mindset predisposition	Well-being
Perspective taking	Mindset predisposition	Well-being
Emotional loneliness	Mood/affect/emotions	Ill-being
Positive affect	Mood/affect/emotions	Well-being
Negative affect	Mood/affect/emotions	Ill-being
Anxiety	Mood/affect/emotions	Ill-being
Depression	Mood/affect/emotions	Ill-being
Overall loneliness	Mood/affect/emotions	Ill-being
Social loneliness	Mood/affect/emotions	Ill-being
Personal distress	Stress	Ill-being
Greer et al [89]		
Anxiety	Mood/affect/emotions	Ill-being
Depression	Mood/affect/emotions	Ill-being
Positive emotion	Mood/affect/emotions	Well-being
Negative emotion	Mood/affect/emotions	Ill-being
Tagalidou et al [90]		
Subjective perceived change: coping humor	Coping	Well-being
Cheerfulness: coping humor	Coping	Well-being
Coping humor: coping humor	Coping	Well-being
Depression: coping humor	Coping	Ill-being
Bad mood: coping humor	Coping	Ill-being
Happiness: coping humor	Coping	Well-being
Seriousness: coping humor	Coping	Ill-being
Seriousness: 3 funny things	3 funny things/3 good things	Ill-being
Seriousness: 3 good things	3 funny things/3 good things	Ill-being
Coping humor: 3 funny things	3 funny things/3 good things	Well-being
Coping humor: 3 good things	3 funny things/3 good things	Well-being
Subjective perceived change: 3 funny things	3 funny things/3 good things	Well-being
Subjective perceived change: 3 good things	3 funny things/3 good things	Well-being
Cheerfulness: 3 funny things	3 funny things/3 good things	Well-being
Cheerfulness: 3 good things	3 funny things/3 good things	Well-being
Depression: 3 funny things	3 funny things/3 good things	Ill-being
Depression: 3 good things	3 funny things/3 good things	Ill-being
Bad mood: 3 funny things	3 funny things/3 good things	Ill-being
Bad mood: 3 good things	3 funny things/3 good things	Ill-being
Happiness: 3 funny things	3 funny things/3 good things	Well-being
Happiness: 3 good things	3 funny things/3 good things	Well-being
Bronk et al [91]		
Hope: purpose	Purpose/gratitude/hope	Well-being
Gratitude: purpose	Purpose/gratitude/hope	Well-being
Hope: gratitude	Purpose/gratitude/hope	Well-being
Gratitude: gratitude	Purpose/gratitude/hope	Well-being

Studies and original construct	Category	Outcome
Identified purpose: gratitude	Purpose/gratitude/hope	Well-being
Identified purpose: purpose	Purpose/gratitude/hope	Well-being
Prosocial intentions: gratitude	Purpose/gratitude/hope	Well-being
Prosocial intentions: purpose	Purpose/gratitude/hope	Well-being
Searching for purpose: gratitude	Purpose/gratitude/hope	Well-being
Searching for purpose: purpose	Purpose/gratitude/hope	Well-being
Gu et al [92]		
Compassion-focused theory: self-criticism	Cognitive bias/flexibility	Ill-being
Compassion-focused theory: sensitivity to others	Cognitive bias/flexibility	Ill-being
Compassion-focused theory: shame	Cognitive bias/flexibility	Ill-being
Rational emotive behavior therapy: self-criticism	Cognitive bias/flexibility	Ill-being
Rational emotive behavior therapy: sensitivity to others	Cognitive bias/flexibility	Well-being
Rational emotive behavior therapy: shame	Cognitive bias/flexibility	Ill-being
Rational emotive behavior therapy: tolerance of uncomfortable things	Cognitive bias/flexibility	Well-being
Compassion-focused theory: tolerance of uncomfortable things	Cognitive bias/flexibility	Well-being
Compassion-focused theory: compassion	Compassion	Well-being
Compassion-focused theory: self-compassion	Compassion	Well-being
Rational emotive behavior therapy: compassion	Compassion	Well-being
Rational emotive behavior therapy: self-compassion	Compassion	Well-being
Compassion-focused theory: kindness to others	Mindset predisposition	Well-being
Compassion-focused theory: kindness to self	Mindset predisposition	Well-being
Rational emotive behavior therapy: kindness to others	Mindset predisposition	Well-being
Rational emotive behavior therapy: kindness to self	Mindset predisposition	Well-being
Compassion-focused theory: anxiety	Mood/affect/emotions	Ill-being
Compassion-focused theory: depression	Mood/affect/emotions	Ill-being
Rational emotive behavior therapy: anxiety	Mood/affect/emotions	Ill-being
Rational emotive behavior therapy: depression	Mood/affect/emotions	Ill-being
Alexiou et al [93]		
Depersonalization	Cognitive bias/flexibility	Ill-being
Personal accomplishment	Mindset predisposition	Well-being
Depression	Mood/affect/emotions	Ill-being
Anxiety	Mood/affect/emotions	Ill-being
Emotional exhaustion	Mood/affect/emotions	Ill-being
Positive emotions	Mood/affect/emotions	Well-being
Negative emotions	Mood/affect/emotions	Ill-being
Satisfaction with life	Satisfaction/quality	Well-being
Stress	Stress	Ill-being
Manicavasagar et al [94]		
Depression	Mood/affect/emotions	Ill-being
Anxiety	Mood/affect/emotions	Ill-being

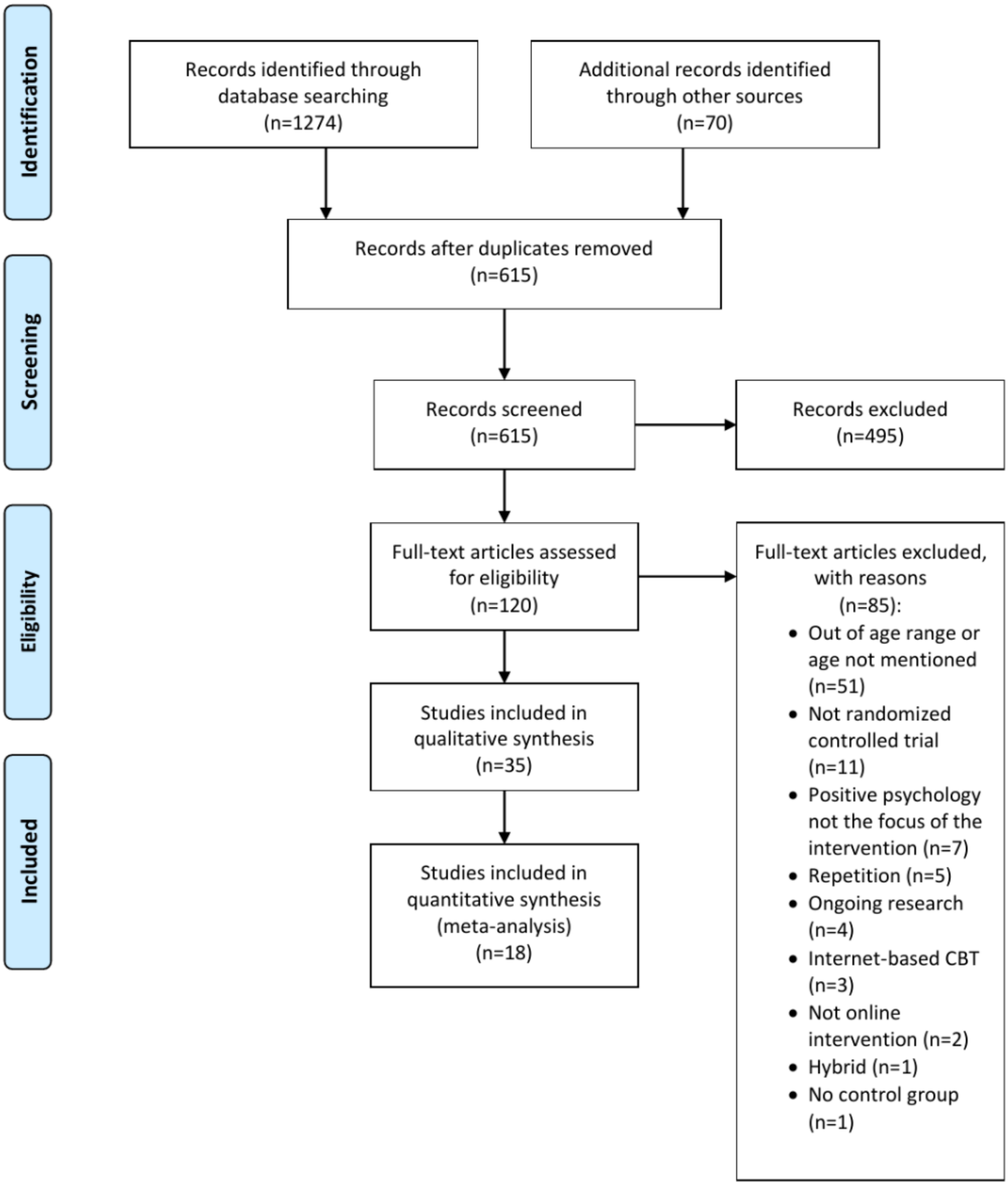
Studies and original construct	Category	Outcome
Well-being	Mood/affect/emotions	Well-being
Stress	Stress	Ill-being

Results

Overview

The study selection process is reported in the PRISMA flowchart (Figure 1). The initial database and hand search returned 1344 publications, of which 729 (54.24%) were duplicates, which were removed. After title and abstract screening of 615 (45.76%) records, we assessed 120 full-text articles for eligibility. We then excluded 85 (70.8%) articles, resulting in a qualitative assessment of 35 (29.2%) articles and a meta-analysis of 18 studies. The reasons for exclusion include out of age range or the age was not explicitly mentioned (n=51, 60%), positive psychology was not the main focus of the intervention (n=7, 8%), the CBT was internet based (n=3, 4%), the intervention was not web based (n=2, 2%), the trials were not RCTs (n=11, 13%) or CTs (n=1, 1%), the intervention was hybrid (n=1, 1%), repetitions (n=5, 6%), and the research was still ongoing (n=4, 5%). Cohen κ was calculated as a measure on intercoder reliability, and it was excellent (κ=0.95).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of included studies in systematic review and meta-analysis [95]. CBT: cognitive behavioral therapy.



Study Characteristics

The systematic review is based on 35 studies (Multimedia Appendix 1 [77-94,96-112]). Overall, the analytical sample amounts to 7341 participants, of which 19 studies (54%) mentioned young adults aged 20 to 35 years of age; 8 studies (23%) were focused on children, adolescents, and young adults aged up to 20 years of age; and 3 studies (9%) mentioned children, adolescents, and adults aged up to 35 years. A total of 3 studies (9%) mentioned young adults and adult participants, while 1 (3%) study mentioned children, adolescents, and adults, and 1 study (3%) mentioned all age groups, that is, children, adolescents, young adults, and adults. Gender distribution for males and females has been mentioned separately in 28 (80%) studies, while 7 (20%) studies only mentioned the percentage of female participants [79,85,87,96-99]. A total of 27 studies (77%) mentioned the ethnicity of the participants.

A total of 13 (37%) studies were conducted in Europe (Sweden, Slovakia, London, the Netherlands, Norway, Turkey, Finland, Germany, Greece, Spain, and Austria); 11 (31%) studies were conducted in North America (both United States and Canada combined); 7 (20%) in Asia (Turkey, Singapore, India, China, Japan, and South Korea); 3 (9%) in Australia, and 1 (3%) in Africa (Tunisia). The duration of the interventions varied from 1 week to a maximum of 12 weeks. Of all the studies, 23 (66%) studies had follow-up assessments. In particular, 19 (54.3%) studies had 1 follow-up, 2 (6%) studies had 2 follow-ups, and 2 (6%) had 3 follow-ups (Table 2). The duration of the follow-ups ranged from 2 to 12 weeks. The range of the interventions was 11 weeks (from 1 to 12 weeks), while the median duration was 12 weeks.

Data from the participants were collected through web-based surveys and questionnaires as well as in person. A total of 26 (74%) studies used convenience sampling techniques, and 7 (20%) studies used purposive sampling procedures. In 1 (3%) study, experience sampling was used, while in 1 (3%) study, the details of sampling were not clearly mentioned; 29 (83%) studies conducted the trial on the general population (nonclinical), while 6 (17%) studies conducted the trial on clinical population (ie, individuals who engaged in nonsuicidal self-injury, scored high on scales of depression, had anxiety and stress, had mental disorder or psychological stress, had autism spectrum disorder, were undergoing active cancer treatment). A total of 28 (80%) studies conducted the intervention in a web-based setting. In addition, 4 (11%) studies conducted the intervention through smartphone-based applications or SMS text messaging; 1 (3%) study used a hybrid setting, and 1 (3%) conducted the intervention in a telehealth setting [100,101]. Interestingly, 1 (3%) study conducted the intervention through Instagram, and 1 (3%) study used the Vivibot chatbot [89,102].

Table 2. Study characteristics.

Study	Study design	Duration of intervention	Follow-up and timing of follow-up	Experimental group, n	Control group, n	Characteristics of control group	Sampling
Mahalik et al [112]	RCT[a]	—[b]	—	Psychoeducation=61, psychoeducation and purpose reflection=70	52	Waitlist control	—
Krifa et al [77]	2-armed RCT, pretest and posttest	8 weeks	1 (12 weeks)	183	183	Waitlist control	Convenience sampling
Drabu et al [78]	RCT, pretest and posttest	T1-postsession one; post one- week training-T2	1 (2 weeks after completion of second session)	30	33	Digital control	Convenience and purposive sampling
Lennard et al [79]	RCT	—	1 (8 weeks)	231	239	Waitlist control	Convenience sampling
Andersson et al [80]	RCT, pretest and posttest	6 weeks	—	Compassion and mindfulness group=25 each	15	Waitlist control	Convenience sampling
Beshai et al [81]	RCT, pretest and posttest	4 weeks	—	227	229	Digital control	Convenience sampling
Chilver and Gatt [96]	RCT, pretest and posttest	6 weeks	1 (7 weeks)	205	204	Digital control	Convenience sampling
Hussong et al [104]	RCT, pretest and posttest	1 week (parents asked to complete the program within the week)	1 (4 weeks)	53	51	Waitlist control	Convenience and purposive sampling
Halamová et al [105]	RCT, pretest and posttest	2 weeks (14 days)	1 (8 weeks)	70	53	Waitlist control	Convenience sampling through snowballing technique
Kelman et al [82]	RCT, pretest and posttest	2 weeks	1 (2 weeks)	69	68	Digital control	Convenience sampling
Hamm et al [97]	RCT, pretest and posttest	4 weeks (1 month)	3 (12 weeks)	628	628	Waitlist control	Convenience sampling
Daugherty et al [98]	Quasi-experimental, RCT, pretest and posttest	1 month (28 days)	—	66	46	Digital control	Convenience sampling
Halamová et al [108]	RCT, pretest and posttest	2 weeks (14 days)	1 (8 weeks)	69	53	Waitlist control	Convenience sampling
Kappen et al [83]	RCT, pretest and posttest	2 weeks (12 days)	—	56	57	Digital control	Convenience sampling
Galante et al [99]	RCT, pretest and posttest	4 weeks	—	409	400	Digital control	Convenience sampling
Halamová et al [109]	RCT, pretest and posttest	2 weeks (15 days)	1 (8 weeks)	93	53	Waitlist control	Convenience sampling
Drozd et al [110]	RCT	4 weeks	3 (4 weeks, 8 weeks, and 24 weeks)	112	94	Waitlist control	Convenience sampling
Koydemir and Sun-Selışık [84]	RCT	8 weeks	—	48	44	Waitlist control	Convenience sampling
Sergeant and Mongrain [85]	RCT	3 weeks	2 (4 weeks and 8 weeks)	253	213	Digital control	Convenience sampling

Study	Study design	Duration of intervention	Follow-up and timing of follow-up	Experimental group, n	Control group, n	Characteristics of control group	Sampling
Lappalainen et al [86]	RCT, pretest and posttest	5 weeks	—	iACT[c] student coach+virtual coach group =116 and iACT virtual coach group=116	116	Waitlist control	Convenience sampling
Tay [106]	RCT, pretest and posttest	2 weeks	1 (8 weeks)	97	78	Digital control	Convenience sampling and purposive sampling
Paetzold et al [100]	RCT, pretest and posttest	6 weeks	1 (4 weeks)	46	46	Waitlist control	Experience sampling
Qu et al [101]	RCT	12 weeks	—	Program evaluation=56.25%; focus group interview=70.8%	Program evaluation=43.75%; focus group intervein=29.2%	Digital control	Convenience sampling and snowball sampling
Webb et al [87]	RCT, pretest and posttest	4 weeks	1	159	129	Waitlist control	Purposive sampling
Nawa and Yamagishi [103]	RCT, pretest and posttest	2 weeks	2 (4 weeks and 12 weeks)	42	42	Digital control	Convenience sampling
Brouzos et al [88]	Quasi-experimental, pretest and posttest	2 weeks	1 (2 weeks)	44	38	Not explicit	Convenience sampling
Pizarro-Ruiz et al [107]	RCT, pretest and posttest	2 weeks=14 days	—	89	75	Digital control	Convenience sampling
Halamová et al [111]	RCT, pretest and posttest	13 days=2 weeks	1 (8 weeks)	91	53	Waitlist control	Convenience sampling
Sampson et al [102]	RCT, pretest and posttest	Recruitment: 5 months=20 weeks	—	71	61	Digital control	Convenience sampling
Greer et al [89]	RCT	4 weeks=28 days	1 (8 weeks)	25	20	Waitlist control	Convenience and snowball sampling
Tagalidou et al [90]	RCT, pretest and posttest	1 week	1 (4 weeks)	Coping humor=35, three funny things=46, three good things=52	Early memories=49	Waitlist control	Convenience sampling
Bronk et al [91]	RCT, pretest and posttest, and lagged posttest	3 days	1 (1 week)	Gratitude condition=74; purpose condition=79	71 in the control condition	Waitlist control	Convenience sampling
Gu et al [92]	RCT, pretest and posttest	4 weeks	1 (2 weeks)	CFI[d]=10 and REBT[e]=10	12	Waitlist control	Convenience sampling
Alexiou et al [93]	RCT, pretest and posttest	3 weeks	1 (4 weeks)	19	19	Digital control	Purposive and convenience sampling
Manicavasagar et al [94]	RCT, pretest and posttest	6 weeks	—	120	115	Digital control	Convenience sampling

[a]RCT: randomized controlled trial.

[b]Not applicable.

[c]iACT: internet-based acceptance and commitment therapy.

[d]CFI: compassion-focused therapy-based intervention.

[e]REBT: rational emotive behavior therapy.

Theoretical Background

The broaden-and-build theory was the widely used theory in 5 (14%) studies [84,89,91,99,103]. The broaden-and-build theory of positive emotion states that certain discrete positive emotions, for example, joy, contentment, pride, and love, share the ability to broaden momentary thought-action repertoires of people and build their enduring personal resources, ranging from physical and intellectual resources to social and psychological resources [113]. Other studies frequently mentioned in the articles included acceptance and commitment theory [79] (oriented toward the development of psychological flexibility), affect theory [80] (emotions generating weak or strong ties to relations), attachment theory [80,92,100] (individuals born with innate behaviors function to attract proximity to attachment figures), Eisenberg's theory of parent emotion socialization [104] (parents' emotion socialization behaviors driving children's emotion socialization), emotion-focused therapy theory [105] (integrating person-focused care with modern emotion theory), motivational theory of life span development (Heckausen's theory) [97] (role of individual in lifespan development), Bandura's self-efficacy theory [106] (belief that one can execute needed steps to achieve a goal), social mentality theory [92,100] (both care-seeking and caregiving mentalities are activated when one is being self-compassionate and reassuring), embodiment theory [87] (psychological processes influenced by body), self-determination theory [103] (internalizing regulation and self-regulation), theory of mindfulness [107] (being actively engaged is beneficial for a rigid and judgmental mindset), Festinger's social comparison theory [102] (based on social comparison), stress and coping theory [89] (coping with stressful situations), and humor theory [90] (cognitive view of humor).

Quality Assessment

Among all 35 articles, the summary score of the quality assessment ranged from 12 [78] to 29 [77,102], with a median of 22.5 points. Among all criteria, most of the papers (31/35, 89%) did not report all important harms or unintended effects in each group and protocol of the full trial. Blinding was either not performed or not explicitly reported by 86% (n=30) of the studies. Also, most papers (26/35, 74%) lacked more detailed information about any changes to trial outcomes after the trial started with reasons, and 24 (69%) studies lacked information regarding the mechanism used to implement the random allocation sequence (such as sequentially numbered containers) and the description of any steps taken to conceal the sequence until interventions were assigned; 23 (66%) studies did not report explanation of any interim analyses and stopping guidelines, and 22 (63%) studies did not mention essential changes to methods after trial commencement (such as eligibility criteria) with reasons. Finally, 21 (60%) studies did not calculate both absolute and relative effect sizes for binary outcomes. A detailed description of each study evaluation is reported in Multimedia Appendix 2 [77-94,96-112].

PPIs Used in the Studies

A diverse range of PPIs were conducted among the study participants and is reported in detail in Multimedia Appendix 3 [77-94,96-112].

Web-Based Meditation and Mindfulness

Drabu et al [78] used web-based, self-compassion–based, guided meditation for nonsuicidal self-injury. Krifa et al [77] experimented with a web-based multicomponent intervention that included lectures, expert videos, psychoeducation, and positive psychology practices to assist Tunisian students with mental health. Kelman et al [82] compared web-based compassion mind training and CBT for perinatal and pregnant women. Beshai et al [81] used web-based psychoeducational videos, guided meditations, and exercises to reduce anxiety and depression. Halamová et al [105,108,109] studied self-compassion and self-criticism through various web-based and smartphone-assisted exercises. Pizarro-Ruiz et al [107] conducted guided mindfulness sessions via a smartphone app (Aire Fresco). Gu et al [92] experimented on Chinese students to mitigate depression and anxiety through web-based individual counseling sessions regarding mindfulness meditation (MP3 files).

Positive Psychology and Self-Compassion

Lennard et al [79] provided web-based self-compassion training for mothers. Drozd et al [110] experimented with an internet-based program, "Better Days," which included psychoeducational exercises to increase happiness. Halamová et al [111] explored emotion-focused training and loving-kindness meditation. Galante et al [99] practiced loving-kindness meditation through web-based videos. Webb et al [87] used a web-based yoga program to improve body image satisfaction and self-compassion. Nawa and Yamagishi [103] assessed academic motivation using journal writing and web-based self-assessments, including gratitude and other daily life aspects. Brouzos et al [88] tested the "Staying Home—Feeling Positive" web-based PPI during the COVID-19 pandemic. Andersson et al [80] provided compassion mindset intervention training via smartphone app among university students. Alexiou et al [93] assessed burnout and depression among Greek health care professionals by conducting a PPI. Manicavasagar et al [94] used "Bite Back," a multicomponent web-based positive psychology to increase well-being among young adults.

Gratitude and Acts of Kindness

Chilver and Gatt [96] explored self-compassion and acts of kindness through web-based modules. Hussong et al [104] examined parent-child gratitude conversations using web-based tools. Tagalidou et al [90] used web-based humorous diary writing techniques to address happiness and depression.

Optimism and Positive Emotion

Sergeant and Mongrain [85] analyzed optimism among participants through web-based diary writing exercises. Tay [106] assessed a web-based Hope, Optimism, and Positive Emotion intervention. Sampson et al [102] used Instagram images to assess body, facial, and smile dissatisfaction. Koydemir and Sun-Selışık [84] analyzed optimism among participants by using alternating web-based diary writing exercises. Hamm et al [97] focused on improving goal engagement and optimism among university students.

Relationship Satisfaction and Acceptance

Kappen et al [83] assessed relationship satisfaction and partner acceptance through web-based psychoeducation. Qu et al [101] analyzed sensory social routines, attention, dyadic engagement, and nonverbal communication in children with autism using synchronous group-based parent coaching sessions via telehealth.

Purpose and Well-Being

Mahalik et al [112] assessed the Father Project webpage's intervention for fathers' sense of purpose. Paetzold et al [100] aimed to enhance resilience through web-based ecologic momentary interventions and face-to-face sessions. Lappalainen et al [86] used ACT intervention to increase self-compassion skills and psychological flexibility during the COVID-19pandemic. Bronk et al [91] conducted the Purpose Toolkit and Gratitude Toolkit to increase a sense of purpose among participants. Greer et al [89] studied psychological well-being among patients with cancer using Vivibot, a chatbot designed to deliver positive psychological skills. Daugherty et al [98] used a smartphone app for hope and well-being.

Digital Control Versus Waitlist Control

The control groups were categorized into 2 groups: digital controls (15/35, 43%) and waitlist controls (19/35, 54%; Table 2). Digital controls involved some form of digital or web-based interaction but did not include the full intervention content. They had access to certain activities or features but did not receive the complete intervention that the experimental group received. For example, digital control groups included audio recording; video watching; internet-based communication [78,81,82]; psychoeducation; web-based daily registry of relationship experiences; web-based diary writing activities without positive psychology components; and digital placebo activities (writing daily events, early memories, and life events) [83,85,93,94]. Other control measures included elements of positive psychology that differed from the focus of the intervention (the digital control group had identical initial app assessment as the intervention group but did not receive the complete treatment or they had access to a website featuring inspirational phrases) [96,98,99,106]. In other cases, the control group even performed daily self-evaluations without an equivalent active task, downloaded the Lumosity smartphone app, or used neutral Instagram images of nature [101-103,107].

The waitlist controls referred to the control groups in which participants did not receive the active intervention during the initial phase of the study but were promised or scheduled to receive it at a later time. The waitlist control group participants do not receive the full intervention immediately and instead are placed on "waitlist" to receive the intervention after a specified waiting period. RCT control groups either received no treatment or were given access to full digital intervention content after the trial. Among the 35 studies, 1 (3%) study did not explicitly mention the category of the control group [88].

Outcomes of PPIs

Outcomes were related to both ill-being and well-being components (Table 1). In particular, 27% (15/18) of studies focused on ill-being, among others prominently including depression (11/15, 73%), anxiety (9/15, 60%), stress (8/15, 53%), and loneliness (3/15, 20%) using measures such as the Depression, Anxiety, and Stress Scales, the Generalized Anxiety Disorder scale, the short and long forms of the Spielberger State-Trait Anxiety Inventory, and De Jong Gierveld Loneliness Scale.

Well-being outcomes included compassion (6/18, 33%), satisfaction (7/18, 39%), optimism (1/18, 6%), happiness (3/18, 17%), resilience (1/18, 6%), emotion regulation and emotion awareness (6/18, 33%), hope (3/18, 17%), mindfulness (2/18, 11%), purpose (1/18, 6%), quality of life (1/18, 6%), gratitude (1/18, 6%), empathy (1/18, 6%), forgiveness (1/18, 6%), motivation (1/18, 6%), and kindness (1/18, 6%) using the Self‐Compassion Scale (Self‐Compassion Scale‐Short Form), Satisfaction with Life Scale, Life Orientation Test-Revised, Authentic Happiness Inventory, Connor-Davidson Resilience Scale, Profile of Emotional Competence, Snyder Hope scale, The Five Facet Mindfulness Questionnaire-15, Claremont Purpose Scale, "Psychological Health" and "Social Relationships" subscales of WHO Quality of Life-Brief Version, Gratitude Questionnaire, Interpersonal Reactivity Index, Heartland Forgiveness Scale, and Chinese Compassion Scale.

Both ill-being and well-being included components of self-criticism and self-reassurance (6/18, 33%), well-being (both positive and negative; 5/18, 28%), positive and negative effect (4/18, 22%) using Forms of Self-Criticism and Reassuring Scale, Warwick-Edinburgh Mental Well-being Scale and Positive and Negative Affect Scale.

Meta-Analytic Results

Meta-analyses on well-being outcomes showed that PPIs improved purpose, gratitude, and hope with a medium-to-large effect size (k=12; Hedges g=0.555, 95% CI 0.348-0.761; P<.001; I^2=70%). Only 1 (%) study involved a digital control group, for which the reported effect was significantly smaller (Hedges g=0.09). In addition, PPIs augmented the levels of compassion (k=13; Hedges g=0.447, 95% CI 0.210-0.684; P=.001; I^2=62%), with no significant differences (P=.34) between the waiting list (k=11; Hedges g=0.356) and the digital control group (k=2; g=0.670). In addition, PPIs augmented the positive coping behaviors (k=6; Hedges g=0.421; 95% CI 0.072-0.770; P=.003; I^2=72%) with a medium effect size. PPI interventions also improved body image–related outcomes with a medium effect (k=7; Hedges g=0.238, 95% CI 0.090-0.388; P=.007; I^2=0%). A small-to-medium effect was found for mindset predisposition (k=13; Hedges g=0.304, 95% CI 0.072-0.537; P=.02; I^2=74%), with a significant difference between the control groups (P=.01). In particular, effect size was larger and significant when a waiting list (k=6; Hedges g=0.534) was included as the control group when compared with the digital controls (k=7; Hedges g=0.092). Also, a small-to-medium effect was also found for the variable 3 funny things/3 good things (k=8; Hedges g=–0.206, 95% CI –0.328 to –0.083; P=.005; I^2=0%) with all studies including the waiting list control groups. A nonsignificant effect was found for cognitive flexibility (k=7; Hedges g=0.054, 95% CI –0.265 to 0.372; P=.69; I^2=75%) and mood/affect/emotions (k=8; Hedges *g*=0.364, 95% CI –0.120 to 0.849; *P*=.12; I^2=81%), the latter with no difference in the control groups (*P*=.40) although participants in the waiting list showed a larger effect size (*k*=5; Hedges *g*=0.570) when compared with the digital control (*k*=3; Hedges *g*=0.088). Also, satisfaction and quality of life (*k*=7; Hedges *g*=0.338, 95% CI –0.119 to 0.793; *P*=.12; I^2=81%) showed a nonsignificant effect, with no differences between subgroups (*P*=.65; Figure 2).

Ill-being outcomes were less represented in the included studies (Figure 3). Meta-analyses showed a large negative effect for the reduction of cognitive biases (*k*=14; Hedges *g*=–0.637, 95% CI 1.309 to –0.036; *P*=.05; I^2=94%), with no group differences between the waiting list and the digital control group (*P*=.54), although once again effect sizes tended to be larger when a waiting list (*k*=7; Hedges *g*=–0.799) was considered with respect to digital control groups (*k*=5; Hedges *g*=–0.405). PPIs showed a medium-to-large effect on the reduction of negative emotions and mood problems (*k*=30; Hedges *g*=–0.369, 95% CI –0.513 to –0.225; *P*<.001; I^2=60%). Interestingly, subgroup differences showed that the effect size was significantly (*P*=.03) larger for studies including a waiting list (*k*=21; Hedges *g*=–0.456) when compared with studies including digital control groups (*k*=9; Hedges *g*=–0.200). PPIs also diminished stress levels (*k*=8; Hedges *g*=–0.342, 95% CI –0.677 to –0.007; *P*=.045; I^2=81%), with no significant differences in the effect size (*P*=.35) between studies including a waiting list (*k*=5; Hedges *g*=–0.441) versus digital control groups (*k*=3; Hedges *g*=–0.157). A very large effect was found for coping (*k*=3; Hedges *g*=–0.939,95% CI –1.151 to –0.728; *P*=.003; I^2=0%); however, interpretation of this result would be limited due to the low number of studies. While the effect sizes were not significant for body image–related outcomes (*k*=4; Hedges *g*=–0.305, 95% CI –0.851 to –0.240; *P*=.17; I^2=17%) and 3 funny things/3 good things (*k*=6; Hedges *g*=–0.048, 95% CI –0.289 to –0.192; *P*=.63; I^2=28.5%).

Funnel plots were symmetrical for the overall meta-analysis of both the well-being and ill-being outcomes (Figures 4 and 5), and the regression test for funnel plot asymmetry was not significant in both cases as well, thus confirming the absence of publication biases. Influence analyses revealed that that a single study did not account for a significant part of the variance in the final effect.

Finally, meta-regression analyses showed that PPIs tended to show a larger effect size on well-being outcomes in studies including young adults (β=.322; *P*=.008), while no specific effect was found for ill-being outcomes. Figures S1 and S2 in Multimedia Appendices 4 and 5 report additional detailed information of each meta-analysis.

Figure 2. Meta-analyses on well-being outcomes for positive psychology interventions (PPIs).

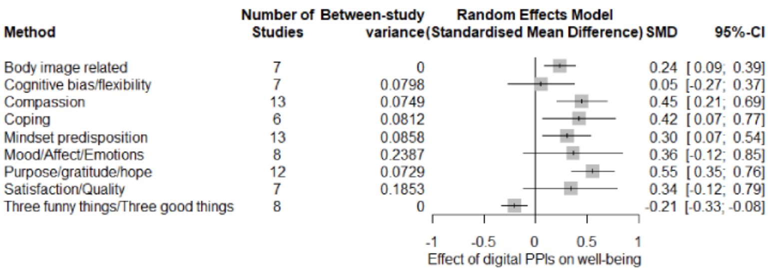


Figure 3. Meta-analyses on ill-being outcomes for positive psychology interventions (PPIs).

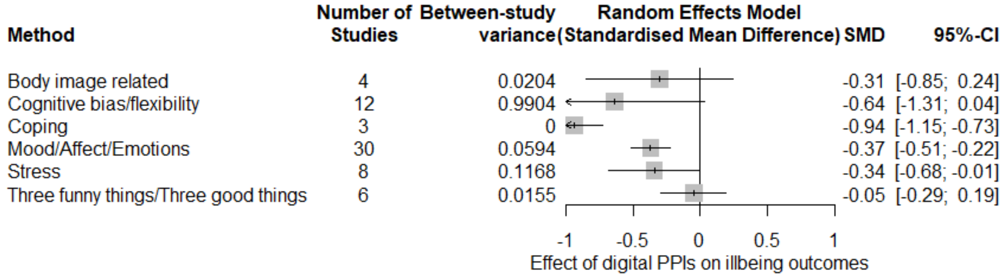


Figure 4. Funnel plot of well-being outcomes for positive psychology interventions.

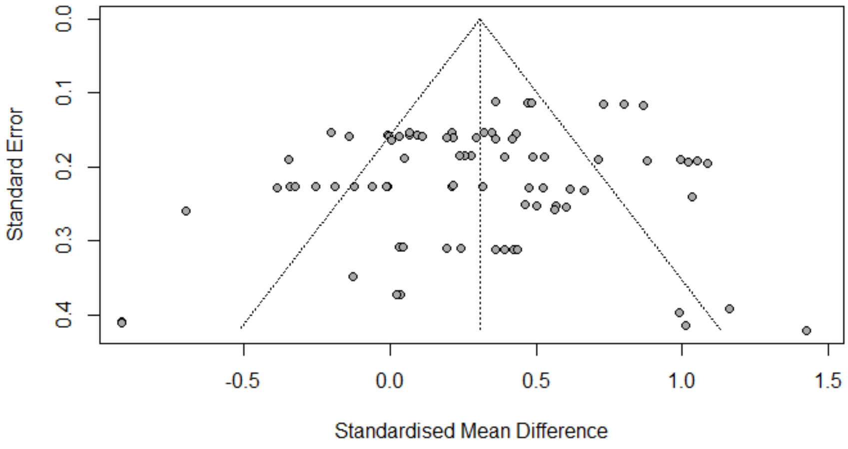
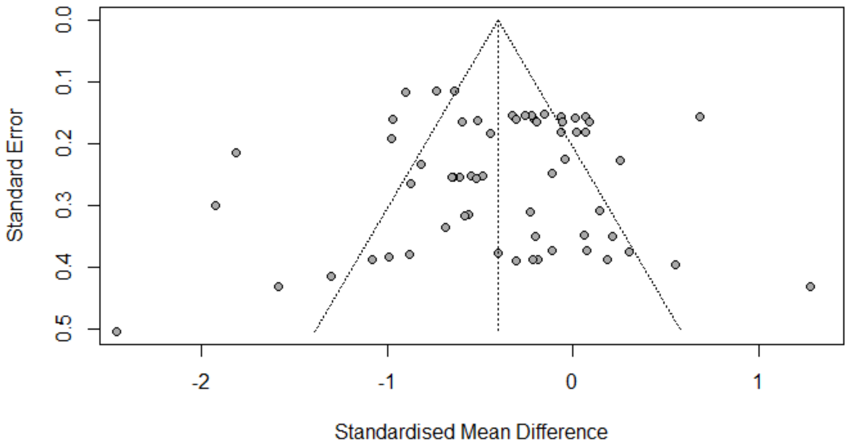


Figure 5. Funnel plot of ill-being outcomes for positive psychology interventions.



Discussion

Principal Findings

In our systematic review and meta-analysis of 35 studies and 18 studies, respectively, we examined the impact of digital interventions grounded in positive psychology on the well-being and ill-being of children, adolescents, and young adults. Our results showed 4 main findings. First, when it comes to well-being outcomes, PPIs enhanced various facets of well-being, notably purpose, gratitude, and hope, with a medium-to-large effect, as well as compassion, positive coping (eg, coping with humor), and body image–related concerns with a medium effect. Smaller effects were found for mindset predisposition and 3 funny things/3 good things, while PPIs did not seem to improve mood and positive emotions, satisfaction and quality of life, and cognitive flexibility. This aligns with prior investigations of PPIs in traditional settings [17]. These interventions seem to provide robust support in enhancing aspects of well-being that involve an individual's outlook on life and their ability to foster a sense of personal achievement and satisfaction.

Second, when we looked at ill-being outcomes, the picture was different. In particular, the larger effect was found for diminishing cognitive biases, including self-criticism and fears, followed by a decrement in negative emotions and mood problems, especially when participants of the experimental group were compared with the waiting list. Hence, PPIs can be a useful tool in reducing cognitive biases typical of, for example, mood problems, and stress levels [83]. To note, it is crucial to differentiate the control groups in the analyses. Indeed, although we could not make subgroup comparisons for all the outcomes due to the paucity of studies in each group, we showed that effects sizes tended to be consistently larger in studies including a waiting list rather than a digital control group (eg, including some sort of web-based interactions). Digital control groups, such as those engaging in nonspecific digital activities or using general health apps, could serve as valuable benchmarks. This would allow us to distinguish the specific contributions of PPIs from broader digital engagement effects. Such comparisons would shed light on the specific psychological mechanisms activated by PPIs compared with general digital exposure, helping to isolate the unique elements of PPIs that contribute to improved well-being outcomes.

Third, several studies within our review highlighted the efficacy of interventions tailored to specific settings and contexts. For example, digital interventions aimed at fostering hope and optimism were found to be particularly beneficial for college students prone to failure and those with low optimism levels [63]. Also, interventions focusing on self-compassion were found to be especially beneficial for mothers of infants, offering them a respite from the unique challenges of early parenthood [61]. By contrast, interventions that used smartphone delivery, such as the hope intervention, showcased the adaptability and accessibility of digital platforms, making well-being practices more integrated into daily routines [64]. Another noteworthy finding was the positive impact of multicomponent PPIs delivered web-based for subjective well-being of young adults [62].

Fourth, when age was considered as a moderator, studies with young adult participants showed larger effect sizes in the meta-analysis with well-being outcomes, but no differences emerged with respect to ill-being indicators. This is an important consideration since young adults might be more inclined to understand the importance of promoting well-being and thus more willing to take part in interventions and experiencing the positive effects. However, while some demographic groups appeared to benefit more from certain types of interventions, the overall evidence was not strong enough to conclusively determine that these effects were consistently replicated across different age groups, such as children, adolescents, or young adults [17,31].

Finally, the variability in intervention efficacy highlights the critical role of intervention design and implementation in achieving desired outcomes. This variation underscores the need for carefully tailored interventions that consider the unique needs and circumstances of the target demographic to optimize efficacy. Therefore, while PPIs hold promise, the evidence suggests that a nuanced approach to their application is necessary, where factors such as intervention type, target population, and desired well-being outcome are all carefully considered to maximize benefits [17,31].

When compared with the findings of existing literature, our findings provide a nuanced view that aligns with some previous studies but also highlights the complexity of applying PPIs across diverse populations and settings [6,114,115]. Unlike some optimistic narratives, our results suggest that while PPIs can be beneficial, their efficacy is not universal and depends on specific intervention types and target populations. Furthermore, our findings diverge from studies such as the MYRIAD trial [116], underscoring the need for cautious interpretation of PPI efficacy and the potential for adverse effects.

Future Directions

The findings from this systematic review and meta-analysis provide a solid foundation for understanding the effectiveness of PPIs in young populations. However, as with all research, there are avenues that remain unexplored and warrant further investigation. One primary recommendation is to conduct more rigorous RCTs with larger and more diverse samples. This would not only enhance the generalizability of the findings but also allow for a more in-depth exploration of the nuances and specific components of the interventions that are most effective. Also, we suggest that PPIs should be integrated in interventions that also collect biological information to further assess their efficacy.

Another crucial area for future research is the examination of the long-term effects of these digital interventions. While our review captured the immediate and short-term benefits, understanding the sustainability of these positive outcomes over extended periods is essential. This would provide insights into whether these interventions lead to lasting changes in well-being and mental health or whether periodic "booster" sessions are required to maintain the benefits. In addition, given the rapid advancements in technology, exploring the integration of emerging technologies, such as virtual reality or augmented reality, into PPIs could offer innovative ways to engage and support adolescents.

From a practical implementation perspective, stakeholders in the field of youth mental health should consider incorporating evidence-based digital interventions into broader mental health programs and curricula. Schools, community centers, and mental health organizations can benefit from these scalable and accessible tools, especially in regions where traditional face-to-face interventions might be limited. Collaborations between researchers, technologists, and educators can further refine and optimize these interventions, ensuring that they remain relevant and effective in the ever-evolving digital landscape.

Limitations

Our study, while extensive, exhibits limitations that warrant attention for a comprehensive understanding of the scope and applicability of our findings. One major limitation is the heterogeneity in study settings and targeted age groups, which ranged from school environments to clinical settings and included diverse demographic categories from children to young adults [30]. This wide variability complicates the task of uniformly generalizing the results across different settings and age demographics. In addition, the medium-to-low quality of some included studies potentially undermines the reliability and robustness of our findings. The varying methodological rigor and potential biases in the study design across the analyzed studies necessitate a cautious interpretation of the effectiveness and applicability of PPIs based on this evidence base. We did not divide the results or their interpretation between interventions aiming at preventing versus treating mental health problems since our aim was to explore the literature and treatment effects of PPIs in general; however, we do acknowledge that the absence of improvement in a prevention intervention may not be the evidence that an intervention is ineffective; hence we should consider this interpretation to avoid biasing the findings of the meta-analysis. Hence, we suggest that future studies should look more carefully at this differentiation.

In addition, an important aspect that was not covered in our review is the assessment of the safety and potential adverse effects of PPIs. Not including an evaluation of harms, as highlighted by the findings from larger trials such as the MYRIAD trial, which documented no significant effects and even potential harm in certain subgroups, poses a noteworthy gap in our analysis [117]. This aspect highlights an area for further investigation, particularly considering the intricate nature of mental health interventions and their varied effects across diverse individuals. In addition, the absence of data from lower-middle-income countries and the lack of studies not published in languages other than English limit the generalizability of our conclusions globally, raising concerns about the effectiveness and safety of PPIs in these regions where cultural, economic, and health care contexts may differ significantly from those in high-income countries [30,31]. Finally, although we calculated the intercoder reliability for the screening process, we were not able to provide a measure of reliability for the quality assessment of the studies; hence, we encourage future studies to consider conducting the assessment blind and calculate a measure of agreement.

Conclusions

In conclusion, our systematic review suggests that while PPIs can enhance certain aspects of well-being among children, adolescents, and young adults, the effects are not consistent across all domains or demographic groups. The evidence supports the effectiveness of specific types of PPIs, particularly those that enhance gratitude, purpose, and hope. However, these benefits are not uniform, and the impact varies by the type of well-being outcome and the population segment. Moreover, given the significant variability in the intervention settings, the diversity of outcomes, and the medium-to-low quality of the studies reviewed, any conclusions about the efficacy of PPIs should be viewed as tentative. The findings underscore the necessity for further rigorous research to better understand the mechanisms and effectiveness of PPIs, assess their safety, and evaluate their applicability in different geographical and clinical contexts. Future studies should also explore how digital platforms might uniquely influence the success of these interventions and consider the theoretical underpinnings of PPIs in more depth to enhance their practical and academic contributions.

Acknowledgments

This study was funded by Swiss National Science Foundation (grant P500PS_202974) and supported by the National Institute of Child Health and Human Development NICHD (grant 1R21HD115354-01).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Basic characteristics.
[DOCX File , 30 KB - mental_v11i1e56045_app1.docx]

Multimedia Appendix 2
Quality assessment.
[DOCX File , 19 KB - mental_v11i1e56045_app2.docx]

Multimedia Appendix 3
Characteristics of the positive psychology intervention.
[DOCX File , 33 KB - mental_v11i1e56045_app3.docx]

Multimedia Appendix 4
Additional detailed information of each meta-analysis (Figure S1).
[PNG File , 126 KB - mental_v11i1e56045_app4.png]

Multimedia Appendix 5
Additional detailed information of each meta-analysis (Figure S2).
[PNG File , 177 KB - mental_v11i1e56045_app5.png]

References

1. Barker MM, Beresford B, Bland M, Fraser LK. Prevalence and incidence of anxiety and depression among children, adolescents, and young adults with life-limiting conditions: a systematic review and meta-analysis. JAMA Pediatr 2019 Sep 01;173(9):835-844 [FREE Full text] [doi: 10.1001/jamapediatrics.2019.1712] [Medline: 31282938]
2. World mental health report: transforming mental health for all. World Health Organization. 2022 Jun 16. URL: https://www.who.int/publications/i/item/9789240049338 [accessed 2023-07-02]
3. Mental health of adolescents. World Health Organization. 2021 Nov 17. URL: https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health [accessed 2023-08-08]
4. Surgeon general issues new advisory about effects social media use has on youth mental health. U.S. Department of Health and Human Services. 2023 May 23. URL: https://www.hhs.gov/about/news/2023/05/23/surgeon-general-issues-new-advisory-about-effects-social-media-use-has-youth-mental-health.html [accessed 2023-07-02]

5. Twenge JM, Haidt J, Blake AB, McAllister C, Lemon H, Le Roy A. Worldwide increases in adolescent loneliness. J Adolesc 2021 Dec;93(1):257-269 [FREE Full text] [doi: 10.1016/j.adolescence.2021.06.006] [Medline: 34294429]
6. Marciano L, Ostroumova M, Schulz PJ, Camerini AL. Digital media use and adolescents' mental health during the Covid-19 pandemic: a systematic review and meta-analysis. Front Public Health 2021;9:793868 [FREE Full text] [doi: 10.3389/fpubh.2021.793868] [Medline: 35186872]
7. Blum RW, Lai J, Martinez M, Jessee C. Adolescent connectedness: cornerstone for health and wellbeing. BMJ 2022 Oct 27;379:e069213 [FREE Full text] [doi: 10.1136/bmj-2021-069213] [Medline: 36302526]
8. Eisenstadt M, Liverpool S, Infanti E, Ciuvat RM, Carlsson C. Mobile apps that promote emotion regulation, positive mental health, and well-being in the general population: systematic review and meta-analysis. JMIR Ment Health 2021 Nov 08;8(11):e31170 [FREE Full text] [doi: 10.2196/31170] [Medline: 34747713]
9. Thieme A, Wallace J, Meyer TD, Olivier P. Designing for mental wellbeing: towards a more holistic approach in the treatment and prevention of mental illness. In: Proceedings of the 2015 British HCI Conference. 2015 Presented at: British HCI '15; July 13-17, 2015; Lincoln, UK. [doi: 10.1145/2783446.2783586]
10. Kobau R, Seligman ME, Peterson C, Diener E, Zack MM, Chapman D, et al. Mental health promotion in public health: perspectives and strategies from positive psychology. Am J Public Health 2011 Aug;101(8):e1-e9. [doi: 10.2105/ajph.2010.300083]
11. Tomé G, Almeida A, Ramiro L, Gaspar T, Gaspar de Matos M. Intervention in schools promoting mental health and wellbeing: a systematic review. Global J Community Psychol Pract 2021;12(1):1-23 [FREE Full text] [doi: 10.1093/med/9780198792994.003.0042]
12. Geerling B, Kraiss JT, Kelders SM, Stevens AW, Kupka RW, Bohlmeijer ET. The effect of positive psychology interventions on well-being and psychopathology in patients with severe mental illness: a systematic review and meta-analysis. J Posit Psychol 2020 Jul 10;15(5):572-587. [doi: 10.1080/17439760.2020.1789695]
13. Williams E, Dingle GA, Clift S. A systematic review of mental health and wellbeing outcomes of group singing for adults with a mental health condition. Eur J Public Health 2018 Dec 01;28(6):1035-1042. [doi: 10.1093/eurpub/cky115] [Medline: 29982515]
14. APA dictionary of psychology. American Psychological Association. URL: https://dictionary.apa.org/ [accessed 2023-06-20]
15. Ciarrochi J, Atkins PW, Hayes LL, Sahdra BK, Parker P. Contextual positive psychology: policy recommendations for implementing positive psychology into schools. Front Psychol 2016 Oct 10;7:1561 [FREE Full text] [doi: 10.3389/fpsyg.2016.01561] [Medline: 27777564]
16. Rashid T. Positive psychotherapy. In: Maggino F, editor. Encyclopedia of Quality of Life and Well-Being Research. Cham, Switzerland: Springer; 2020.
17. Arslan G, Yıldırım M, Zangeneh M, Ak İ. Benefits of positive psychology-based story reading on adolescent mental health and well-being. Child Indic Res 2022 Jan 06;15(3):781-793 [FREE Full text] [doi: 10.1007/s12187-021-09891-4] [Medline: 35013685]
18. Di Fabio A, Palazzeschi L. Hedonic and eudaimonic well-being: the role of resilience beyond fluid intelligence and personality traits. Front Psychol 2015 Sep 11;6:1367 [FREE Full text] [doi: 10.3389/fpsyg.2015.01367] [Medline: 26441743]
19. Park N, Peterson C, Szvarca D, Vander Molen RJ, Kim ES, Collon K. Positive psychology and physical health: research and applications. Am J Lifestyle Med 2016 Sep 26;10(3):200-206 [FREE Full text] [doi: 10.1177/1559827614550277] [Medline: 30202275]
20. Scorsolini-Comin F, Fontaine AM, Koller SH, Antônio dos Santos M. From authentic happiness to well-being: the flourishing of positive psychology. Psicol Reflex Crit 2012;26(4):663-670. [doi: 10.1590/s0102-79722013000400006]
21. Seligman ME. Positive psychology: a personal history. Annu Rev Clin Psychol 2019 May 07;15(1):1-23. [doi: 10.1146/annurev-clinpsy-050718-095653] [Medline: 30525996]
22. Goh PS, Goh YW, Jeevanandam L, Nyolczas Z, Kun A, Watanabe Y, et al. Be happy to be successful: a mediational model of PERMA variables. Asia Pac J Human Res 2021 Feb 12;60(3):632-657. [doi: 10.1111/1744-7941.12283]
23. Leontopoulou S. Measuring well-being in emerging adults: exploring the PERMA framework for positive youth development. Psychology 2020 Nov 22;25(1):72. [doi: 10.12681/psy_hps.25337]
24. Kelders SM, van Zyl LE, Ludden GD. The concept and components of engagement in different domains applied to eHealth: a systematic scoping review. Front Psychol 2020 May 27;11:926 [FREE Full text] [doi: 10.3389/fpsyg.2020.00926] [Medline: 32536888]
25. Kun Á, Balogh P, Krasz KG. Development of the work-related well-being questionnaire based on Seligman's PERMA model. Period Polytech Soc Manag Sci 2016 Dec 05;25(1):56-63. [doi: 10.3311/PPso.9326]
26. Martela F, Steger MF. The three meanings of meaning in life: distinguishing coherence, purpose, and significance. J Posit Psychol 2016 Jan 27;11(5):531-545. [doi: 10.1080/17439760.2015.1137623]
27. Hidayat R, Moosavi Z, Hermandra, Zulhafizh Z, Hadisaputra P. Achievement goals, well-being and lifelong learning: a mediational analysis. Int J Instruct 2022;15(1):89-112 [FREE Full text] [doi: 10.29333/iji.2022.1516a]
28. Hendriks T, Warren MA, Schotanus-Dijkstra M, Hassankhan A, Graafsma T, Bohlmeijer E, et al. How WEIRD are positive psychology interventions? A bibliometric analysis of randomized controlled trials on the science of well-being. J Posit Psychol 2018 Aug 29;14(4):489-501. [doi: 10.1080/17439760.2018.1484941]

29. Owens RL, Waters L. What does positive psychology tell us about early intervention and prevention with children and adolescents? A review of positive psychological interventions with young people. J Posit Psychol 2020 Jul 07;15(5):588-597. [doi: 10.1080/17439760.2020.1789706]
30. Cheung EO, Kwok I, Ludwig AB, Burton W, Wang X, Basti N, et al. Development of a positive psychology program (LAVENDER) for preserving medical student well-being: a single-arm pilot study. Glob Adv Health Med 2021 Jan 28;10:2164956120988481 [FREE Full text] [doi: 10.1177/2164956120988481] [Medline: 33614252]
31. Antoine P, Dauvier B, Andreotti E, Congard A. Individual differences in the effects of a positive psychology intervention: applied psychology. Pers Individ Differ 2018 Feb;122:140-147. [doi: 10.1016/j.paid.2017.10.024]
32. Kor A, Pirutinsky S, Mikulincer M, Shoshani A, Miller L. A longitudinal study of spirituality, character strengths, subjective well-being, and prosociality in middle school adolescents. Front Psychol 2019 Feb 27;10:377 [FREE Full text] [doi: 10.3389/fpsyg.2019.00377] [Medline: 30873082]
33. Kaplan S, Bradley-Geist JC, Ahmad A, Anderson A, Hargrove AK, Lindsey A. A test of two positive psychology interventions to increase employee well-being. J Bus Psychol 2013 Aug 9;29(3):367-380. [doi: 10.1007/s10869-013-9319-4]
34. Durgante HB, Dell'Aglio DD. Adaptation for online implementation of a positive psychology intervention for health promotion. Ciencias Psicológicas 2022;16(2):1-15 [FREE Full text]
35. Briggs S, Netuveli G, Gould N, Gkaravella A, Gluckman NS, Kangogyere P, et al. The effectiveness of psychoanalytic/psychodynamic psychotherapy for reducing suicide attempts and self-harm: systematic review and meta-analysis. Br J Psychiatry 2019 Jun 28;214(6):320-328. [doi: 10.1192/bjp.2019.33] [Medline: 30816079]
36. Abbass AA, Kisely SR, Town JM, Leichsenring F, Driessen E, De Maat S, et al. Short-term psychodynamic psychotherapies for common mental disorders. Cochrane Database Syst Rev 2014 Jul 01;2014(7):CD004687 [FREE Full text] [doi: 10.1002/14651858.CD004687.pub4] [Medline: 24984083]
37. Caldwell DM, Davies SR, Hetrick SE, Palmer JC, Caro P, López-López JA, et al. School-based interventions to prevent anxiety and depression in children and young people: a systematic review and network meta-analysis. The Lancet Psychiatry 2019 Dec;6(12):1011-1020. [doi: 10.1016/s2215-0366(19)30403-1]
38. Werner-Seidler A, Spanos S, Calear AL, Perry Y, Torok M, O'Dea B, et al. School-based depression and anxiety prevention programs: an updated systematic review and meta-analysis. Clin Psychol Rev 2021 Nov;89:102079 [FREE Full text] [doi: 10.1016/j.cpr.2021.102079] [Medline: 34571372]
39. Lee A, Gage NA. Updating and expanding systematic reviews and meta‐analyses on the effects of school‐wide positive behavior interventions and supports. Psychol Sch 2020 May;57(5):783-804. [doi: 10.1002/pits.22336]
40. Versluis A, Verkuil B, Spinhoven P, van der Ploeg MM, Brosschot JF. Changing mental health and positive psychological well-being using ecological momentary interventions: a systematic review and meta-analysis. J Med Internet Res 2016 Jun 27;18(6):e152 [FREE Full text] [doi: 10.2196/jmir.5642] [Medline: 27349305]
41. Malouff JM, Schutte NS. Can psychological interventions increase optimism? A meta-analysis. J Posit Psychol 2016 Aug 16;12(6):594-604. [doi: 10.1080/17439760.2016.1221122]
42. Brown L, Ospina JP, Celano CM, Huffman JC. The effects of positive psychological interventions on medical patients' anxiety: a meta-analysis. Psychosom Med 2019 Sep;81(7):595-602. [doi: 10.1097/PSY.0000000000000722] [Medline: 31232915]
43. Firth J, Torous J, Nicholas J, Carney R, Rosenbaum S, Sarris J. Can smartphone mental health interventions reduce symptoms of anxiety? A meta-analysis of randomized controlled trials. J Affect Disord 2017 Aug 15;218:15-22 [FREE Full text] [doi: 10.1016/j.jad.2017.04.046] [Medline: 28456072]
44. Lim WL, Tierney S. The effectiveness of positive psychology interventions for promoting well-being of adults experiencing depression compared to other active psychological treatments: a systematic review and meta-analysis. J Happiness Stud 2023 Nov 05;24(1):249-273 [FREE Full text] [doi: 10.1007/s10902-022-00598-z] [Medline: 36373089]
45. Sin NL, Lyubomirsky S. Enhancing well-being and alleviating depressive symptoms with positive psychology interventions: a practice-friendly meta-analysis. J Clin Psychol 2009 May;65(5):467-487. [doi: 10.1002/jclp.20593] [Medline: 19301241]
46. White CA, Uttl B, Holder MD. Meta-analyses of positive psychology interventions: the effects are much smaller than previously reported. PLoS One 2019 May 29;14(5):e0216588 [FREE Full text] [doi: 10.1371/journal.pone.0216588] [Medline: 31141537]
47. Bolier L, Haverman M, Westerhof GJ, Riper H, Smit F, Bohlmeijer E. Positive psychology interventions: a meta-analysis of randomized controlled studies. BMC Public Health 2013 Feb 08;13:119 [FREE Full text] [doi: 10.1186/1471-2458-13-119] [Medline: 23390882]
48. Santos V, Paes F, Pereira V, Arias-Carrión O, Silva AC, Carta MG, et al. The role of positive emotion and contributions of positive psychology in depression treatment: systematic review. Clin Pract Epidemiol Ment Health 2013 Nov 28;9(1):221-237 [FREE Full text] [doi: 10.2174/1745017901309010221] [Medline: 24358052]
49. Marshall JM, Dunstan D, Bartik W. Positive psychology mobile applications for increasing happiness and wellbeing - a systematic app store review. R U appy? Eur J Appl Posit Psychol 2020 Oct 1;4:1-10 [FREE Full text]
50. Donaldson SI, Lee JY, Donaldson SI. Evaluating positive psychology interventions at work: a systematic review and meta-analysis. Int J Appl Posit Psychol 2019 Sep 10;4(3):113-134. [doi: 10.1007/s41042-019-00021-8]

51. Avey JB, Reichard RJ, Luthans F, Mhatre KH. Meta‐analysis of the impact of positive psychological capital on employee attitudes, behaviors, and performance. Human Resour Dev Q 2011;22(2):127-152. [doi: 10.1002/hrdq.20070]
52. Meyers MC, van Woerkom M, Bakker AB. The added value of the positive: a literature review of positive psychology interventions in organizations. Eur J Work Organ Psychol 2013 Oct;22(5):618-632. [doi: 10.1080/1359432x.2012.694689]
53. Tejada-Gallardo C, Blasco-Belled A, Torrelles-Nadal C, Alsinet C. Effects of school-based multicomponent positive psychology interventions on well-being and distress in adolescents: a systematic review and meta-analysis. J Youth Adolesc 2020 Oct 18;49(10):1943-1960. [doi: 10.1007/s10964-020-01289-9] [Medline: 32683592]
54. Carr A, Cullen K, Keeney C, Canning C, Mooney O, Chinseallaigh E, et al. Effectiveness of positive psychology interventions: a systematic review and meta-analysis. J Posit Psychol 2020 Sep 10;16(6):749-769. [doi: 10.1080/17439760.2020.1818807]
55. Hoppen TH, Morina N. Efficacy of positive psychotherapy in reducing negative and enhancing positive psychological outcomes: a meta-analysis of randomised controlled trials. BMJ Open 2021 Sep 06;11(9):e046017 [FREE Full text] [doi: 10.1136/bmjopen-2020-046017] [Medline: 34489269]
56. Li Pira G, Aquilini B, Davoli A, Grandi S, Ruini C. The use of virtual reality interventions to promote positive mental health: systematic literature review. JMIR Ment Health 2023 Jul 06;10:e44998 [FREE Full text] [doi: 10.2196/44998] [Medline: 37410520]
57. Groot J, MacLellan A, Butler M, Todor E, Zulfiqar M, Thackrah T, et al. The effectiveness of fully automated digital interventions in promoting mental well-being in the general population: systematic review and meta-analysis. JMIR Ment Health 2023 Oct 19;10:e44658 [FREE Full text] [doi: 10.2196/44658] [Medline: 37856172]
58. Braunwalder C, Müller R, Glisic M, Fekete C. Are positive psychology interventions efficacious in chronic pain treatment? A systematic review and meta-analysis of randomized controlled trials. Pain Med 2022 Jan 03;23(1):122-136. [doi: 10.1093/pm/pnab247] [Medline: 34347095]
59. DuBois CM, Lopez OV, Beale EE, Healy BC, Boehm JK, Huffman JC. Relationships between positive psychological constructs and health outcomes in patients with cardiovascular disease: a systematic review. Int J Cardiol 2015 Sep 15;195:265-280 [FREE Full text] [doi: 10.1016/j.ijcard.2015.05.121] [Medline: 26048390]
60. Chakhssi F, Kraiss JT, Sommers-Spijkerman M, Bohlmeijer ET. The effect of positive psychology interventions on well-being and distress in clinical samples with psychiatric or somatic disorders: a systematic review and meta-analysis. BMC Psychiatry 2018 Jun 27;18(1):211 [FREE Full text] [doi: 10.1186/s12888-018-1739-2] [Medline: 29945603]
61. Pina I, Braga CM, de Oliveira TF, de Santana CN, Marques RC, Machado L. Positive psychology interventions to improve well-being and symptoms in people on the schizophrenia spectrum: a systematic review and meta-analysis. Braz J Psychiatry 2021 Aug;43(4):430-437 [FREE Full text] [doi: 10.1590/1516-4446-2020-1164] [Medline: 33331497]
62. Mazzucchelli TG, Kane RT, Rees CS. Behavioral activation interventions for well-being: a meta-analysis. J Posit Psychol 2010 Mar;5(2):105-121 [FREE Full text] [doi: 10.1080/17439760903569154] [Medline: 20539837]
63. Salameh JP, Bossuyt PM, McGrath TA, Thombs BD, Hyde CJ, Macaskill P, et al. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. BMJ 2020 Aug 14;370:m2632 [FREE Full text] [doi: 10.1136/bmj.m2632] [Medline: 32816740]
64. Solmi M, Radua J, Olivola M, Croce E, Soardo L, Salazar de Pablo G, et al. Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. Mol Psychiatry 2022 Jan 02;27(1):281-295 [FREE Full text] [doi: 10.1038/s41380-021-01161-7] [Medline: 34079068]
65. Lijster JM, Dierckx B, Utens EM, Verhulst FC, Zieldorff C, Dieleman GC, et al. The age of onset of anxiety disorders. Can J Psychiatry 2017 Apr;62(4):237-246 [FREE Full text] [doi: 10.1177/0706743716640757] [Medline: 27310233]
66. Fusar-Poli P, Bauer M, Borgwardt S, Bechdolf A, Correll CU, Do KQ, et al. European college of neuropsychopharmacology network on the prevention of mental disorders and mental health promotion (ECNP PMD-MHP). Eur Neuropsychopharmacol 2019 Dec;29(12):1301-1311 [FREE Full text] [doi: 10.1016/j.euroneuro.2019.09.006] [Medline: 31606303]
67. Fusar-Poli P, Salazar de Pablo G, De Micheli A, Nieman DH, Correll CU, Kessing LV, et al. What is good mental health? A scoping review. Eur Neuropsychopharmacol 2020 Feb;31:33-46 [FREE Full text] [doi: 10.1016/j.euroneuro.2019.12.105] [Medline: 31901337]
68. Salazar de Pablo G, De Micheli A, Nieman DH, Correll CU, Kessing LV, Pfennig A, et al. Universal and selective interventions to promote good mental health in young people: systematic review and meta-analysis. Eur Neuropsychopharmacol 2020 Dec;41:28-39 [FREE Full text] [doi: 10.1016/j.euroneuro.2020.10.007] [Medline: 33162291]
69. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. Enhancing the QUAlity and Transparency Of health Research Network. URL: https://www.equator-network.org/reporting-guidelines/consort/ [accessed 2023-07-19]
70. Altman DG. Better reporting of randomised controlled trials: the CONSORT statement. BMJ 1996 Sep 07;313(7057):570-571 [FREE Full text] [doi: 10.1136/bmj.313.7057.570] [Medline: 8806240]
71. Higgins JP, Green S. Cochrane Handbook for Systematic Reviews of Interventions. London, UK: Cochrane Collaboration; 2008.
72. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Cambridge, MA: Academic Press; 1969.

73. Partlett C, Riley RD. Random effects meta-analysis: coverage performance of 95% confidence and prediction intervals following REML estimation. Stat Med 2017 Jan 30;36(2):301-317 [FREE Full text] [doi: 10.1002/sim.7140] [Medline: 27714841]
74. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 06;327(7414):557-560 [FREE Full text] [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
75. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634 [FREE Full text] [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
76. Sterne JA, Egger M. Regression methods to detect publication and other bias in meta-analysis. In: Rothstein HR, Sutton AJ, Borenstein M, editors. Publication Bias in Meta‐Analysis: Prevention, Assessment and Adjustments. Hoboken, NJ: John Wiley & Sons; 2005.
77. Krifa I, Hallez Q, van Zyl LE, Braham A, Sahli J, Ben Nasr S, et al. Effectiveness of an online positive psychology intervention among Tunisian healthcare students on mental health and study engagement during the Covid-19 pandemic. Appl Psychol Health Well Being 2022 Nov 22;14(4):1228-1254. [doi: 10.1111/aphw.12332] [Medline: 34939332]
78. Drabu S, Sündermann O, Hong RY. A one-week online self-compassion training reduces self-criticism and pain endurance in adults with non-suicidal self-injury ideation: a randomized-waitlist controlled study. Mindfulness 2022 Apr 06;13(5):1232-1245. [doi: 10.1007/S12671-022-01870-2]
79. Lennard GR, Mitchell AE, Whittingham K. Randomized controlled trial of a brief online self-compassion intervention for mothers of infants: effects on mental health outcomes. J Clin Psychol 2021 Mar 15;77(3):473-487. [doi: 10.1002/jclp.23068] [Medline: 33063321]
80. Andersson C, Bergsten KL, Lilliengren P, Norbäck K, Rask K, Einhorn S, et al. The effectiveness of smartphone compassion training on stress among Swedish university students: a pilot randomized trial. J Clin Psychol 2021 Apr 27;77(4):927-945. [doi: 10.1002/jclp.23092] [Medline: 33245161]
81. Beshai S, Bueno C, Yu M, Feeney JR, Pitariu A. Examining the effectiveness of an online program to cultivate mindfulness and self-compassion skills (Mind-OP): randomized controlled trial on Amazon's Mechanical Turk. Behav Res Ther 2020 Nov;134:103724. [doi: 10.1016/j.brat.2020.103724] [Medline: 32942203]
82. Kelman AR, Evare BS, Barrera AZ, Muñoz RF, Gilbert P. A proof-of-concept pilot randomized comparative trial of brief internet-based compassionate mind training and cognitive-behavioral therapy for perinatal and intending to become pregnant women. Clin Psychol Psychother 2018 Feb 23 (forthcoming). [doi: 10.1002/cpp.2185] [Medline: 29473698]
83. Kappen G, Karremans JC, Burk WJ. Effects of a short online mindfulness intervention on relationship satisfaction and partner acceptance: the moderating role of trait mindfulness. Mindfulness 2019 Jul 2;10(10):2186-2199. [doi: 10.1007/s12671-019-01174-y]
84. Koydemir S, Sun-Selışık ZE. Well-being on campus: testing the effectiveness of an online strengths-based intervention for first year college students. Brit J Guid Counsel 2015 Nov 16;44(4):434-446. [doi: 10.1080/03069885.2015.1110562]
85. Sergeant S, Mongrain M. An online optimism intervention reduces depression in pessimistic individuals. J Consult Clin Psychol 2014 Apr;82(2):263-274. [doi: 10.1037/a0035536] [Medline: 24417602]
86. Lappalainen P, Lappalainen R, Keinonen K, Kaipainen K, Puolakanaho A, Muotka J, et al. In the shadow of COVID-19: a randomized controlled online ACT trial promoting adolescent psychological flexibility and self-compassion. J Contextual Behav Sci 2023 Jan;27:34-44 [FREE Full text] [doi: 10.1016/j.jcbs.2022.12.001] [Medline: 36514308]
87. Webb JB, Padro MP, Thomas EV, Davies AE, Etzel L, Rogers CB, et al. Yoga at every size: a preliminary evaluation of a brief online size-inclusive yoga and body gratitude journaling intervention to enhance positive embodiment in higher weight college women. Front Glob Womens Health 2022 May 26;3:852854 [FREE Full text] [doi: 10.3389/fgwh.2022.852854] [Medline: 35720811]
88. Brouzos A, Vassilopoulos SP, Baourda VC, Tassi C, Stavrou V, Moschou K, et al. "Staying home - feeling positive": effectiveness of an on-line positive psychology group intervention during the COVID-19 pandemic. Curr Psychol 2023 Mar 20;42(4):2749-2761 [FREE Full text] [doi: 10.1007/s12144-021-01613-x] [Medline: 33776381]
89. Greer S, Ramo D, Chang YJ, Fu M, Moskowitz J, Haritatos J. Use of the chatbot "Vivibot" to deliver positive psychology skills and promote well-being among young people after cancer treatment: randomized controlled feasibility trial. JMIR Mhealth Uhealth 2019 Oct 31;7(10):e15018 [FREE Full text] [doi: 10.2196/15018] [Medline: 31674920]
90. Tagalidou N, Baier J, Laireiter AR. The effects of three positive psychology interventions using online diaries: a randomized-placebo controlled trial. Internet Interv 2019 Sep;17:100242 [FREE Full text] [doi: 10.1016/j.invent.2019.100242] [Medline: 30923680]
91. Bronk KC, Baumsteiger R, Mangan S, Riches B, Dubon V, Benavides C, et al. Fostering purpose among young adults: effective online interventions. J Character Educ 2019;15(2):21-38.
92. Gu X, Li S, Hyun MH. Does compassion-focused therapy-based online intervention work for Chinese international students with high self-criticism? A randomized controlled trail. SAGE Open 2022 Sep 15;12(3). [doi: 10.1177/21582440221121596]
93. Alexiou E, Kotsoni A, Stalikas A. The effectiveness of an online positive psychology intervention among healthcare professionals with depression, anxiety or stress symptoms and burnout. Psychology 2021;12(03):392-408. [doi: 10.4236/psych.2021.123025]

94. Manicavasagar V, Horswood D, Burckhardt R, Lum A, Hadzi-Pavlovic D, Parker G. Feasibility and effectiveness of a web-based positive psychology program for youth mental health: randomized controlled trial. J Med Internet Res 2014 Jun 04;16(6):e140 [FREE Full text] [doi: 10.2196/jmir.3176] [Medline: 24901900]
95. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [Medline: 19621072]
96. Chilver MR, Gatt JM. Six-week online multi-component positive psychology intervention improves subjective wellbeing in young adults. J Happiness Stud 2022 Sep 05;23(3):1267-1288 [FREE Full text] [doi: 10.1007/s10902-021-00449-3] [Medline: 34512122]
97. Hamm JM, Perry RP, Chipperfield JG, Parker PC, Heckhausen J. A motivation treatment to enhance goal engagement in online learning environments: assisting failure-prone college students with low optimism. Motiv Sci 2019 Jun;5(2):116-134. [doi: 10.1037/mot0000107]
98. Daugherty DA, Runyan JD, Steenbergh TA, Fratzke BJ, Fry BN, Westra E. Smartphone delivery of a hope intervention: another way to flourish. PLoS One 2018 Jun 1;13(6):e0197930 [FREE Full text] [doi: 10.1371/journal.pone.0197930] [Medline: 29856805]
99. Galante J, Bekkers MJ, Mitchell C, Gallacher J. Loving-kindness meditation effects on well-being and altruism: a mixed-methods online RCT. Appl Psychol Health Well Being 2016 Nov 23;8(3):322-350. [doi: 10.1111/aphw.12074] [Medline: 27333950]
100. Paetzold I, Schick A, Rauschenberg C, Hirjak D, Banaschewski T, Meyer-Lindenberg A, et al. A hybrid ecological momentary compassion-focused intervention for enhancing resilience in help-seeking young people: prospective study of baseline characteristics in the EMIcompass trial. JMIR Form Res 2022 Nov 04;6(11):e39511 [FREE Full text] [doi: 10.2196/39511] [Medline: 36331526]
101. Qu L, Chen H, Miller H, Miller A, Colombi C, Chen W, et al. Assessing the satisfaction and acceptability of an online parent coaching intervention: a mixed-methods approach. Front Psychol 2022 Jul 28;13:859145 [FREE Full text] [doi: 10.3389/fpsyg.2022.859145] [Medline: 35967644]
102. Sampson A, Jeremiah HG, Andiappan M, Newton JT. The effect of viewing idealised smile images versus nature images via social media on immediate facial satisfaction in young adults: a randomised controlled trial. J Orthod 2020 Mar 07;47(1):55-64. [doi: 10.1177/1465312519899664] [Medline: 32031041]
103. Nawa NE, Yamagishi N. Enhanced academic motivation in university students following a 2-week online gratitude journal intervention. BMC Psychol 2021 May 13;9(1):71 [FREE Full text] [doi: 10.1186/s40359-021-00559-w] [Medline: 33980290]
104. Hussong AM, Coffman JL, Thomas TE. Gratitude conversations: an experimental trial of an online parenting tool. J Posit Psychol 2020 May 01;15(2):267-277 [FREE Full text] [doi: 10.1080/17439760.2019.1610484] [Medline: 32477421]
105. Halamová J, Kanovský M, Varšová K, Kupeli N. Randomised controlled trial of the new short-term online emotion focused training for self-compassion and self-protection in a nonclinical sample. Curr Psychol 2021 Jul 30;40(1):333-343 [FREE Full text] [doi: 10.1007/s12144-018-9933-4] [Medline: 33488040]
106. Tay JL. Online HOPE intervention on help-seeking attitudes and intentions among young adults in Singapore: a randomized controlled trial and process evaluation. Arch Psychiatr Nurs 2022 Dec;41:286-294. [doi: 10.1016/j.apnu.2022.09.008] [Medline: 36428062]
107. Pizarro-Ruiz JP, Ordóñez-Camblor N, Del-Líbano M, Escolar-LLamazares MC. Influence on forgiveness, character strengths and satisfaction with life of a short mindfulness intervention via a Spanish smartphone application. Int J Environ Res Public Health 2021 Jan 19;18(2):802 [FREE Full text] [doi: 10.3390/ijerph18020802] [Medline: 33477831]
108. Halamová J, Kanovský M, Jakubcová K, Kupeli N. Short online compassionate intervention based on mindful self-compassion program. Československá Psychologie 2020;64(2):236-250 [FREE Full text]
109. Halamová J, Kanovský M, Jurková V, Kupeli N. Effect of a short-term online version of a mindfulness-based intervention on self-criticism and self-compassion in a nonclinical sample. Stud Psychol 2018 Dec;60(4):259-273 [FREE Full text] [doi: 10.21909/sp.2018.04.766]
110. Drozd F, Mork L, Nielsen B, Raeder S, Bjørkli CA. Better days – a randomized controlled trial of an internet-based positive psychology intervention. J Posit Psychol 2014 Apr 22;9(5):377-388. [doi: 10.1080/17439760.2014.910822]
111. Halamová J, Kanovský M, Pačutová A, Kupeli N. Randomised controlled trial of an online version of compassion mind training in a nonclinical sample. Eur J Psychol 2020 May 29;16(2):262-279 [FREE Full text] [doi: 10.5964/ejop.v16i2.1683] [Medline: 33680182]
112. Mahalik JR, Di Bianca M, Martin NG. Evaluation of brief online interventions to increase sense of purpose for fathers living in the United States. Psychol Men Masculinities 2022 Dec 05;24(1):26-33. [doi: 10.1037/men0000414]
113. Biber DD, Melton B, Czech DR. The impact of COVID-19 on college anxiety, optimism, gratitude, and course satisfaction. J Am Coll Health 2022 Oct 30;70(7):1947-1952. [doi: 10.1080/07448481.2020.1842424] [Medline: 33253625]
114. Marino C, Gini G, Vieno A, Spada MM. The associations between problematic Facebook use, psychological distress and well-being among adolescents and young adults: a systematic review and meta-analysis. J Affect Disord 2018 Jan 15;226:274-281. [doi: 10.1016/j.jad.2017.10.007] [Medline: 29024900]

115. Raeside R, Jia SS, Todd A, Hyun K, Singleton A, Gardner LA, et al. Are digital health interventions that target lifestyle risk behaviors effective for improving mental health and wellbeing in adolescents? A systematic review with meta-analyses. Adolescent Res Rev 2023 Aug 12;9(2):193-226. [doi: 10.1007/s40894-023-00224-w]
116. Kuyken W, Ball S, Crane C, Ganguli P, Jones B, Montero-Marin J, The Myriad Team, et al. Effectiveness and cost-effectiveness of universal school-based mindfulness training compared with normal school provision in reducing risk of mental health problems and promoting well-being in adolescence: the MYRIAD cluster randomised controlled trial. Evid Based Ment Health 2022 Jul 12;25(3):99-109 [FREE Full text] [doi: 10.1136/ebmental-2021-300396] [Medline: 35820992]
117. Montero-Marin J, Hinze V, Crane C, Dalrymple N, Kempnich ME, Lord L, MYRIAD Team, et al. Do adolescents like school-based mindfulness training? Predictors of mindfulness practice and responsiveness in the MYRIAD trial. J Am Acad Child Adolesc Psychiatry 2023 Nov;62(11):1256-1269 [FREE Full text] [doi: 10.1016/j.jaac.2023.02.016] [Medline: 37236303]

Abbreviations

CBT: cognitive behavioral therapy
CONSORT: Consolidated Standards of Reporting Trials
CT: controlled trial
PPI: positive psychology intervention
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial

Edited by J Torous; submitted 03.01.24; peer-reviewed by P Batterham, S Liu; comments to author 19.03.24; revised version received 14.05.24; accepted 17.05.24; published 14.08.24.

Please cite as:
Saboor S, Medina A, Marciano L
Application of Positive Psychology in Digital Interventions for Children, Adolescents, and Young Adults: Systematic Review and Meta-Analysis of Controlled Trials
JMIR Ment Health 2024;11:e56045
URL: https://mental.jmir.org/2024/1/e56045
doi:10.2196/56045
PMID:

Review

Self-Administered Interventions Based on Natural Language Processing Models for Reducing Depressive and Anxiety Symptoms: Systematic Review and Meta-Analysis

David Villarreal-Zegarra[1,2*], MPH; C Mahony Reategui-Rivera[2*], MD; Jackeline García-Serna[1], BSc; Gleni Quispe-Callo[1], BSc; Gabriel Lázaro-Cruz[1], Mg; Gianfranco Centeno-Terrazas[1], MPH; Ricardo Galvez-Arevalo[3], Mg; Stefan Escobar-Agreda[4], MD; Alejandro Dominguez-Rodriguez[5], PhD; Joseph Finkelstein[2], MD, PhD

[1]Instituto Peruano de Orientación Psicológica, Lima, Peru
[2]Department of Biomedical Informatics, School of Medicine, University of Utah, Salt Lake City, UT, United States
[3]Instituto Nacional de Salud del Niño San Borja, Lima, Peru
[4]Telehealth Unit, Universidad Nacional Mayor de San Marcos, Lima, Peru
[5]Department of Psychology, Health, and Technology, University of Twente, Enschede, Netherlands
[*]these authors contributed equally

Corresponding Author:
C Mahony Reategui-Rivera, MD
Department of Biomedical Informatics
School of Medicine
University of Utah
421 Wakara Way
Salt Lake City, UT, 84108
United States
Phone: 1 (801) 581 4080
Email: mahony.reategui@utah.edu

Abstract

Background: The introduction of natural language processing (NLP) technologies has significantly enhanced the potential of self-administered interventions for treating anxiety and depression by improving human-computer interactions. Although these advances, particularly in complex models such as generative artificial intelligence (AI), are highly promising, robust evidence validating the effectiveness of the interventions remains sparse.

Objective: The aim of this study was to determine whether self-administered interventions based on NLP models can reduce depressive and anxiety symptoms.

Methods: We conducted a systematic review and meta-analysis. We searched Web of Science, Scopus, MEDLINE, PsycINFO, IEEE Xplore, Embase, and Cochrane Library from inception to November 3, 2023. We included studies with participants of any age diagnosed with depression or anxiety through professional consultation or validated psychometric instruments. Interventions had to be self-administered and based on NLP models, with passive or active comparators. Outcomes measured included depressive and anxiety symptom scores. We included randomized controlled trials and quasi-experimental studies but excluded narrative, systematic, and scoping reviews. Data extraction was performed independently by pairs of authors using a predefined form. Meta-analysis was conducted using standardized mean differences (SMDs) and random effects models to account for heterogeneity.

Results: In all, 21 articles were selected for review, of which 76% (16/21) were included in the meta-analysis for each outcome. Most of the studies (16/21, 76%) were recent (2020-2023), with interventions being mostly AI-based NLP models (11/21, 52%); most (19/21, 90%) delivered some form of therapy (primarily cognitive behavioral therapy: 16/19, 84%). The overall meta-analysis showed that self-administered interventions based on NLP models were significantly more effective in reducing both depressive (SMD 0.819, 95% CI 0.389-1.250; *P*<.001) and anxiety (SMD 0.272, 95% CI 0.116-0.428; *P*=.001) symptoms compared to various control conditions. Subgroup analysis indicated that AI-based NLP models were effective in reducing depressive symptoms (SMD 0.821, 95% CI 0.207-1.436; *P*<.001) compared to pooled control conditions. Rule-based NLP models showed effectiveness in reducing both depressive (SMD 0.854, 95% CI 0.172-1.537; *P*=.01) and anxiety (SMD 0.347, 95% CI 0.116-0.578; *P*=.003) symptoms. The meta-regression showed no significant association between participants' mean age and treatment outcomes (all

P>.05). Although the findings were positive, the overall certainty of evidence was very low, mainly due to a high risk of bias, heterogeneity, and potential publication bias.

Conclusions: Our findings support the effectiveness of self-administered NLP-based interventions in alleviating depressive and anxiety symptoms, highlighting their potential to increase accessibility to, and reduce costs in, mental health care. Although the results were encouraging, the certainty of evidence was low, underscoring the need for further high-quality randomized controlled trials and studies examining implementation and usability. These interventions could become valuable components of public health strategies to address mental health issues.

Trial Registration: PROSPERO International Prospective Register of Systematic Reviews CRD42023472120; https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42023472120

(JMIR Ment Health 2024;11:e59560) doi:10.2196/59560

KEYWORDS

natural language processing; depression; anxiety; systematic review; artificial intelligence; AI

Introduction

Background

Depression and anxiety pose a substantial worldwide burden. In 2020, depression and anxiety affected approximately 246 million and 374 million people, respectively [1]. Moreover, these conditions reduce individuals' quality of life and have significant economic repercussions [2]. The World Health Organization estimates that depression and anxiety result in a loss of US $1 trillion annually due to loss of productivity [3]. In addition, their increasing incidence and a lack of health resources challenge the health care systems and workforce to meet the growing demand for mental health care services adequately [4].

In response, self-administered technology-based interventions have emerged as promising solutions for managing these conditions. These self-guided interventions enable users to progress through treatments independently, without external support [4], and they have demonstrated the potential to reduce costs; save health providers' time; and improve satisfaction and access to care, especially during crises and quarantine periods, for patients with mental health conditions living in remote areas, those with disabilities, or those unable to afford traditional care [5]. However, despite the potential of self-directed interventions to manage mental health problems, many of these interventions face important challenges in user engagement and adherence [6].

Self-administered interventions that are effective vary by delivery format, including web-based platforms, mobile apps, and virtual or augmented reality [7,8]. These interventions can be integrated within a professional intervention package or be completely independent of any external support [9,10]. Furthermore, they can be based solely on the presentation of relevant therapeutic information, typically based on a behavioral cognitive approach [10-12], or rely on machine learning (ML) models to process the natural language of clients' responses [13].

Natural language processing (NLP) offers a promising avenue for enhancing the efficacy of self-administered interventions. Defined as a cross-disciplinary field focused on enabling computers to comprehend, process, and interact with human language [14], NLP has the potential to make self-directed interventions more cost-effective and accessible and facilitate fidelity and engagement of patients through better interaction [15].

Moreover, NLP can be categorized into 2 main approaches: rule based and artificial intelligence (AI) based. Rule-based NLP uses predefined linguistic rules to guide text interpretation, offering high explainability but limited flexibility in handling complex language nuances [16]. Conversely, AI-based NLP, encompassing ML and deep learning techniques, learns from extensive data to process language. It has shown remarkable success in various NLP tasks due to its scalability and ability to manage linguistic ambiguities [17].

The advent of large language models and multimodal large language models has further enhanced the capabilities of NLP-based health interventions. These advances are not limited to enhanced user interaction but extend to personalizing therapeutic modalities to the patient's unique requirements, as demonstrated in specific psychotherapeutic settings [18].

Previously, other systematic reviews, such as those conducted by Le Glaz et al [19] and Zhang et al [20], analyzed the impact of NLP on mental health. However, these reviews primarily focused on the general applications of NLP in mental health. In addition, another systematic review demonstrated promising results for NLP-based interventions in mental health, but the findings encompassed a broad range of mental health disorders and did not specifically address self-administered interventions [15].

Objectives

Although these advances are highly promising, analysis of their effectiveness and safety in managing mental health concerns such as depression and anxiety remains fragmented [21]. This study aims to systematically review available literature to determine the effect of self-administered NLP-based interventions on symptoms of depression and anxiety.

Methods

Design and Protocol Registration

This study systematically searched available literature in the principal health databases and synthesized the main quantitative results in a meta-analysis. Our study adheres to the PRISMA

(Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (refer to Multimedia Appendix 1 for the PRISMA 2020 checklist) and the Cochrane Collaboration recommendations for meta-analyses [22]. The protocol for this systematic review was registered in the PROSPERO repository (CRD42023472120).

Eligibility Criteria

Our study follows the PICOS (Population, Intervention, Comparison, Outcomes, and Study Design) framework to evaluate whether interventions based on NLP models can effectively reduce depressive and anxiety symptoms. We define these symptoms as follows: (1) depressive symptoms are defined as a mood disorder characterized by the persistent presence of a profound sense of sadness, loss of interest or pleasure in daily activities, and a general lack of energy; and (2) anxiety symptoms are characterized by the anticipation of imagined events that are perceived as potential threats, causing emotional distress and physiological tension.

The eligibility criteria for our review are presented in Textbox 1.

Textbox 1. Eligibility criteria for the review determined using the PICOS (Population, Intervention, Comparison, Outcomes, and Study Design) framework.

Review eligibility criteria

- Population: we included studies with participants of any age group (child, adolescent, adult, and older adult) with or without previous comorbidities. Eligible studies must report participants who have been diagnosed with depression or anxiety through an interview or consultation with a mental health professional (eg, physician, psychologist, or psychiatrist) or assessed using validated psychometric instruments.
- Intervention: the intervention must be based on natural language processing (NLP) models such as large language models, multimodal large language models, artificial intelligence–led systems (ie, digital conversational agent, chatbot, or interactive voice response), and other NLP models. We included interventions regardless of their primary design purpose, provided they were self-administered.
- Comparison: we considered both passive (ie, waiting lists, nonintervention control groups, or placebos) and active (ie, web-based or face-to-face psychological interventions, virtual reality, serious games, biofeedback for mental health problems, pharmacological therapies to treat symptoms of depression and anxiety, or animal-assisted therapies) comparators.
- Outcomes: we included studies measuring depressive and anxiety symptom scores using validated psychometric questionnaires (eg, Patient Health Questionnaire-9, Beck Depression Inventory, Hamilton Depression Rating Scale, Generalized Anxiety Disorder-7, Beck Anxiety Inventory, Hamilton Anxiety Rating Scale, or similar instruments).
- Study Design: we included randomized controlled trials and quasi-experimental studies (without a control arm or randomization groups) that assessed the effect of NLP-based interventions on depressive and anxiety symptoms. We excluded narrative reviews, systematic reviews, scoping reviews, and other nonoriginal research designs. Only peer-reviewed publications (original articles or briefs) were included; proceedings, posters, and other similar items were excluded. There were no exclusion criteria based on language, publication date, or setting (ie, clinical or community settings).

Information Sources and Search Strategy

The databases we used for the systematic review were Web of Science, Scopus, MEDLINE (by PubMed), PsycINFO (by EBSCO), IEEE Xplore, Embase, and Cochrane Library. The search strategy included terms related to NLP as well as depression and anxiety, along with health science descriptors (refer to Multimedia Appendix 2 for the search strategy). Our search included any document available from inception to November 3, 2023.

Selection Process

We downloaded all records identified by the search strategy in RIS format and compiled them into an EndNote (Clarivate) file, which served as a repository for all retrieved records. Next, we used automated and manual methods to remove duplicate records. We exported the list of unique records from EndNote to Rayyan (Rayyan Systems Inc) for the selection process. First, 2 pairs of authors (JG-S with RG-A and GQ-C with GL-C) independently assessed the abstracts and titles of the studies to ensure that they met the inclusion criteria. Two pairs of authors reviewed the resulting retrieved text independently (JG-S with RG-A and GQ-C with GL-C). Any excluded studies were recorded along with the reasons for their exclusion (refer to Multimedia Appendix 3 for a list of the excluded studies). If disagreements arose between the reviewers at either stage, they were resolved by discussion. A third reviewer (DV-Z) was consulted if disagreement persisted to decide whether the study met the inclusion criteria. Records were included or excluded depending on whether they met the inclusion criteria. At the title and abstract stage, if it was unclear whether a record met all the inclusion criteria, it could proceed to the full-text stage, where a more detailed review was carried out (a sensitive approach). However, at the full-text stage, all inclusion criteria had to be met for final acceptance.

The title and abstract review were performed in English because this is the language in which the databases save the metadata. The full-text review and results extraction were mainly performed in English and Spanish (the languages the reviewers speak). When studies in other languages were found, the reviewers used DeepL Translator (DeepL SE) to translate the documents into English before proceeding with the review and extraction. Therefore, our review had no language limitations. It is important to note that all papers evaluated in the full-text review and extraction were in English.

Data Collection

Two pairs of authors (JGS with RGA and GQC with GLC) independently collected the information from the included studies using a predefined collection form in a Microsoft Excel sheet. Initially, a pilot data extraction process was conducted on 5 data sets reviewed by all raters with 85% agreement. Subsequently, minor changes were made to the final version of the extraction form to improve the clarity of the extracted data, which included the following: (1) general information (ie, authors, year of publication, title, country, and language); (2) participant characteristics (ie, age range, sex, number of participants, and diagnosis); (3) intervention characteristics (ie, type of NLP model, duration, frequency, and brief description of the intervention); (4) comparator (passive or active); and (5) main outcomes (ie, means, SDs, preintervention and postintervention measures, and the effect size for control and intervention groups).

Risk-of-Bias Assessment and Certainty of Evidence

We used the JBI critical appraisal tools to identify potential biases that may have occurred during the design, conduct, and analysis of the studies. For quasi-experimental studies, we used the JBI critical appraisal checklist for quasi-experimental studies [23], which is a checklist with 9 questions for assessing potential bias. For randomized controlled trials (RCTs), we used the JBI critical appraisal tool for the assessment of risk of bias in RCTs [24], which is a 13-question checklist evaluating the internal and statistical validity of the conclusions of RCTs. On the basis of the answers from both assessment tools, reviewers decide whether to include the reviewed study. Two reviewers used these tools independently to assess the risk of bias in the studies included in the meta-analysis. Any disagreement between the reviewers about whether to include or exclude a study was resolved by discussion. If the disagreement persisted, a third reviewer was asked to arbitrate.

We used the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) methodology to assess the certainty of evidence regarding the intervention’s effects. This methodology evaluates the certainty of evidence based on several criteria, including risk of bias, inconsistency, indirectness, and imprecision [25]. Given that the GRADE approach is primarily focused on RCTs, and the GRADE working group has not yet reached a consensus on the combination of results from randomized and nonrandomized trials, we applied this evaluation exclusively to the RCTs included in our review.

Synthesis Methods

Narrative Synthesis

To address the multifaceted nature of the factors involved in self-administered NLP-based interventions for symptoms of depression and anxiety, we adopted a comprehensive framework for data synthesis based on an adaptation of the categories from the framework for NLP applications for mental health interventions proposed by Malgaroli et al [15] in the context of self-administered NLP interventions. This systematic approach thoroughly integrates all relevant factors, providing a coherent structure for our analysis. We categorized data from eligible studies into four primary domains: (1) demographic and sample descriptions, (2) NLP technical aspects, (3) clinical categories, and (4) intervention results. Due to the nature of our study, the last category is presented through the findings of the meta-analysis and analysis of subgroups.

Meta-Analysis

We performed analyses using Stata (version 18.0; StataCorp LLC). Meta-analysis was only performed if at least 3 studies of the same design type (ie, randomized or quasi-experimental controlled trials) assessing the same outcome were available. The analysis was differentiated by outcome and by study type. Standardized mean differences (SMDs) with 95% CIs were used for meta-analyses and summary statistics of the studies because the results of the included studies were measured using different scales. SMD is the mean difference between the intervention and control groups divided by the pooled SD.

The standard measure of effect size to be considered for the Hedges *g* analyses includes small (SMD 0.2), moderate (SMD 0.5), and large (SMD >0.8) effect sizes. These thresholds were used to evaluate the combined effect of the analyzed interventions using Hedges *g*. Hedges *g*, unlike Cohen *d*, corrects for possible risk of bias associated with small sample sizes, making it a more appropriate measure for our analyses [26].

Heterogeneity Analysis

The assessment of statistical heterogeneity involved the following tests: the Cochran Q test statistic to detect the presence of heterogeneity between studies, the I^2 Higgins and H^2 index statistics to measure the extent of variability between studies due to heterogeneity, and the between-study variance (τ^2) to assess the variance between the effects observed across the studies. If the overall assessment indicated high heterogeneity, random effects models were used to estimate the effect of the interventions in general.

Publication Bias Analysis

If there were >10 studies in the meta-analysis, we conducted both visual and quantitative tests to detect biases. Our visual examination used a funnel plot; the quantitative test used was the Egger regression test, which can capture the effects of small studies and other potential information biases [27]. We identified selection bias if we observed an asymmetric funnel plot distribution and a significant Egger regression test result (P<.05). In cases of asymmetry, the trim-and-fill method proposed by Duval and Tweedie [28] was implemented as a bias correction technique to estimate the number of missing studies for the meta-analysis.

Analysis of Subgroups

If the meta-analysis data allowed, we assessed intervention effects using the NLP-based models from the selected studies. Such models could include rule-based NLP, AI-based NLP, or other NLP. In addition, we assessed the impact of interventions on subgroups, including gender, disease severity, prior therapies, concurrent depression and anxiety disorders, and age ranges.

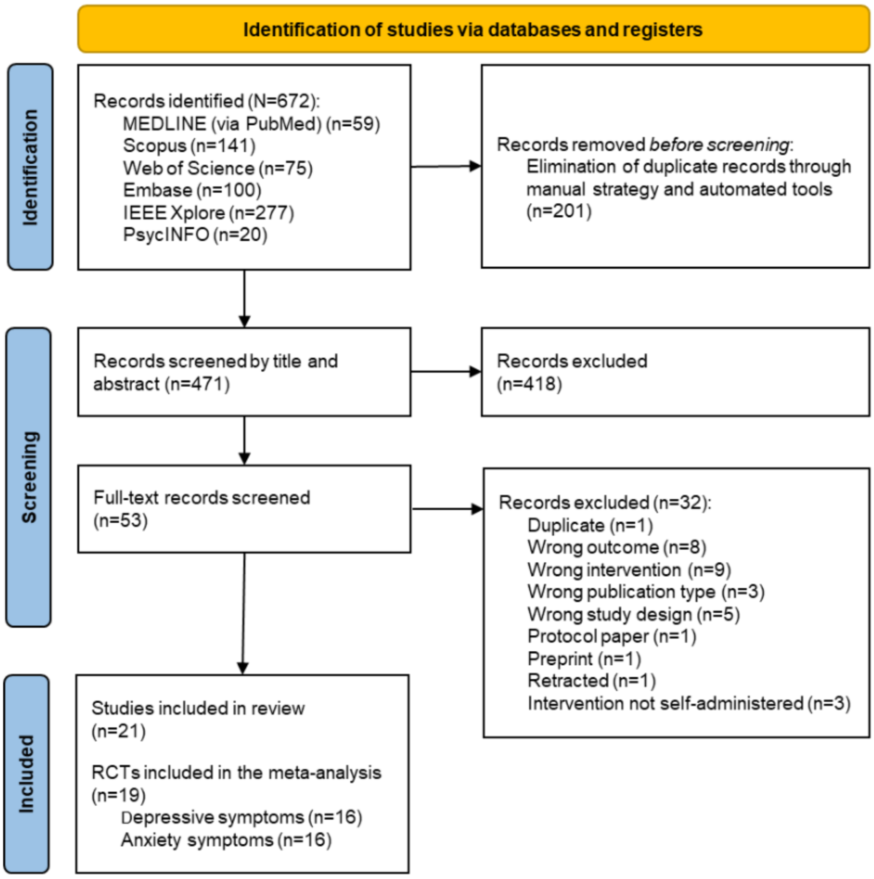
We performed a random effects meta-regression using aggregate-level data. Our analysis specified the variables containing the SE within each study using the *metareg* command and the *wsse* option in Stata. The meta-regression was a function of the mean age of the participants and was only applied to the overall meta-analysis. Our analysis yielded a meta-regression coefficient with 95% CI.

Results

Study Selection

Initially, 672 records were identified in the different databases; after eliminating 201 (29.9%) duplicates, 471 (70.1%) records advanced to title and abstract review. Of these 471 records, 418 (88.7%) were discarded, leaving 53 (11.3%) records for full-text review. Subsequently, 32 (60%) of the 53 records were excluded, resulting in 21 (40%) articles selected for review. Of these 21 articles, 19 (90%) were included in the meta-analysis on depressive and anxiety symptoms. Of the 19 studies included in the meta-analysis, 16 (84%) reported sufficient data for the meta-analysis of depressive symptoms, and another 16 (84%) reported sufficient data for the meta-analysis of anxiety symptoms. Figure 1 shows the complete review process, and Multimedia Appendices 3 and 4 [29-49] list the articles excluded and included, respectively.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart of the selection process. RCT: randomized clinical trial.



Characteristics of the Included Studies

Of the 21 studies identified, 19 (90%) were RCTs [29-47], and 2 (10%) were quasi-experimental studies without a control group [48,49]. Most of the studies (16/21, 76%) were published between 2020 and 2023, and 81% (17/21) were conducted in high-income countries. The United States was the country with the most publications among the selected studies (10/21, 48%). Regarding the characteristics of the populations studied, the majority (16/21, 76%) focused on adults. With regard to the outcomes assessed, depressive symptoms were analyzed in 95% (20/21) of the studies and anxiety symptoms in 90% (19/21).

We found 29 potential comparisons between interventions and controls because 5 (24%) of the 21 studies reported ≥3 arms. AI-based NLP applications were the most common intervention (11/21, 52%), while the most common control conditions were waiting list or no intervention (8/21, 38%) and information, psychoeducation, or bibliotherapy (8/21, 38%). The most commonly used scales to measure depressive and anxiety symptoms were the Patient Health Questionnaire (PHQ; PHQ-9 and PHQ-8; 13/21, 62%) and the Generalized Anxiety Disorder-7 (GAD-7; 10/21, 48%), respectively. Table 1 shows the characteristics of the studies, divided into RCTs and uncontrolled quasi-experimental studies.

Table 1. Characteristics of the included studies (n=21).

Characteristics	Randomized controlled trials (n=19), n (%)	Uncontrolled quasi-experimental studies (n=2), n (%)
Publication year		
2014-2015	1 (5)	0 (0)
2016-2019	4 (21)	0 (0)
2020-2023	14 (74)	2 (100)
Country income level		
Upper-middle income	4 (21)	0 (0)
High income	15 (79)	2 (100)
Country		
Argentina	1 (5)	0 (0)
China	3 (16)	0 (0)
Italy	1 (5)	1 (50)
Japan	1 (5)	0 (0)
South Korea	2 (11)	0 (0)
United Kingdom	2 (11)	0 (0)
United States	9 (47)	1 (50)
Study design		
Crossover	5 (26)	0 (0)
Parallel	14 (74)	0 (0)
Not applicable	0 (0)	2 (100)
Participants' life stage		
Adolescent	2 (11)	0 (0)
Adult	14 (74)	2 (100)
Older adult	2 (11)	0 (0)
Pregnant	1 (5)	0 (0)
Included in meta-analysis		
Depressive symptoms	16 (84)	0 (0)
Anxiety symptoms	16 (84)	0 (0)
Not applicable	0 (0)	2 (100)
Depressive symptoms		
Main outcome	11 (58)	2 (100)
Secondary outcome	7 (37)	0 (0)
Not evaluated	1 (5)	0 (0)
Anxiety symptoms		
Main outcome	11 (58)	2 (100)
Secondary outcome	6 (32)	0 (0)
Not evaluated	2 (11)	0 (0)
Funding		
Corporations	8 (42)	1 (50)
Government	4 (21)	0 (0)
Self-financed	2 (11)	1 (50)
Not reported	5 (26)	0 (0)
Conflicts of interest		

Characteristics	Randomized controlled trials (n=19), n (%)	Uncontrolled quasi-experimental studies (n=2), n (%)
Yes	4 (21)	2 (100)
No	12 (63)	0 (0)
Not reported	3 (16)	0 (0)
Study has ≥3 arms		
No	14 (74)	2 (100)
Yes	5 (26)	0 (0)
Control group[a]		
Waiting list or no intervention	8 (42)	0 (0)
Usual treatment	2 (11)	0 (0)
Information, psychoeducation, or bibliotherapy	8 (42)	0 (0)
Conversational computer-based intervention	5 (26)	0 (0)
Not applicable	0 (0)	2 (100)
Type of NLP[b] **application**[a]		
Rule based	10 (53)	1 (50)
AI[c] based	11 (58)	1 (50)
Focus of intervention[a]		
Depressive symptoms	8 (42)	1 (50)
Anxiety symptoms	7 (37)	1 (50)
Other mental health problems	13 (68)	2 (100)
Therapeutical approach[a]		
Cognitive behavioral therapy	15 (79)	2 (100)
Other	3 (16)	0 (0)
Unclear	1 (5)	0 (0)
Scale used to measure depression[a]		
PHQ[d]-9 and PHQ-8	13 (68)	2 (100)
DASS-21[e]	2 (11)	0 (0)
Other	3 (16)	0 (0)
Not evaluated	1 (5)	0 (0)
Scale used to measure anxiety[a]		
GAD-7[f]	10 (53)	2 (100)
DASS-21	3 (16)	0 (0)
Other	4 (21)	0 (0)
Not evaluated	2 (11)	0 (0)

[a]The totals do not add up to 100% because there are studies with 3 and 4 arms that evaluated >1 type of intervention at the same time.
[b]NLP: natural language processing.
[c]AI: artificial intelligence.
[d]PHQ: Patient Health Questionnaire.
[e]DASS-21: Depression, Anxiety, and Stress Scale-21.
[f]GAD-7: Generalized Anxiety Disorder-7.

NLP Technical Aspects

Of the 21 included studies, 10 (48) used rule-based approaches, while 11 (52%) used AI-based techniques. Within the AI-based category, of the 11 studies, 4 (36%) implemented deep learning methods, 6 (55%) did not specify the AI technique used, and 1 (9%) used ML algorithms. Regarding the specific NLP techniques used, sentiment analysis was used in 18% (2/11) of the studies, and natural language understanding was used in 18% (2/11). Notably, 7 (64%) of the 11 studies did not specify the NLP techniques used in their interventions. This distribution highlights a diverse application of NLP methods in addressing symptoms of depression and anxiety, with more than half of the studies (11/21, 52%) leveraging advanced AI techniques, albeit often without detailed specification (7/11, 64%).

The input modality for the NLP interventions was primarily text based in 19 (90%) of the 21 studies, while 1 (5%) study used either text or voice, and 1 (5%) study used voice alone. Regarding output modalities, text was predominantly used in 20 (95%) of the 21 studies, while only 1 (5%) study used voice. The language of the NLP input and output varied among the studies. Of the 21 studies, 7 (33%) used English, and 3 (14%) used Chinese, while Japanese, Spanish, and Italian were used in 1 (5%) study each. However, 38% (8/21) of the studies did not specify the language used for the NLP input and output.

Demographics and Sample Descriptions

Overview

The study participants' demographic characteristics were analyzed for rule-based NLP studies and AI-based NLP studies. All 21 studies provided demographic information regarding the sample or testing data set used for the intervention. Demographic data for rule-based NLP studies are reported only for the intervention samples. By contrast, AI-based NLP studies were expected to provide demographic information for the training data used to develop the AI-based models and the participants involved in the intervention or experiment.

Training Sample Description

None of the AI-based NLP studies provided detailed demographic information regarding the training data. While 3 (27%) of the 11 AI-based NLP studies mentioned the source of their training data (Stanford Sentiment Treebank data set, ad hoc user utterances from an unspecified source, and Emotion Support Conversation data set), they did not describe the demographic characteristics of these data sets.

Testing Data or Intervention Sample Description

Across all studies, gender distribution varied significantly. Of the 21 studies, in 3 (14%), only women participated; in 16 (76%), >50% of the participants were women; and in 2 (10%), >50% of the participants were men. Regarding the age of the participants, 20 (95%) of the 21 studies reported the mean age of their samples. Of these 20 studies, 9 (45%) involved participants aged >30 years, 10 (50%) included participants aged between 18 and 29 years, and 1 (5%) included participants aged <18 years. Participants' special conditions were also considered in the analysis. Of the 21 studies, 4 (19%) included participants with chronic diseases, 7 (33%) focused on individuals with mental disorders, and 7 (33%) included university students, while 4 (19%) involved participants with other conditions. Specifically, among the 7 studies that focused on mental disorders, 2 (29%) included participants with a positive screening for depression, and 2 (29%) focused on participants with a positive screening for substance use disorder. Among the 4 studies that included participants with chronic diseases, there were diverse conditions, such as diabetes mellitus (n=1, 25%), cancer (n=1, 25%), inflammatory bowel disease (n=1, 25%), and dementia (n=1, 25%).

Focusing on the 11 AI-based NLP studies, the gender distribution of the intervention samples was as follows: in 9 (82%) studies, the majority of the participants were women; and in 2 (18%) studies, the majority of the participants were men. Regarding age distribution, of the 10 studies that reported mean ages, 5 (50%) involved participants aged >30 years, and 5 (50%) included participants aged between 18 and 29 years. With regard to special conditions in the intervention samples, of the 11 studies, 2 (18%) included participants with chronic diseases, 2 (18%) focused on individuals with mental disorders, 6 (55%) included university students, and 2 (18%) involved participants with other conditions (participants with panic disorder: n=1, 50%; and participants with a positive screening for depression: n=1, 50%). For chronic conditions, of the 2 studies, 1 (50%) involved patients with dementia, and 1 (50%) included patients with diabetes mellitus.

Clinical Categories

The included studies were evaluated for their focus on clinical presentation and the delivery of therapeutic interventions. Only 1 (5%) of the 21 studies reported having a component of diagnosis and screening for mental health problems, although it did not specify the disease or the methods used for diagnosis.

Most of the studies (19/21, 90%) declared that they delivered some form of therapy through their NLP interventions. By contrast, 2 (10%) of the 21 studies did not include any therapeutic component. Among the 19 studies that delivered therapy, 16 (84%) implemented cognitive behavioral therapy, 1 (5%) combined cognitive behavioral therapy with dialectical behavioral therapy, and 2 (11%) reported delivering therapy but did not specify the therapeutic approach used.

Meta-Analysis Findings

Main Meta-Analysis

Only 16 (76%) of the 21 studies were included in the meta-analysis, excluding the uncontrolled quasi-experimental studies (n=2, 10%) and the RCTs with insufficient data for meta-analysis (n=3, 14%). The rationale for excluding the 2 quasi-experimental studies was that a meta-analysis specific to this study design required at least 3 studies of the same design type assessing the same outcome. For the depressive symptoms (Figure 2 [29-37,39-45]), the overall meta-analysis showed that self-administered interventions based on NLP models were significantly more effective in reducing depressive symptoms compared to various control conditions (waiting list or no intervention, treatment as usual, psychoeducation, and other computer-based conversational interventions; SMD 0.819, 95% CI 0.389-1.250; *P*<.001). In addition, high heterogeneity was observed in the overall meta-analysis (I^2=92.7%, 95% CI 78.3%-96.4%; H^2=3.71, 95% CI 2.15-5.27; τ^2=0.97; *P*<.001). Regarding publication bias, the funnel plot analysis showed evidence of bias (Egger test coefficient=3.61, 95% CI 0.45-6.78; *P*=.03; Multimedia Appendix 5).

For the outcome of anxiety symptoms (Figure 3 [30-34,36-46]), the global meta-analysis showed that self-administered NLP model–based interventions were significantly more effective in reducing depressive symptoms compared to various control conditions (waitlist or no intervention, treatment as usual, psychoeducation, and other conversational computer-based interventions; SMD 0.272; 95% CI 0.116-0.428; *P*=.001). In addition, high heterogeneity was observed in the overall meta-analysis (I^2=64%, 95% CI 0.5%-81.6%; H^2=1.67, 95% CI 1.00-2.33; τ^2=0.07; *P*<.001). Regarding publication bias, the funnel plot analysis showed no evidence of bias (Egger test coefficient=–0.22, 95% CI –1.55 to 1.11; *P*=.73; Multimedia Appendix 5).

Figure 2. Forest plot for control conditions versus self-administered interventions based on natural language processing models to reduce depressive symptoms.

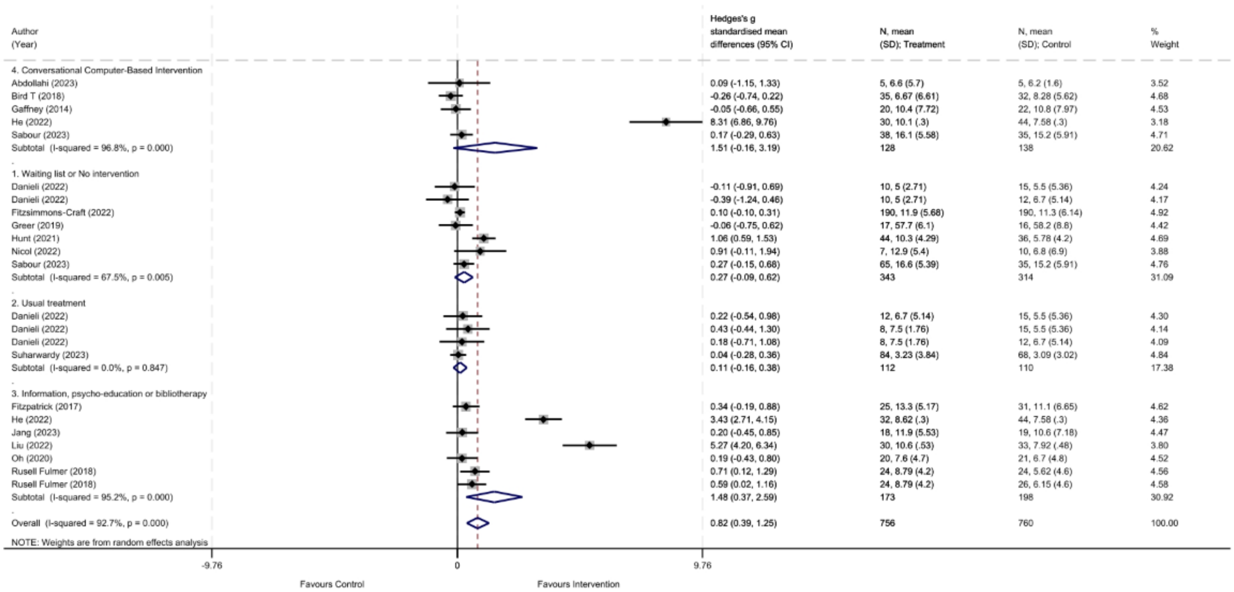
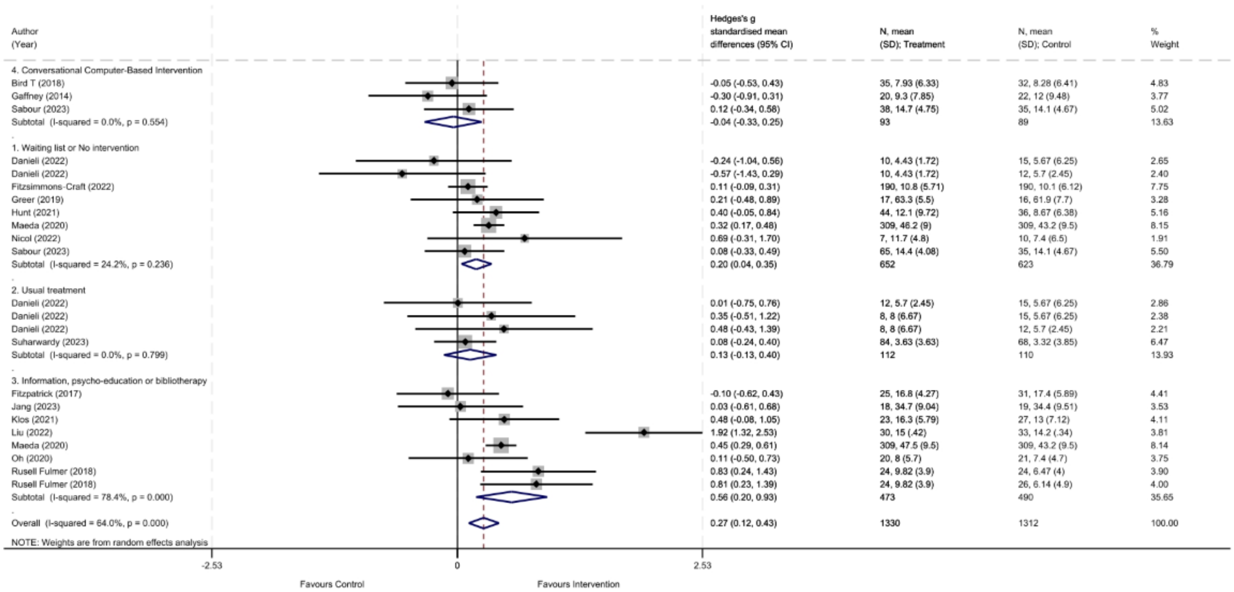


Figure 3. Forest plot for control conditions versus self-administered interventions based on natural language processing models to reduce anxiety symptoms.



Subgroup Analyses

We also conducted a detailed analysis according to the type of comparator, intervention, and the scale used, evaluating the results for depressive symptoms and anxiety symptoms separately. For depressive symptoms, self-administered interventions based on NLP models were found to be more effective than information, psychoeducation, or bibliotherapy (SMD 1.481, 95% CI 0.368-2.594; *P*=.009). Similarly, AI-based NLP models were more effective than the set of control conditions (SMD 1.059, 95% CI 0.520-1.597; *P*<.001) for reducing depressive symptoms. Regarding the scale used, studies using the PHQ-9 or PHQ-8 showed that self-administered interventions based on NLP outperformed the set of control conditions (SMD 0.914, 95% CI 0.417-1.410; *P*<.001).

For the outcome of anxiety symptoms, self-administered interventions based on NLP models were more effective than waitlist or no intervention (SMD 0.196, 95% CI 0.042-0.351; *P*=.01) and information, psychoeducation, or bibliotherapy (SMD 0.561, 95% CI 0.195-0.927; *P*=.003). In addition, the use of AI-based NLP models had a higher effect than the average of the control conditions (SMD 0.302, 95% CI 0.073-0.532; *P*=.01) in reducing anxiety symptoms. Regarding the scale used, studies using the GAD-7 showed that self-administered interventions based on NLP had a higher effect than the average of the control conditions in reducing anxiety symptoms (SMD 0.333, 95% CI 0.074-0.592; *P*=.01). Full details of this subgroup analysis are presented in Table 2.

Given that factors such as age may influence the outcomes of depressive and anxiety symptoms, we performed a meta-regression to assess whether the mean age of participants affected the overall meta-analysis results. Our analysis revealed that the mean age was not significantly associated with the point estimates for either depressive symptoms (coefficient=–0.037, 95% CI –0.092 to 0.019; *P*=.18) or anxiety symptoms (coefficient=–0.010, 95% CI –0.030 to 0.010; *P*=.29). Detailed results of the meta-regression are presented in Table 3.

Table 2. Meta-analysis by subgroup for depressive and anxiety symptoms.

Symptoms and subgroups	Studies, n (%); groups, n	SMD[a] (95% CI)	*P* value	Heterogeneity (I^2; %)	Cochran Q test (*P* value)
Depressive symptoms (n=16)					
By control group					
Waiting list or no intervention	6 (38); 7	0.267 (–0.085 to 0.620)	.14	67.5	.005
Usual treatment	2 (12); 4	0.111 (–0.155 to 0.378)	.41	0	.85
Information, psychoeducation, or bibliotherapy	6 (38); 7	*1.481 (0.368 to 2.594)* [b]	*.009*	95.2	<.001
Conversational computer-based intervention	5 (31); 5	1.513 (–0.162 to 3.188)	.08	96.8	<.001
By intervention group					
Rule-based NLP[c] model	7 (44); 7	*0.854 (0.172 to 1.537)*	*.01*	94	<.001
AI[d]-based NLP model	9 (56); 16	*0.821 (0.207 to 1.436)*	*.009*	92.5	<.001
By scale used					
PHQ[e]-9 and PHQ-8	11 (69); 17	*0.914 (0.417 to 1.410)*	*<.001*	92.8	<.001
DASS-21[f]	2 (12); 2	—[g]	—	—	—
Anxiety symptoms (n=16)					
By control group					
Waiting list or no intervention	7 (44); 8	*0.196 (0.042 to 0.351)*	*.01*	24.2	.24
Usual treatment	2 (12); 4	0.133 (–0.134 to 0.400)	.33	0	.80
Information, psychoeducation, or bibliotherapy	7 (44); 8	*0.561 (0.195 to 0.927)*	*.003*	78.4	<.001
Conversational computer-based intervention	3 (19); 3	–0.041 (–0.333 to 0.250)	.78	0	.55
By intervention group					
Rule-based NLP model	8 (50); 9	*0.347 (0.116 to 0.578)*	*.003*	79.7	<.001
AI-based NLP model	8 (50); 14	0.198 (–0.011 to 0.406)	.06	34.4	.10
By scale used					
GAD-7[h]	9 (56); 15	*0.333 (0.074 to 0.592)*	*.01*	71.7	<.001
DASS-21	3 (19); 3	0.050 (–0.352 to 0.453)	.81	47.3	.15

[a]SMD: standardized mean difference.

[b]Italicized values are significant. Only meta-analyses with at least 3 measurements are presented in this study.

[c]NLP: natural language processing.

[d]AI: artificial intelligence.

[e]PHQ: Patient Health Questionnaire.

[f]DASS-21: Depression, Anxiety, and Stress Scale-21.

[g]There are not enough trials to do a meta-analysis.

[h]GAD-7: Generalized Anxiety Disorder-7.

Table 3. Meta-regression analysis by overall meta-analysis of depressive and anxiety symptoms.

Variable	Coefficient (SE; 95% CI)	*t (df)*	*P* value
Depressive symptoms			
Age, mean	–0.037 (0.026; –0.092 to 0.019)	–1.390 (18)	.18
Intercept	2.108 (1.033; –0.063 to 4.279)	2.040 (18)	.06
Anxiety symptoms			
Age, mean	–0.010 (0.009; –0.030 to 0.010)	–1.080 (16)	.29
Intercept	0.553 (0.329; –0.145 to 1.251)	1.680 (16)	.11

Risk of Bias and Certainty of Evidence

In the overall analysis of the risk of bias for the outcome of depressive symptoms, the majority of the studies (9/16, 56%) had an overall low risk of bias, while only 19% (3/16) had an overall high risk of bias (Figure 4A). Regarding the dimensions assessed, the lowest risk of bias was observed in reporting and analysis strategies (15/16, 94%), followed by participant loss or missing data (14/16, 88%). However, intervention delivery showed an unclear risk of bias due to limited reporting in the reviewed manuscripts. By contrast, for the outcome of anxiety symptoms, half of the studies (8/16, 50%) had an overall low risk of bias, while only 12% (2/16) had an overall high risk of bias (Figure 4B). At the level of each dimension assessed, all studies had a low risk of bias in reporting and analysis strategies, and 81% (13/16) had a low risk of bias in outcome measurement and retention throughout the study. Detailed risk-of-bias analyses for each study are available in Multimedia Appendix 6 for depressive symptoms and Multimedia Appendix 7 for anxiety symptoms.

We found that, for the outcomes studied (depressive symptoms and anxiety symptoms), the evidence was of very low certainty (Table 4). This was mainly due to several factors. First, there was a high risk of bias, with 3 (19%) of the 16 studies presenting an overall high risk of bias for depressive symptoms and 2 (12%) of the 16 studies presenting an overall high risk of bias for anxiety symptoms. Second, there was significant inconsistency, as indicated by an overall I^2 value of >60%. In addition, indirectness was a major concern due to the high variability in the interventions, controls, and sample characteristics across the studies. Finally, publication bias was strongly suspected due to the marked right-side asymmetry revealed by the funnel plot. Notwithstanding these limitations, the findings provide a preliminary understanding of the potential effects of self-administered NLP-based interventions on depressive and anxiety symptoms.

Figure 4. Risk of bias grouped for the outcomes of (A) depressive symptoms and (B) anxiety symptoms.

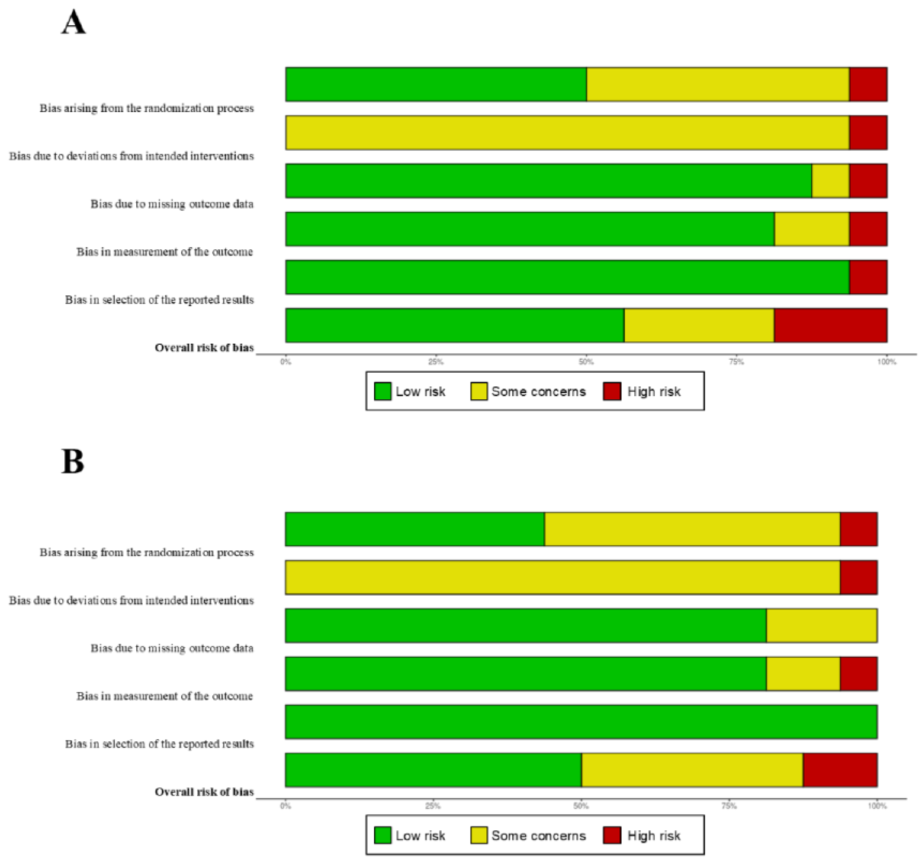


Table 4. Summary of findings and certainty of evidence using the Grading of Recommendations Assessment, Development, and Evaluation methodology.

Outcome	Assessment of certainty of evidence						Effect: Hedges *g*, SMD[a] (95% CI)	Certainty of evidence
	Studies (RCTs[b]), n (participants, n)	Risk of bias	Inconsistency	Indirectness	Imprecision	Publication bias		
Depressive symptoms	16 (1516; control: 760, intervention: 756)	Very serious[c]	Very serious[d]	Very serious[e]	Not serious	Strongly suspected[f]	0.82, lower (0.39-1.25)	⊕OOO[g]
Anxiety symptoms	16 (2642; control: 1312, intervention: 1330)	Very serious[h]	Very serious[d]	Very serious[e]	Not serious	Strongly suspected[f]	0.27, lower (0.12-0.43)	⊕OOO

[a]SMD: standardized mean difference.

[b]RCT: randomized controlled trial.

[c]Of the 16 studies, 3 (19%) present an overall high risk of bias.

[d]Overall I^2 value >60%.

[e]There is a high variability in the interventions, controls, and sample characteristics.

[f]The funnel plot reveals a marked right-side asymmetry.

[g]Very low (each filled circle [⊕] signifies a higher level of certainty, while each empty circle [O] indicates a lower level of certainty).

[h]Of the 16 studies, 2 (12%) present an overall high risk of bias.

Discussion

Principal Findings

Our results indicate that self-administered interventions based on NLP models have a significant overall effect on reducing depressive and anxiety symptoms compared to various control conditions. Our study used random effects models to estimate this overall effect, thus accounting for heterogeneity among the interventions analyzed. Therefore, we consider the results to be robust. At the level of each intervention group and control group, we observed variability in their effectiveness in reducing symptoms of depression and anxiety, which could be due to the limited number of studies available for meta-analysis. In particular, conversational computer-based interventions were shown to be effective in reducing depressive and anxiety symptoms compared to pooled control conditions. In addition, NLP-based interventions overall outperformed psychoeducation and bibliotherapy in reducing both depressive and anxiety symptoms. Furthermore, these interventions were more effective than waitlist or no intervention in reducing anxiety symptoms.

These findings support the usefulness of self-administered NLP-based interventions in alleviating such common mental health problems as depressive and anxiety symptoms. Thus, they have the potential to be implemented in primary care settings, where they could represent a valuable public health strategy to improve the mental health of the population.

Comparison With Other Studies

Our findings are consistent with previous research that has examined the application of NLP-based models at various stages of mental health care in both clinical and community settings [50-52], indicating that NLP-based interventions may effectively alleviate symptoms of emotional disorders. The robustness of our research is strengthened by the fact that most of the studies included in the meta-analysis of depressive (9/16, 56%) and anxiety symptoms (8/16, 50%) have a low risk of bias, indicating that our findings are derived from rigorous and reliable research.

A previous scoping review highlighted the heterogeneity of the tools used to assess the effects of dialogue interventions on mental health [53]. However, our review found that in the case of RCTs focusing on depressive and anxiety symptoms, validated instruments such as the PHQ-9 and GAD-7 were used, reducing the risk of bias and making the results more robust. Nevertheless, we highlight the lack of studies using experiential sampling or real-time measures to assess depressive and anxiety symptoms, which could provide a more accurate assessment of the impact of these self-administered NLP-based interventions.

The subgroup analysis showed variability in the effectiveness of the interventions in reducing depressive and anxiety symptoms, which may be due to the limited number of studies analyzed. Another possible explanation lies in the variety of NLP-based models used and their level of sophistication. Interventions using conversational agents based on advanced deep ML models showed significant results compared to other strategies, such as rule-based chatbots [54]. Unlike simpler NLP-based models, conversational agents offer better performance on various tasks [54]. However, more complex models also require high computational costs and large amounts of data for optimization [55,56], which may limit their adaptability to the different linguistic and cultural needs of different regions [57]. It is important to note that high-income countries have led research in this field and have advanced technological resources for developing these AI-based models compared to low- and middle-income countries [58,59]. This situation represents a challenge and a potential source of inequity in access to, and the implementation of, NLP-based interventions within public health systems.

Implications for Clinical Practice and Public Health

A previous systematic review on the general use of NLP and ML in mental health also identified the potential of NLP-based interventions to improve population mental health [19]. However, our study differs in that it focuses only on self-applied interventions to reduce depressive and anxiety symptoms, thus contributing to a specific aspect of NLP-based interventions. Our study provides a valuable starting point for future research to confirm the effectiveness of NLP-based interventions in the real world and their ability to be implemented within the public health system. There is a need to evaluate the implementation and promotion of these interventions as part of mental health strategies because this could be an effective strategy to reduce depressive and anxiety symptoms in health service users [60,61]. Given their accessibility through digital platforms, these interventions have the potential to reduce the burden of depressive and anxiety disorders at the population level [62,63] while also being cost-effective and a way to optimize mental health resources [64]. To ensure successful implementation within the public health system, using the Artificial Intelligence–Quality Implementation Framework could be beneficial [65]. However, it is crucial to develop protocols that ensure confidentiality and respect for the ethics and privacy of patient data at all stages of implementation and use [66]. In addition, it is important to consider the digital determinants of health [67], such as access to appropriate devices, the internet, and stable connectivity, because these factors pose challenges for implementation in low- and middle-income countries.

Strengths and Limitations

The main strength of our study is that we conducted an exhaustive review of available literature on the subject and that the main meta-analysis was based on RCTs, which is the most robust design for determining the effect of an intervention. However, our study has several limitations. First, the methodological variability of the included studies led to high heterogeneity in both outcomes, which could affect the interpretation of our findings despite using random effects models for their management. Second, the various measurement tools used in the studies could introduce measurement bias. However, we believe that our study minimized this risk by including only studies that used validated instruments and an effect size that controls for heterogeneity among measures such as the SMD. Third, the lack of clarity in the description of the studied groups may have introduced a risk of bias in assessing their effectiveness because there is no clear taxonomy for grouping NLP-based interventions. Fourth, the global meta-analysis for depressive symptoms identified the potential existence of publication bias, which could overestimate results in favor of trials with positive effects. Therefore, we encourage researchers to report their studies, even if they have negative results, to understand the effect of these interventions better. Fifth, variability in the standards for diagnosing and treating depression and anxiety, as well as in the criteria for determining recovery among the included studies, may have affected the interpretation of the efficacy of the interventions and the generalizability of the findings to different populations. This heterogeneity highlights the importance of considering the context in which NLP-based interventions are applied and the need to adapt them to the characteristics of different populations [11]. Finally, the GRADE assessment shows that the evidence for self-administered NLP-based interventions on depressive and anxiety symptoms is of very low certainty. This suggests caution in interpreting these potential benefits. High risk of bias, significant inconsistency (high I^2 values), and high indirectness complicate the findings. Suspected publication bias further skews the results because studies with nonsignificant or negative outcomes may be underreported. To overcome these limitations in future reviews, we recommend focusing on specific interventions and encouraging researchers to share their primary data to strengthen the quality and reliability of meta-analytic analyses.

Conclusions

Our systematic review and meta-analysis support the use of self-administered interventions based on NLP models to reduce depressive and anxiety symptoms. These findings enhance the theoretical understanding of how advanced NLP tools can effectively deliver psychological therapy, improving cognitive and emotional self-regulation in individuals. By demonstrating the efficacy of various NLP-based interventions, our study advances the theoretical framework by elucidating the mechanisms through which these technologies can replicate and potentially enhance traditional therapeutic processes.

The integration of NLP with different therapeutic modalities offers a novel approach to mental health treatment, expanding the accessibility and scalability of evidence-based interventions. However, the certainty of evidence for the effectiveness of these interventions remains very low, primarily due to a high risk of bias, significant inconsistency, and indirectness in the included studies. Therefore, there is a crucial need for RCTs with larger sample sizes and rigorous methodologies to strengthen the inferential power of future meta-analyses.

Moreover, while our findings are encouraging, there is a need for systematic reviews that examine the implementation processes of these interventions in depth, as well as qualitative studies that evaluate their usability and feasibility. Such research will be essential for effectively recommending the adoption of NLP-based self-administered interventions in public health systems.

Our study provides a valuable starting point for future research to validate the efficacy and practical implementation of these interventions as components of standard mental health care. Ensuring their integration into public health strategies could enhance the mental health outcomes of diverse populations, particularly those who may have limited access to traditional therapeutic resources.

Acknowledgments

This project was in part funded by the National Institutes of Health (R33HL143317). The authors thank Piero Segobia and Carlos García-Navarrete for their collaboration in the initial stages of drafting the manuscript. We used DeepL [68] to translate specific sections of the manuscript and Grammarly [69] to improve the wording of certain sections.

Authors' Contributions

DVZ was responsible for conceptualization, methodology, validation, formal analysis, investigation, data curation, writing the original draft, and visualization. CMRR was responsible for conceptualization, methodology, investigation, writing the original draft, and reviewing and editing the manuscript. JGS was responsible for methodology, validation, investigation, and data curation. GQC, GLC, and RGA were responsible for validation and data curation. GCT was responsible for validation, investigation, writing the original draft, and reviewing and editing the manuscript. ADR was responsible for investigation, writing the original draft, reviewing and editing the manuscript, and supervision. SEA was responsible for investigation, writing the original draft, and reviewing and editing the manuscript. Joseph Finkelstein was responsible for investigation, resources, reviewing and editing the manuscript, and supervision. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[DOCX File , 32 KB - mental_v11i1e59560_app1.docx]

Multimedia Appendix 2
Search strategy.
[DOCX File , 22 KB - mental_v11i1e59560_app2.docx]

Multimedia Appendix 3
Excluded records.
[DOCX File , 25 KB - mental_v11i1e59560_app3.docx]

Multimedia Appendix 4
Included records.
[DOCX File , 21 KB - mental_v11i1e59560_app4.docx]

Multimedia Appendix 5
Funnel plot by depressive and anxiety symptoms.
[DOCX File , 295 KB - mental_v11i1e59560_app5.docx]

Multimedia Appendix 6
Risk of bias for individual studies for the outcome of depressive symptoms.
[DOCX File , 154 KB - mental_v11i1e59560_app6.docx]

Multimedia Appendix 7
Risk of bias for individual studies for the outcome of anxiety symptoms.
[DOCX File , 160 KB - mental_v11i1e59560_app7.docx]

References

1. COVID-19 Mental Disorders Collaborators. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 2021 Nov 06;398(10312):1700-1712 [FREE Full text] [doi: 10.1016/S0140-6736(21)02143-7] [Medline: 34634250]
2. Arias D, Saxena S, Verguet S. Quantifying the global burden of mental disorders and their economic value. EClinicalMedicine 2022 Dec;54:101675 [FREE Full text] [doi: 10.1016/j.eclinm.2022.101675] [Medline: 36193171]
3. Mental health at work. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/mental-health-at-work [accessed 2024-04-05]

4. Edge D, Watkins ER, Limond J, Mugadza J. The efficacy of self-guided internet and mobile-based interventions for preventing anxiety and depression - a systematic review and meta-analysis. Behav Res Ther 2023 May;164:104292 [FREE Full text] [doi: 10.1016/j.brat.2023.104292] [Medline: 37003138]
5. Saad A, Bruno D, Camara B, D'Agostino J, Bolea-Alamanac B. Self-directed technology-based therapeutic methods for adult patients receiving mental health services: systematic review. JMIR Ment Health 2021 Nov 26;8(11):e27404 [FREE Full text] [doi: 10.2196/27404] [Medline: 34842556]
6. Bauer AM, Iles-Shih M, Ghomi RH, Rue T, Grover T, Kincler N, et al. Acceptability of mHealth augmentation of collaborative care: a mixed methods pilot study. Gen Hosp Psychiatry 2018;51:22-29 [FREE Full text] [doi: 10.1016/j.genhosppsych.2017.11.010] [Medline: 29272712]
7. Philippe TJ, Sikder N, Jackson A, Koblanski ME, Liow E, Pilarinos A, et al. Digital health interventions for delivery of mental health care: systematic and comprehensive meta-review. JMIR Ment Health 2022 May 12;9(5):e35159 [FREE Full text] [doi: 10.2196/35159] [Medline: 35551058]
8. Shelton CR, Kotsiou A, Hetzel-Riggin MD. Digital mental health interventions: impact and considerations. In: Lum H, editor. Human Factors Issues and the Impact of Technology on Society. Hershey, PA: IGI Global; 2021.
9. Pineda BS, Mejia R, Qin Y, Martinez J, Delgadillo LG, Muñoz RF. Updated taxonomy of digital mental health interventions: a conceptual framework. Mhealth 2023;9:28 [FREE Full text] [doi: 10.21037/mhealth-23-6] [Medline: 37492117]
10. Cuijpers P, Schuurmans J. Self-help interventions for anxiety disorders: an overview. Curr Psychiatry Rep 2007 Aug;9(4):284-290 [FREE Full text] [doi: 10.1007/s11920-007-0034-6] [Medline: 17880859]
11. Psychological interventions implementation manual: integrating evidence-based psychological interventions into existing services. World Health Organization. URL: https://www.who.int/publications/i/item/9789240087149 [accessed 2024-04-29]
12. Consolidated telemedicine implementation guide. World Health Organization. URL: https://www.who.int/publications-detail-redirect/9789240059184 [accessed 2024-06-07]
13. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. Annu Rev Clin Psychol 2018 May 07;14:91-118. [doi: 10.1146/annurev-clinpsy-032816-045037] [Medline: 29401044]
14. Caldarini G, Jaf S, McGarry K. A literature survey of recent advances in chatbots. Information 2022 Jan 15;13(1):41. [doi: 10.3390/info13010041]
15. Malgaroli M, Hull TD, Zech JM, Althoff T. Natural language processing for mental health interventions: a systematic review and research framework. Transl Psychiatry 2023 Oct 06;13(1):309 [FREE Full text] [doi: 10.1038/s41398-023-02592-2] [Medline: 37798296]
16. Friedman C, Hripcsak G. Natural language processing and its future in medicine. Acad Med 1999 Aug;74(8):890-895. [doi: 10.1097/00001888-199908000-00012] [Medline: 10495728]
17. Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing [review article]. IEEE Comput Intell Mag 2018 Aug;13(3):55-75. [doi: 10.1109/MCI.2018.2840738]
18. Nie J, Shao H, Fan Y, Shao Q, You H, Preindl M, et al. LLM-based conversational AI therapist for daily functioning screening and psychotherapeutic intervention via everyday smart devices. arXiv Preprint posted online on March 16, 2024 [FREE Full text] [doi: 10.1145/3560905.3568074]
19. Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res 2021 May 04;23(5):e15708 [FREE Full text] [doi: 10.2196/15708] [Medline: 33944788]
20. Zhang T, Schoene AM, Ji S, Ananiadou S. Natural language processing applied to mental illness detection: a narrative review. NPJ Digit Med 2022 Apr 08;5(1):46 [FREE Full text] [doi: 10.1038/s41746-022-00589-7] [Medline: 35396451]
21. de Choudhury M, Pendse SR, Kumar N. Benefits and harms of large language models in digital mental health. arXiv Preprint posted online on November 7, 2023 [FREE Full text] [doi: 10.31234/osf.io/y8ax9]
22. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions. London, UK: The Cochrane Collaboration; 2023.
23. Tufanaru C, Munn Z, Aromataris E, Campbell J, Hopp L. Systematic reviews of effectiveness. In: Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, editors. JBI Manual for Evidence Synthesis. Sydney, Australia: Joanna Briggs Institute; 2020.
24. Barker TH, Stone JC, Sears K, Klugar M, Tufanaru C, Leonardi-Bee J, et al. The revised JBI critical appraisal tool for the assessment of risk of bias for randomized controlled trials. JBI Evid Synth 2023 Mar 01;21(3):494-506. [doi: 10.11124/JBIES-22-00430] [Medline: 36727247]
25. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011 Apr;64(4):383-394. [doi: 10.1016/j.jclinepi.2010.04.026] [Medline: 21195583]
26. Ferguson CJ. An effect size primer: a guide for clinicians and researchers. Prof Psychol Res Pract 2009 Oct;40(5):532-538. [doi: 10.1037/a0015808]
27. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634 [FREE Full text] [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]

28. Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 2000 Jun;56(2):455-463. [doi: 10.1111/j.0006-341x.2000.00455.x] [Medline: 10877304]
29. Abdollahi H, Mahoor MH, Zandie R, Siewierski J, Qualls SH. Artificial emotional intelligence in socially assistive robots for older adults: a pilot study. IEEE Trans Affect Comput 2023 Jul 1;14(3):2020-2032 [FREE Full text] [doi: 10.1109/taffc.2022.3143803] [Medline: 37840968]
30. Sabour S, Zhang W, Xiao X, Zhang Y, Zheng Y, Wen J, et al. A chatbot for mental health support: exploring the impact of Emohaa on reducing mental distress in China. Front Digit Health 2023;5:1133987 [FREE Full text] [doi: 10.3389/fdgth.2023.1133987] [Medline: 37214342]
31. Suharwardy S, Ramachandran M, Leonard SA, Gunaseelan A, Lyell DJ, Darcy A, et al. Feasibility and impact of a mental health chatbot on postpartum mental health: a randomized controlled trial. AJOG Glob Rep 2023 Aug;3(3):100165 [FREE Full text] [doi: 10.1016/j.xagr.2023.100165] [Medline: 37560011]
32. Danieli M, Ciulli T, Mousavi SM, Silvestri G, Barbato S, Di Natale L, et al. Assessing the impact of conversational artificial intelligence in the treatment of stress and anxiety in aging adults: randomized controlled trial. JMIR Ment Health 2022 Sep 23;9(9):e38067 [FREE Full text] [doi: 10.2196/38067] [Medline: 36149730]
33. Fitzsimmons-Craft EE, Chan WW, Smith AC, Firebaugh ML, Fowler LA, Topooco N, et al. Effectiveness of a chatbot for eating disorders prevention: a randomized clinical trial. Int J Eat Disord 2022 Mar;55(3):343-353. [doi: 10.1002/eat.23662] [Medline: 35274362]
34. Nicol G, Wang R, Graham S, Dodd S, Garbutt J. Chatbot-delivered cognitive behavioral therapy in adolescents with depression and anxiety during the COVID-19 pandemic: feasibility and acceptability study. JMIR Form Res 2022 Nov 22;6(11):e40242 [FREE Full text] [doi: 10.2196/40242] [Medline: 36413390]
35. He Y, Yang L, Zhu X, Wu B, Zhang S, Qian C, et al. Mental health chatbot for young adults with depressive symptoms during the COVID-19 pandemic: single-blind, three-arm randomized controlled trial. J Med Internet Res 2022 Nov 21;24(11):e40719 [FREE Full text] [doi: 10.2196/40719] [Medline: 36355633]
36. Liu H, Peng H, Song X, Xu C, Zhang M. Using AI chatbots to provide self-help depression interventions for university students: a randomized trial of effectiveness. Internet Interv 2022 Mar;27:100495 [FREE Full text] [doi: 10.1016/j.invent.2022.100495] [Medline: 35059305]
37. Hunt M, Miguez S, Dukas B, Onwude O, White S. Efficacy of Zemedy, a mobile digital therapeutic for the self-management of irritable bowel syndrome: crossover randomized controlled trial. JMIR Mhealth Uhealth 2021 May 20;9(5):e26152 [FREE Full text] [doi: 10.2196/26152] [Medline: 33872182]
38. Klos MC, Escoredo M, Joerin A, Lemos VN, Rauws M, Bunge EL. Artificial intelligence-based chatbot for anxiety and depression in university students: pilot randomized controlled trial. JMIR Form Res 2021 Aug 12;5(8):e20678 [FREE Full text] [doi: 10.2196/20678] [Medline: 34092548]
39. Oh J, Jang S, Kim H, Kim JJ. Efficacy of mobile app-based interactive cognitive behavioral therapy using a chatbot for panic disorder. Int J Med Inform 2020 Aug;140:104171. [doi: 10.1016/j.ijmedinf.2020.104171] [Medline: 32446158]
40. Greer S, Ramo D, Chang YJ, Fu M, Moskowitz J, Haritatos J. Use of the chatbot "Vivibot" to deliver positive psychology skills and promote well-being among young people after cancer treatment: randomized controlled feasibility trial. JMIR Mhealth Uhealth 2019 Oct 31;7(10):e15018 [FREE Full text] [doi: 10.2196/15018] [Medline: 31674920]
41. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. JMIR Ment Health 2018 Dec 13;5(4):e64 [FREE Full text] [doi: 10.2196/mental.9782] [Medline: 30545815]
42. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 06;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
43. Bird T, Mansell W, Wright J, Gaffney H, Tai S. Manage your life online: a web-based randomized controlled trial evaluating the effectiveness of a problem-solving intervention in a student sample. Behav Cogn Psychother 2018 Sep;46(5):570-582 [FREE Full text] [doi: 10.1017/S1352465817000820] [Medline: 29366432]
44. Gaffney H, Mansell W, Edwards R, Wright J. Manage Your Life Online (MYLO): a pilot trial of a conversational computer-based intervention for problem solving in a student sample. Behav Cogn Psychother 2014 Nov;42(6):731-746. [doi: 10.1017/S135246581300060X] [Medline: 23899405]
45. Jang S, Kim JJ, Kim SJ, Hong J, Kim S, Kim E. Mobile app-based chatbot to deliver cognitive behavioral therapy and psychoeducation for adults with attention deficit: a development and feasibility/usability study. Int J Med Inform 2021 Jun;150:104440. [doi: 10.1016/j.ijmedinf.2021.104440] [Medline: 33799055]
46. Maeda E, Miyata A, Boivin J, Nomura K, Kumazawa Y, Shirasawa H, et al. Promoting fertility awareness and preconception health using a chatbot: a randomized controlled trial. Reprod Biomed Online 2020 Dec;41(6):1133-1143. [doi: 10.1016/j.rbmo.2020.09.006] [Medline: 33039321]
47. Prochaska JJ, Vogel EA, Chieng A, Baiocchi M, Maglalang DD, Pajarito S, et al. A randomized controlled trial of a therapeutic relational agent for reducing substance misuse during the COVID-19 pandemic. Drug Alcohol Depend 2021 Oct 01;227:108986 [FREE Full text] [doi: 10.1016/j.drugalcdep.2021.108986] [Medline: 34507061]

48. Bassi G, Giuliano C, Perinelli A, Forti S, Gabrielli S, Salcuni S. A virtual coach (Motibot) for supporting healthy coping strategies among adults with diabetes: proof-of-concept study. JMIR Hum Factors 2022 Jan 21;9(1):e32211 [FREE Full text] [doi: 10.2196/32211] [Medline: 35060918]
49. Prochaska JJ, Vogel EA, Chieng A, Kendra M, Baiocchi M, Pajarito S, et al. A therapeutic relational agent for reducing problematic substance use (Woebot): development and usability study. J Med Internet Res 2021 Mar 23;23(3):e24850 [FREE Full text] [doi: 10.2196/24850] [Medline: 33755028]
50. Lázaro E, Yepez JC, Marín-Maicas P, López-Masés P, Gimeno T, de Paúl S, et al. Efficiency of natural language processing as a tool for analysing quality of life in patients with chronic diseases. A systematic review. Comput Human Behav Rep 2024 May;14:100407. [doi: 10.1016/j.chbr.2024.100407]
51. Calvo RA, Milner DN, Hussain MS, Christensen H. Natural language processing in mental health applications using non-clinical texts. Nat Lang Eng 2017 Jan 30;23(5):649-685. [doi: 10.1017/S1351324916000383]
52. Harvey D, Lobban F, Rayson P, Warner A, Jones S. Natural language processing methods and bipolar disorder: scoping review. JMIR Ment Health 2022 Apr 22;9(4):e35928 [FREE Full text] [doi: 10.2196/35928] [Medline: 35451984]
53. Jabir AI, Martinengo L, Lin X, Torous J, Subramaniam M, Tudor Car L. Evaluating conversational agents for mental health: scoping review of outcomes and outcome measurement instruments. J Med Internet Res 2023 Apr 19;25:e44548 [FREE Full text] [doi: 10.2196/44548] [Medline: 37074762]
54. Treviso M, Lee JU, Ji T, Aken BV, Cao Q, Ciosici M, et al. Efficient methods for natural language processing: a survey. Trans Assoc Comput Linguist 2023;11:826-860 [FREE Full text] [doi: 10.1162/tacl_a_00577]
55. Lin T, Wang Y, Liu X, Qiu X. A survey of transformers. AI Open 2022;3:111-132. [doi: 10.1016/j.aiopen.2022.10.001]
56. Khan W, Daud A, Khan K, Muhammad S, Haq R. Exploring the frontiers of deep learning and natural language processing: a comprehensive overview of key challenges and emerging trends. Nat Lang Process J 2023 Sep;4:100026. [doi: 10.1016/j.nlp.2023.100026]
57. Chin H, Song H, Baek G, Shin M, Jung C, Cha M, et al. The potential of chatbots for emotional support and promoting mental well-being in different cultures: mixed methods study. J Med Internet Res 2023 Oct 20;25:e51712 [FREE Full text] [doi: 10.2196/51712] [Medline: 37862063]
58. Oduoye MO, Fatima E, Muzammil MA, Dave T, Irfan H, Fariha FN, et al. Impacts of the advancement in artificial intelligence on laboratory medicine in low- and middle-income countries: challenges and recommendations-A literature review. Health Sci Rep 2024 Jan;7(1):e1794 [FREE Full text] [doi: 10.1002/hsr2.1794] [Medline: 38186931]
59. Alami H, Rivard L, Lehoux P, Hoffman SJ, Cadeddu SB, Savoldelli M, et al. Artificial intelligence in health care: laying the foundation for responsible, sustainable, and inclusive innovation in low- and middle-income countries. Global Health 2020 Jun 24;16(1):52 [FREE Full text] [doi: 10.1186/s12992-020-00584-1] [Medline: 32580741]
60. Babel A, Taneja R, Mondello Malvestiti F, Monaco A, Donde S. Artificial intelligence solutions to increase medication adherence in patients with non-communicable diseases. Front Digit Health 2021;3:669869 [FREE Full text] [doi: 10.3389/fdgth.2021.669869] [Medline: 34713142]
61. Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim HC, et al. Artificial intelligence for mental health and mental illnesses: an overview. Curr Psychiatry Rep 2019 Nov 07;21(11):116 [FREE Full text] [doi: 10.1007/s11920-019-1094-0] [Medline: 31701320]
62. Boucher EM, Harake NR, Ward HE, Stoeckl SE, Vargas J, Minkel J, et al. Artificially intelligent chatbots in digital mental health interventions: a review. Expert Rev Med Devices 2021 Dec;18(sup1):37-49. [doi: 10.1080/17434440.2021.2013200] [Medline: 34872429]
63. Pham KT, Nabizadeh A, Selek S. Artificial intelligence and chatbots in psychiatry. Psychiatr Q 2022 Mar;93(1):249-253 [FREE Full text] [doi: 10.1007/s11126-022-09973-8] [Medline: 35212940]
64. Dawoodbhoy FM, Delaney J, Cecula P, Yu J, Peacock I, Tan J, et al. AI in patient flow: applications of artificial intelligence to improve patient flow in NHS acute mental health inpatient units. Heliyon 2021 May;7(5):e06993 [FREE Full text] [doi: 10.1016/j.heliyon.2021.e06993] [Medline: 34036191]
65. Nilsen P, Svedberg P, Neher M, Nair M, Larsson I, Petersson L, et al. A framework to guide implementation of AI in health care: protocol for a cocreation research project. JMIR Res Protoc 2023 Nov 08;12:e50216 [FREE Full text] [doi: 10.2196/50216] [Medline: 37938896]
66. Thakkar A, Gupta A, de Sousa A. Artificial intelligence in positive mental health: a narrative review. Front Digit Health 2024;6:1280235 [FREE Full text] [doi: 10.3389/fdgth.2024.1280235] [Medline: 38562663]
67. Chidambaram S, Jain B, Jain U, Mwavu R, Baru R, Thomas B, et al. An introduction to digital determinants of health. PLOS Digit Health 2024 Jan;3(1):e0000346 [FREE Full text] [doi: 10.1371/journal.pdig.0000346] [Medline: 38175828]
68. Type to translate. DeepL. URL: https://www.deepl.com/en/translator [accessed 2024-04-29]
69. Responsible AI that ensures your writing and reputation shine. Grammarly. URL: https://www.grammarly.com/ [accessed 2024-06-29]

Abbreviations

AI: artificial intelligence

GAD-7: Generalized Anxiety Disorder-7
GRADE: Grading of Recommendations Assessment, Development, and Evaluation
ML: machine learning
NLP: natural language processing
PHQ: Patient Health Questionnaire
PICOS: Population, Intervention, Comparison, Outcomes, and Study Design
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SMD: standardized mean difference

Edited by J Torous; submitted 15.04.24; peer-reviewed by Y Liu, S Zhang; comments to author 17.05.24; revised version received 12.06.24; accepted 02.07.24; published 21.08.24.

Please cite as:
Villarreal-Zegarra D, Reategui-Rivera CM, García-Serna J, Quispe-Callo G, Lázaro-Cruz G, Centeno-Terrazas G, Galvez-Arevalo R, Escobar-Agreda S, Dominguez-Rodriguez A, Finkelstein J
Self-Administered Interventions Based on Natural Language Processing Models for Reducing Depressive and Anxiety Symptoms: Systematic Review and Meta-Analysis
JMIR Ment Health 2024;11:e59560
URL: https://mental.jmir.org/2024/1/e59560
doi:10.2196/59560
PMID:

Review

Machine Learning, Deep Learning, and Data Preprocessing Techniques for Detecting, Predicting, and Monitoring Stress and Stress-Related Mental Disorders: Scoping Review

Moein Razavi[1,2], BSc, ME, PhD; Samira Ziyadidegan[1], BSc, MSc, PhD; Ahmadreza Mahmoudzadeh[3], BSc, MSc, MS, PhD; Saber Kazeminasab[4], BSc, MSc, PhD; Elaheh Baharlouei[5], BSc, MSc, PhD; Vahid Janfaza[2], BSc, MSc, PhD; Reza Jahromi[1,2], BSc, MSc, MCS, PhD; Farzan Sasangohar[1], BA, BCS, MASc, SM, PhD

[1]Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX, United States
[2]Department of Computer Science and Engineering, Texas A&M University, College Station, TX, United States
[3]Zachry Department of Civil and Environmental Engineering, Texas A&M University, College Station, TX, United States
[4]Harvard Medical School, Harvard University, Boston, MA, United States
[5]Department of Computer Science, University of Houston, Houston, TX, United States

Corresponding Author:
Farzan Sasangohar, BA, BCS, MASc, SM, PhD
Department of Industrial and Systems Engineering
Texas A&M University
3131 TAMU
College Station, TX, 77843-3131
United States
Phone: 1 979 458 2337
Email: sasangohar@tamu.edu

Abstract

Background: Mental stress and its consequent mental health disorders (MDs) constitute a significant public health issue. With the advent of machine learning (ML), there is potential to harness computational techniques for better understanding and addressing mental stress and MDs. This comprehensive review seeks to elucidate the current ML methodologies used in this domain to pave the way for enhanced detection, prediction, and analysis of mental stress and its subsequent MDs.

Objective: This review aims to investigate the scope of ML methodologies used in the detection, prediction, and analysis of mental stress and its consequent MDs.

Methods: Using a rigorous scoping review process with PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines, this investigation delves into the latest ML algorithms, preprocessing techniques, and data types used in the context of stress and stress-related MDs.

Results: A total of 98 peer-reviewed publications were examined for this review. The findings highlight that support vector machine, neural network, and random forest models consistently exhibited superior accuracy and robustness among all ML algorithms examined. Physiological parameters such as heart rate measurements and skin response are prevalently used as stress predictors due to their rich explanatory information concerning stress and stress-related MDs, as well as the relative ease of data acquisition. The application of dimensionality reduction techniques, including mappings, feature selection, filtering, and noise reduction, is frequently observed as a crucial step preceding the training of ML algorithms.

Conclusions: The synthesis of this review identified significant research gaps and outlines future directions for the field. These encompass areas such as model interpretability, model personalization, the incorporation of naturalistic settings, and real-time processing capabilities for the detection and prediction of stress and stress-related MDs.

(JMIR Ment Health 2024;11:e53714) doi:10.2196/53714

KEYWORDS

machine learning; deep learning; data preprocessing; stress detection; stress prediction; stress monitoring; mental disorders

Introduction

Background

Mental health has become a public health concern. According to the Institute of Health Metrics and Evaluation, in 2019, about 53 million people in the United States and about 1 in 8 individuals worldwide (about 1 billion people) have at least 1 mental health disorder (MD) [1]. An MD is defined as an impairment in a person's cognition, emotional control, or behavior patterns that has clinical significance and is often linked to distress or functional impairment [2]. MDs severely limit people's daily functioning and can be fatal [3,4]. In 2019, mental health (MH) problems accounted for 6.6% of all disability-adjusted life years in the United States, making it the fifth most significant cause of disability overall [1,5].

Some of the more prevalent MDs are anxiety disorders, depression or mood disorders, bipolar disorders, psychotic disorders (including schizophrenia), eating disorders, social disorders, disruptive behavior, and addictive behaviors [2]. In 2019, anxiety and depression have been the most prevalent forms of MDs (301 and 280 million people affected worldwide, respectively). Anxiety disorder encompasses emotions of concern, anxiety, excessive fear, or associated behavioral problems that are severe enough to affect everyday activities [2]. Symptoms include an unproportionate level of stress compared to the significance of the triggering event, difficulty in putting worries out of one's mind, and nervousness [6,7]. Generalized anxiety disorder, panic attacks, social anxiety disorder, and posttraumatic stress disorder (PTSD) are all examples of different types of anxiety disorders [2,8]. Depression is characterized by a long-lasting sadness and a lack of desire to be active. One of the main symptoms of depression is the inability to enjoy or find pleasure in most of one's daily activities as well as feeling sadness, anger, or emptiness [2,9]. A depressive episode typically lasts for at least 2 weeks. In addition, a loss of self-worth, feelings of hopelessness for the future, and suicidal thoughts are indicators and symptoms of depression. People who are depressed are more prone to commit suicide [2,9,10].

Stress is categorized into distress, which typically has chronic negative effects on health, and eustress, which is short-term and positively influences motivation and development [11]. Throughout this paper, the term stress is specifically used to denote distress rather than eustress. Mental stress has been shown to significantly contribute to developing and worsening anxiety and depression disorders [12-14]. Mental stress is the body's natural response to various events in which a person feels that the demands of their external environment exceed their psychological and physiological resources for dealing with those demands [15]. Mental stress leads to an asynchrony between the sympathetic and parasympathetic nervous systems (SNS and PNS), which are the main divisions of the autonomic nervous system [16] and serve an important role in regulating vital biological activities [17,18]. The SNS is an integrative system that responds to potentially dangerous circumstances. Activation of the SNS is part of the system responsible for controlling "fight-or-flight" responses. The PNS is responsible for the body's "rest-and-digest" processes.

Given the important role and impact of stress in MDs, previous research has investigated various qualitative and quantitative methods to measure and monitor stress to inform effective stress mitigation approaches. While majority of stress literature relies on self-reported measures, recent literature has used physiological variables such as heart rate (HR); HR variability (HRV) [19-23]; and behavioral data (eg, speech, movement, and facial expressions) [24] to understand changes to SNS and PNS associated with stress. The recent advances in sensor and mobile health technologies have resulted in the emergence of big data related to MH, as well as advanced bioinformatics methods, tools, or techniques to use such data for modeling or inference. One such tool that has recently emerged as a robust, rapid, objective, reliable, and cost-efficient technique for studying chronic illnesses and MDs is machine learning (ML). ML uses advanced statistical and probabilistic techniques to construct systems that can automatically learn from data. Several characteristics of ML make it suitable for applications in MH monitoring including significant pattern recognition and forecasting capabilities [25], the capacity to extract crucial information from various data resources and the opportunity to create personalized experiences [25], and the ability to analyze large amounts of data in a short time [26]. As such, ML has gained popularity and has been applied to MH data to enable detection, monitoring, and treatment [27]. The objective of this research is to review the literature to summarize and synthesize the application of ML in the detection, monitoring, or prediction of stress and stress-related MDs, in particular, anxiety and depression. This paper documents method-specific findings such as data types, preprocessing methods, and different algorithms used, as well as the type and characteristics of studies that used ML.

Traditional statistical methods, such as linear regression, logistic regression (LR), 1- or 2-tailed *t* tests, and ANOVA [28], have been widely used in the past to detect and analyze stress and stress-related MDs. These methods have proven useful in specific contexts, such as comparing means of different groups or modeling linear relationships between variables. As demonstrated by Machado et al [21], Adjei et al [22], Yoo et al [23], Chen et al [24], and Jordan and Mitchell [25], these methods have provided valuable insights in situations wherein the data are relatively simple and adhere to the underlying assumptions of the statistical techniques. However, when faced with complex, high-dimensional MH data, which have become increasingly available, thanks to advancements in technology and data collection techniques, these traditional statistical methods might not be sufficient. The limitations of these methods stem from their inherent simplicity and the assumptions they rely on, which might not hold true in the context of MH data. For example, linear regression and LR analyses assume linear relationships between variables, while *t* tests and ANOVA require specific assumptions about the data distribution. These assumptions may not be applicable in the case of intricate and heterogeneous MH data, potentially leading to inaccurate or incomplete conclusions.

Advanced data analytics methods, such as ML, offer a more powerful and flexible alternative to traditional statistical methods. ML algorithms, with their significant pattern recognition and forecasting capabilities [25], are capable of capturing complex, nonlinear relationships between variables and can adapt to various data distributions. These capabilities enable ML techniques to provide more accurate and insightful predictions, classifications, and associations in the context of MH data [29]. In addition, ML algorithms can handle large-scale, high-dimensional data more efficiently than traditional methods, allowing researchers to analyze vast amounts of information from diverse sources, such as eHealth records, wearable devices, and web-based platforms [26]. This capacity for handling big data is crucial for understanding the multifaceted nature of MDs and developing tailored interventions. ML techniques also offer the advantage of automation and adaptability, allowing them to continuously learn and improve as new data become available [25]. This iterative learning process enables the development of more sophisticated and accurate models for detecting, monitoring, and predicting stress and stress-related MDs over time.

While traditional statistical methods have contributed significantly to our understanding of stress and stress-related MDs in specific contexts, the growing complexity and volume of MH data necessitate the adoption of advanced data analytics methods such as ML. By leveraging the power of ML, researchers can gain deeper insights into the underlying patterns and relationships between stress and MDs [29], ultimately leading to the development of more effective stress mitigation approaches and improved care for individuals who have anxiety, depression, and other MDs.

Acknowledging the substantial contributions of traditional statistical methods, it becomes evident that the escalating complexity and scale of MH data demand the adoption of more sophisticated approaches such as ML. This advancement stands not as a replacement but as an essential evolution in the analytical toolbox available to researchers. As this paper delves into the myriad ways that ML has been applied to MH, particularly in the realms of stress, anxiety, and depression, it seeks to consolidate the current knowledge on the subject.

Objectives

By examining the types of data, preprocessing methods, and the algorithms used in existing studies, this review aspires to offer a detailed synthesis of the field. It aims to provide a clearer understanding of ML's effectiveness in the detection, monitoring, and prediction of MDs, setting a foundation for future research and the enhancement of therapeutic strategies for those impacted by these conditions.

Methods

Protocol and Registration

This scoping review adhered to the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [30]. No formal review protocol was registered due to the exploratory nature of this study, which aimed to map out existing research rather than address a prespecified hypothesis. This approach aligns with the methodological flexibility often required in emergent areas of research.

Eligibility Criteria

We included studies published in English from 2017 to 2022 that used ML techniques to evaluate MDs, specifically focusing on stress and stress-related conditions. Studies were excluded if they did not use ML as the primary analysis method or if they were published in languages other than English.

Information Sources

The literature search involved databases such as EI Engineering Village, Web of Science, ACM Digital Library, and IEEE Xplore. Additional sources were identified through contact with experts and review of references in relevant articles.

Search Strategy

A comprehensive search was conducted using a combination of keywords related to ML and MDs (Textbox 1). The search strategy was designed to capture a broad spectrum of ML applications within this field. The full search list from all databases is available in Multimedia Appendix 1 [31-50].

Textbox 1. Keywords and search strategy for articles since 2017 (last 5 years).

Search strategy

First keyword: predict OR detect

AND

Second keyword: mental health OR mental disorder OR depression OR anxiety OR stress

AND

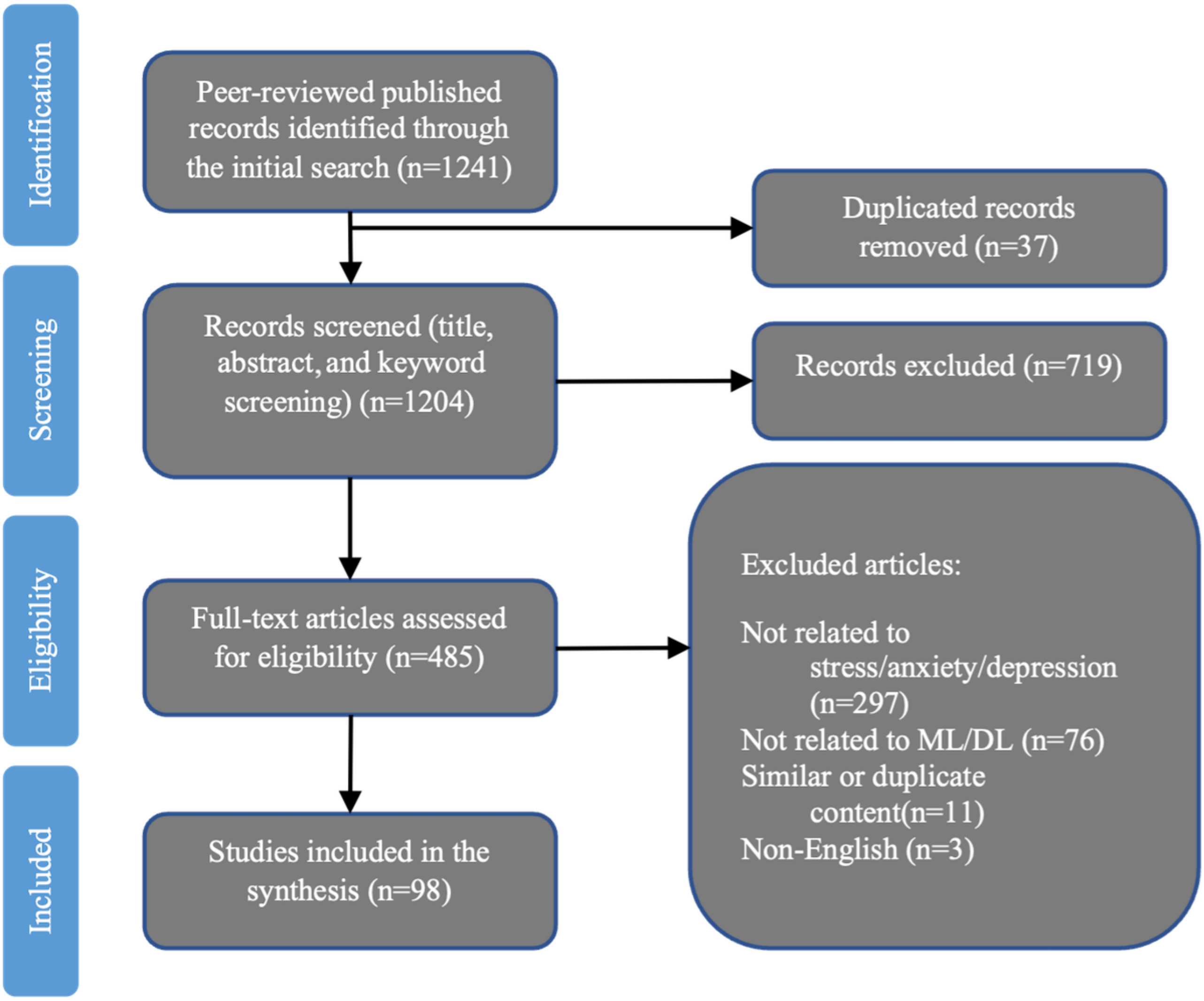
Third keyword: machine learning OR deep learning OR data mining OR pattern classification OR artificial intelligence OR neural networks

Study Selection, Inclusion, and Exclusion Criteria

Articles that did not fully use ML for stress or stress-related MDs evaluations were excluded from the research. Studies published in languages other than English were also excluded.

The initial search yielded 1241 results. After duplicate articles were deleted and eligibility was confirmed using Rayyan (Qatar Computing Research Institute) [51], 1204 (97.02%) articles remained. After applying the exclusion criteria, 98 (8.14%) papers were selected for full review (Figure 1).

Figure 1. Preferred items for scoping literature review and meta-analysis flowchart (modified from Tricco et al [30]). DL: deep learning. ML: machine learning.



Data Charting Process

Data charting was conducted by 2 reviewers independently using a standardized form, which had been pretested on a subset of included studies. Discrepancies were resolved through discussion or consultation with a third reviewer. Study authors were contacted for clarification or additional data where necessary.

Data Items

Data extracted included publication year, study design, population characteristics, ML techniques used, outcomes measured, and key findings. Other variables sought included data preprocessing methods and performance metrics of the ML models. Simplifying assumptions, such as considering different ML algorithms within the same family as a single technique, were made to facilitate synthesis.

Synthesis of Results

Data were synthesized descriptively, and the findings were grouped by ML techniques, data type, and preprocessing techniques. Where possible, quantitative performance metrics were extracted or derived. Results were analyzed in the context of the overall study designs and populations to highlight trends and identify gaps in the current research landscape. No formal critical appraisal or quantitative meta-analysis was conducted due to the diversity of the included studies and the scoping nature of this review.

Results

In this section, types of data, preprosessing techniques, and ML techniques used on the data in the literature have been reviewed and compared with the existing literature.

Types of Data

Overview

Various data types were used in the studies (n=98) that used ML algorithms for stress and stress-related MDs. Studies used questionnaires (n=31, 32%); HRV (n=25, 26%); skin response (eg, skin temperature, skin conductance, etc; n=24, 24%); photoplethysmogram (PPG; n=21, 21%); electrocardiogram (ECG; n=19, 19%); HR (n=17, 17%); electroencephalogram (EEG; n=9, 9%); acceleration or body movement (n=8, 8%); text data (n=7, 7%); respiratory signals (n=7, 7%); electromyogram (EMG; n=3, 3%); eye tracking (n=3, 3%); speech signals (n=3, 3%); and others (n=4, 4%) including audio signals (n=2, 2%), blood pressure (BP; n=1, 1%), and hormones (n=1, 1%). Table 1 shows the distribution of the type of data used for stress detection using ML techniques.

Table 1. Number of articles by type of data (n=98).

Type of data	Articles, n (%)
Questionnaire	31 (32)
HRV[a]	25 (25)
Skin response	24 (24)
HR[b]	17 (17)
EEG[c]	9 (9)
Body movement	8 (8)
Respiratory	7 (7)
Text	7 (7)
EMG[d]	3 (3)
Eye-tracking	3 (3)
Speech signals	3 (3)

[a]HRV: heart rate variability.

[b]HR: heart rate.

[c]EEG: electroencephalogram.

[d]EMG: electromyogram.

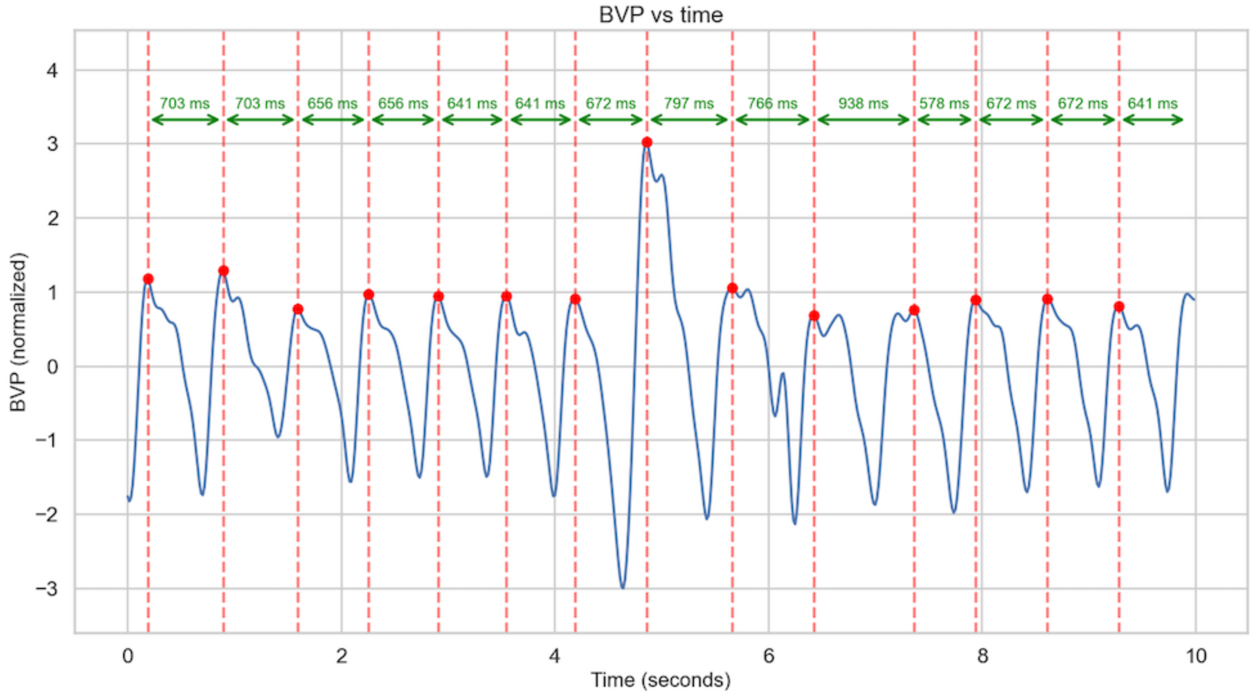
Heart Measures

Heart metrics are primarily used for stress detection and are typically gathered through 2 main methods: ECG and PPG. ECG is a noninvasive diagnostic test that records the heart's electrical activity, while PPG is a noninvasive optical technique that detects changes in blood volume within the tissue's microvascular bed. By using these methods, it is possible to measure various heart-related parameters, including HR, as well as time and frequency domain features of HRV and BP.

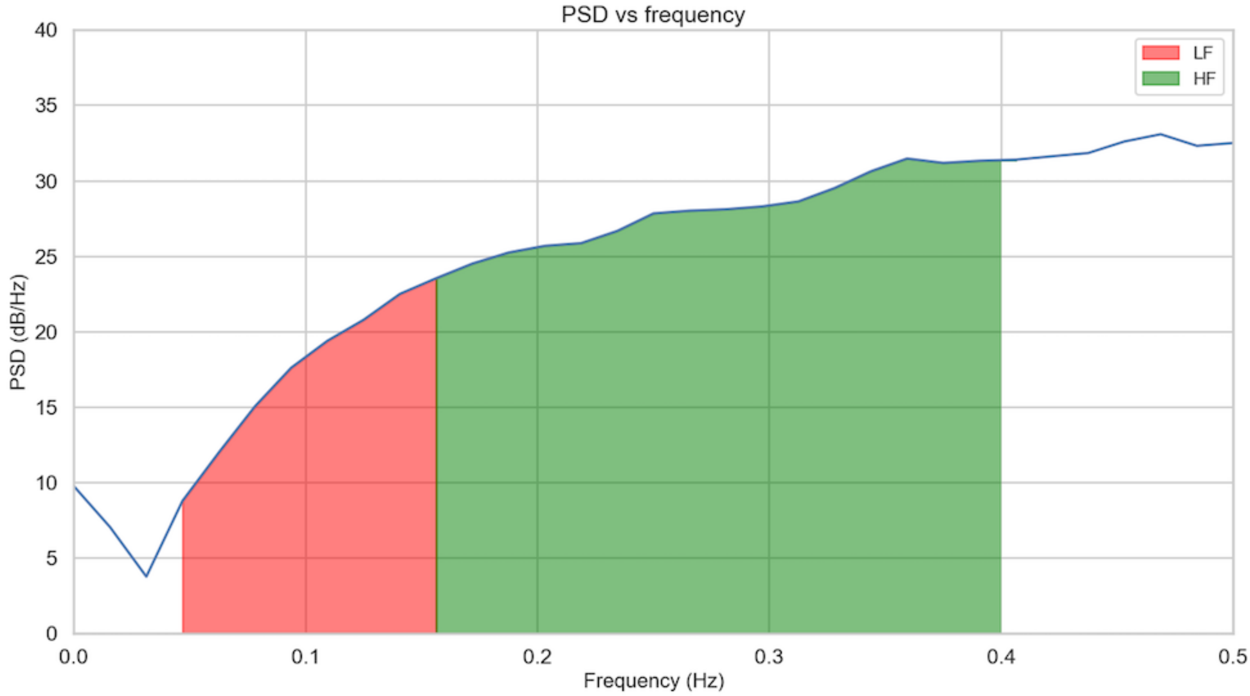
HRV Measures

HRV (n=25, 26%) has been used to assess MH issues, such as stress, anxiety, and depression, due to its rich time and frequency domain features [53]. The blood volume pulse signal is another effective method for capturing HRV features, as it represents the heart's beat-to-beat volume changes. From the blood volume pulse signal, time-domain measures such as the root mean square of successive RR interval differences (RMSSD), SD of neural network (NN) intervals, and SD of RR intervals can be derived. In addition, the frequency domain aspects of HRV, including total power (frequencies <0.4 Hz), low frequencies (LFs; ranging from 0.04 to 0.15 Hz), and high frequencies (HFs; between 0.15 and 0.4 Hz), reflect the autonomic nervous system's dynamics during beat-to-beat measurements of the HR (Figure 2) [54,55]. These HRV measures, both in the time and frequency domains, provide a nuanced view of the physiological underpinnings associated with various MH conditions.

Figure 2. Depiction of heart’s beat-to-beat measurements using blood volume pulse (BVP) signal (A) and power spectral density (PSD) of RR intervals (the signal is bandpass filtered with cutoff frequencies of 0.04 Hz and 0.4 Hz) (B). HF: high frequency; LF: low frequency.



(A)



(B)

HR Measures

One of the most important indicators of stress is an abrupt increase in HR (n=17, 17%). Among the physiological signals, HR is among the top measures that explain stress in ML models, and it has been used in different studies with almost all ML algorithms [56-58].

BP Measures

BP (n=1, 1%) can be obtained by pulse transit time or by pressure cuffs [59]. Stressful conditions create an influx of hormones that increase HR and constrict blood vessels, leading to a temporary BP elevation [60]. In most cases, BP recovers to its prestress level after the stress response diminishes [61]. Schultebraucks et al [62] used systolic BP as one of the measures in predicting one’s level of susceptibility to PTSD.

EEG Data

EEG (n=9, 9%) detects brain electrical activity. Compared with other brain mapping techniques for stress detection, it is more practical due to several factors including affordability, noninvasiveness, nonintrusiveness, and, most importantly, its high temporal resolution [63]. The high temporal resolution of EEG makes it appropriate for real-time stress detection, as well as deep learning (DL) approaches, which require large data sets for training [63-67].

The most commonly used EEG features for the detection of stress are the power of different frequency bands (α, 8-13 Hz; β, 12.5-30.0 Hz; θ, 4.5-7.5 Hz; γ, 30-40 Hz), average and SD of a specific time window of EEG signal, and time-frequency features obtained by discrete wavelet transform algorithm [67-69]. It has also been shown that statistical features of EEG signals, such as kurtosis and entropy, are useful features in stress prediction using ML algorithms [66]. Moreover, power spectral density (PSD), correlation, divisional asymmetry, rational asymmetry, and power spectrum are other EEG features that have been used in different studies for stress detection [70].

Since EEG signals are collected from the scalp, they include excessive noise and have high uncertainty. Therefore, signal processing and feature selection or extraction are important steps while dealing with EEG data. Several well-developed methods are available for treating the EEG data. Among them, latent space derived from auto-encoders and signal reconstruction techniques such as artifact subspace reconstruction (ASR) are well-known methods that can be applied on EEG data to significantly reduce the artifacts [65]. These methods are also fast enough to make real-time detection feasible.

The amygdala and hippocampus are the parts of the brain that have the major responsibility for human reactions to stress [71]. Brain activity caused by stress in those regions would affect the prefrontal cortex. Studies collecting data from the prefrontal cortex have also verified that EEG data from this brain region can be used for stress detection [72]. EEG can be collected from the prefrontal cortex using off-the-shelf EEG recording products such as Muse and Neurosky Mindwave [66,69,70,72].

Eye Tracking

Eye-tracking features (n=3, 3%) can be indicators of stress. For example, to diagnose the level of stress, the changes in the striations of muscle material in the iris as a response to stress can be used as features for ML algorithms. In other words, pupil diameter, which is controlled by iris sphincter muscles, can be used as a feature [73]. Other eye-tracking features that have been used for stress detection are visual fixations, saccade movements, pupil size, microsaccades, and the number of eye blinks in a specific time window during a certain task [74-76].

Skin Response

A skin response (n=24, 24%) can be defined as a stimulus-regulated electrodermal response and is typically measured using electrodes placed on the fingertips or hands. Skin response is usually associated with an increase in sympathetic activity on inducing stress events [77]. The skin becomes a better conductor of electricity when it is stimulated either externally or internally by physiologically stimulating factors, including stressful conditions [78].

Respiratory Signals

Mental stress can affect different respiratory cycle phases and breathing patterns (n=7, 7%) [79,80]. For example, it has been discovered that stress had no impact on overall breath duration (respiration rate) but that exhalation periods were longer and pause periods were shorter in the stress experiment than in the neutral condition [81].

On the basis of the findings of several studies, it can be concluded that respiratory signal is one of the top contributing factors in the explanation of stress in ML models. The most common time-domain respiratory signal features that are extracted for stress detection are root mean square, IQR, and mean square differences between adjacent elements of breathing rate and blood oxygenation levels. The most commonly used frequency domain features of the respiratory signal are the power of LFs (<2 Hz), the power of HFs (>2 Hz), and the ratio of the power of LFs over the power of HFs (LF and HF) [58,62,63,82-84].

EMG Data

EMG (n=3, 3%) detects the electrical activity of muscles at rest, during a modest contraction, and during a strong contraction [85]. Similar to acceleration data, several studies have shown that using EMG data can help increase the performance of ML models trained on ECG data. The action potential intrigued in the EMG during stress can reduce the variance for decision-making of classification models that use ECG [58,86,87].

Hormones

It has been shown that stress can alter the levels of glucocorticoids, catecholamines, growth hormones, and prolactin in the bloodstream. Therefore, in ML models, levels of hormones such as cortisol, dehydroepiandrosterone sulfate, thyroid-stimulating hormone, free triiodothyronine, and free thyroxine can be used as predictors for the detection of stress-related disorders (n=1, 1%) [62].

Acceleration and Body Movement

Mental stress may cause a wide variety of behavioral and body movement symptoms such as shaking hands and feet, which can be measured by the acceleration data (n=8, 8%) [88]. Moreover, research has shown that people with a greater stress score had less variance in their activity level and body movements [89-91]. For example, in older adults, stressful life events can be related to a reduced rate of regular physical exercise [92]. Time and frequency features such as mean absolute deviation from the mean, the total power of acceleration, SD, the mean norm of acceleration, absolute integral, and peak frequency of each axis are the features of hand and body acceleration used for stress detection [31,57,93]. One practical characteristic of motion and acceleration data is that they can be used to identify sources of noise in other signals. For example, motion data can help distinguish stress from physical activity (eg, exercise) when other physiological measures, such as ECG, have uncertainty in prediction [94,95].

Audio and Speech Signals

Speech Signals

Using speech signals (n=3, 3%), it is feasible to diagnose and assess neurological disorders and MDs [96]. Moreover, studies have shown that like body acceleration and EMG, features of speech signals can make stress predictions of heart measurements more robust. The best explanatory parameters of speech signals are frequency domain parameters (eg, PSD, strongest frequency from fast Fourier transform) and time-frequency features such as Mel-frequency cepstral coefficient [56,97,98]. Since time-frequency measures are 2-dimensional measurements with a high number of samples, they make this signal suitable for use in convolutional NN (CNN) models of stress and depression detection [99].

Audio Signals

For laboratory-based studies, audio signals (eg, beeping sounds; n=2, 2%) can be used to stimulate stress events in participants [100,101].

Text Data

Social media content (n=7, 7%) is frequently subjected to reviews, opinions, and influence, as well as sentiment analysis. Natural language processing methods may be used to evaluate social networking posts and comments for mood and emotion to detect whether a user is stressed [102-108].

Questionnaire

Different questionnaires (n=31, 32%) are used for the diagnosis of stress and MDs including anxiety and depression. The scores from different items on these questionnaires can be used as dependent and independent variables in ML studies. The questionnaires mentioned here were selected based on their prevalence in the literature as well as their relevance to the ML outcomes being predicted. For instance, some studies have successfully leveraged scores from multiple questionnaires, such as the Diagnostic and Statistical Manual of Mental Disorders, Depression Anxiety and Stress Scale, Edinburgh Perinatal/Postnatal Depression Scale, Center for Epidemiological Studies-Depression survey, Mean Opinion Score, Hamilton Depression Rating Scale, State-Trait Anxiety Inventory, Posttraumatic Stress Disorder Checklist for Diagnostic and Statistical Manual of Mental Disorders, Beck Depression Inventory, Beck Anxiety Inventory, Hospital Anxiety and Depression Scale, Goldberg's Depression Scale, self-reports, and clinician reports [109-126].

Preprocessing Techniques

In this section, important preprocessing techniques that have yielded significant findings and how they are used to help the detection of stress and its related MDs have been reviewed.

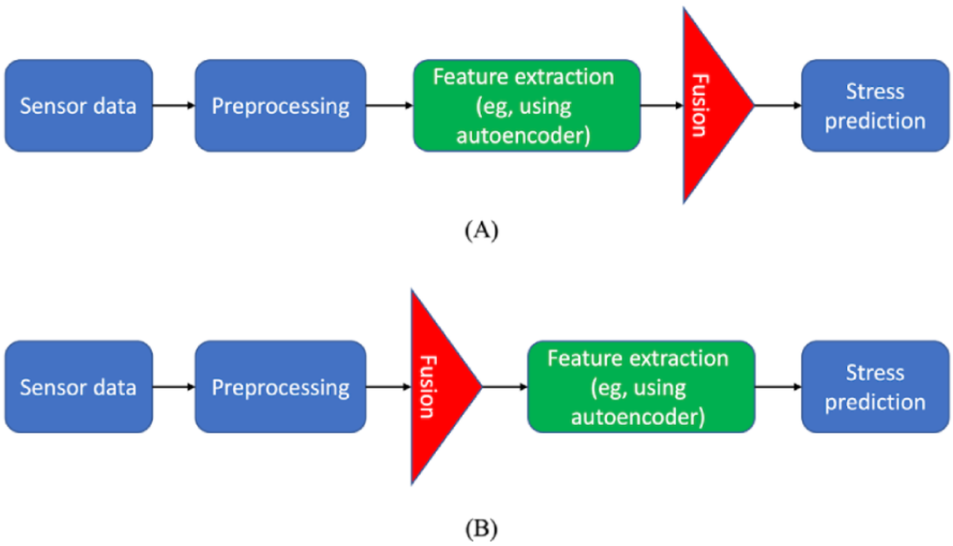
Synthetic Minority Oversampling Technique

In the detection of stress and its related MDs, the number of samples for the stress or MD class is usually significantly lower than the nonstress or non-MD class. This imbalance in the number of samples for each class leads to a bias in prediction (toward the majority class). To correct for data bias, it is possible to oversample the underrepresented group. In stress detection studies using ML models, the synthetic minority oversampling technique is one of the most common approaches to boost the minority class, which creates new samples by synthesizing those already available in the data (by combining their features; n=3, 3%) [31,110,127].

Early Modality Fusion

In ML models used for the prediction of stress with a multimodal approach, it has been shown that early fusion of multimodal data before feature extraction is more effective and achieves a better performance (n=1, 1%). This is because early modality fusion better catches the important characteristics that are coherent with each other. For example, a study showed that combining different measures including skin response, skin temperature, and body acceleration before feature extraction outperforms the approach that extracts the features for each measure separately and combines them afterward (Figure 3) [128].

Figure 3. Early modality fusion (A) and late modality fusion (B).



PSD Method

In physiological signals for stress detection, usually the power of the signal changes during the moments of stress. PSD (n=13, 13%) explains the frequency-based power distribution of a time series and reveals the locations of strong and weak frequency variation. Welch's method is one of the most common approaches for calculating PSD [65]. PSD is often used in studies that include frequency domain HRV features for stress detection such as total HF or LF power [82-84,101,129-136].

ILIOU Method

In the detection of MDs, such as depression and anxiety, using ML techniques, having the least error rate is significantly important so that the person can take further actions appropriately. In this matter, the data preprocessing step has an important role in minimizing the noise and bias toward the false prediction. Iliou et al [114] proposed ILIOU (n=1, 1%), a data mapping and transformation method that identifies useful information for the detection of MDs, especially for depression. This method outperforms common data preprocessing techniques such as principal component analysis (PCA), evolutionary search algorithm, and isomap for the detection of depression.

PCA Method

PCA (n=3, 3%) is a method for lowering the dimensionality of such data sets while maximizing interpretability and minimizing loss of information. It does this by generating new variables that are uncorrelated and progressively optimize variance [58,90,123].

Independent Component Analysis

Independent component analysis (ICA; n=4, 4%) is a computational and statistical method for uncovering hidden elements underlying random variables, observations, or signals. This method is mostly used for removing artifacts from stationary signal noises of the multichannel data. ICA optimizes higher-order statistics such as kurtosis, while PCA optimizes the covariance matrix of the data, which reflects second-order statistics. In stress detection using physiological signals that contain stationary noises (eg, eyeblink noise in EEG) it is recommended to remove noises using ICA [63-65,67].

ASR Approach

ASR (n=1, 1%) is an adaptive approach for removing artifacts from signal recordings on the web or offline, mostly nonstationary signal noises. To identify artifacts based on their statistical qualities in the component subspace, a PCA on covariance matrices is repeatedly computed [137]. Since there are usually substantial nonstationary noises in the EEG data, in order to classify stress at multiple levels using EEG data, using ASR before classification is highly recommended [65].

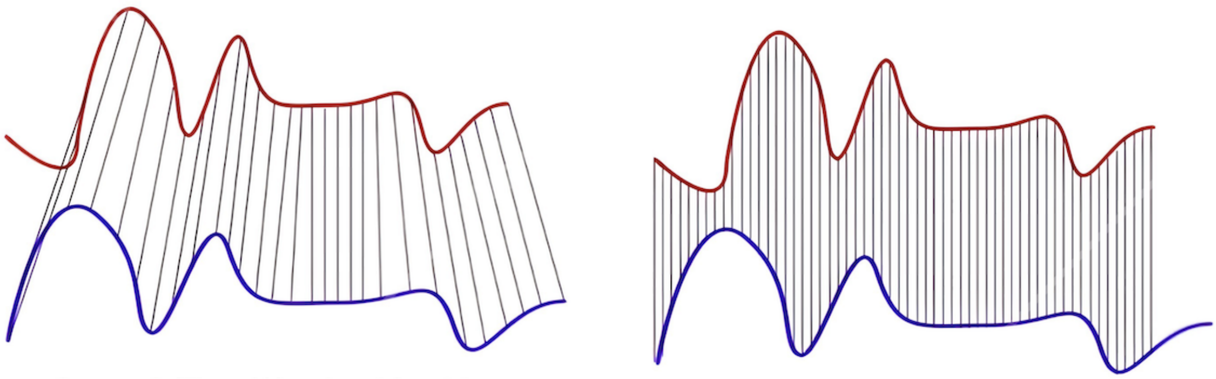
Latent Growth Mixture Modeling

Growth mixture modeling (n=1, 1%) is to discover numerous hidden subpopulations, describe longitudinal development within each hidden subpopulation, and investigate variation in hidden subpopulations' rates of change. Latent growth mixture models are gaining popularity as a statistical tool for estimating individual development over time and for probing the presence of latent trajectories, in which people belong to the trajectories that are not directly observable [62,138,139].

Dynamic Time Warping

It is common practice to transform data from 2 time series into vectors and then compute the Euclidean distance between the resulting points in vector space to determine the degree of similarity or dissimilarity between the series, regardless of whether they vary in time or velocity. Dynamic time warping (DTW) method (n=1, 1%) can be applied to find such similarities that may exist between people in terms of their mood series. As an example, one may compare time series to find whether they match for stress, depression, or anxiety. Moreover, it can be used to forecast the mental condition of persons with substantially comparable series patterns [130,140]. The difference between DTW and Euclidian matching is that unlike Euclidean matching, DTW considers the distance of each point in one sequence to every point in the other sequence to determine the similarity between them (Figure 4).

Figure 4. Dynamic time warping (left) versus Euclidian matching (right; modified from Portilla and Heintz [142]).



Kalman Filter

The Kalman filter (n=2, 2%) is a technique for making predictions about unknown variables (eg, missing data) based on observable data. Kalman filters include 2 iterative steps, predict and update, that are used to estimate states using linear dynamical systems in state-space format. Iterative cycles of predict and update are performed until convergence is achieved [143]. Kalman filter has been used to handle the missing data for stress detection in some studies [144,145].

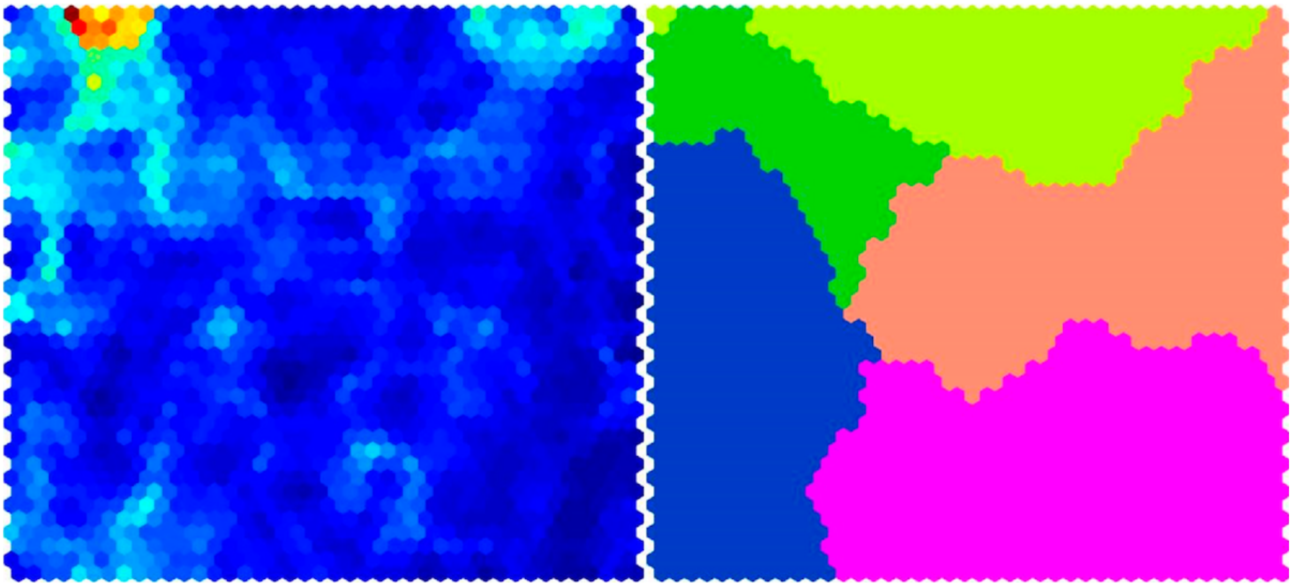
Autoencoders

Autoencoders (n=3, 3%) are a type of NNs that learn a representation of the data in lower dimensions than the original data (encoding) by regenerating the input from the encodings (decoding). For data with very high dimensionality, usually clustering is not optimized because of the noise present in the original data. Hence, it is an appropriate practice to use the encoded representation of the data, obtained by autoencoders, to have lower and more optimized dimensions for clustering [65,108,146].

Self-Organizing Map

In ML, a self-organizing map (n=3, 3%) produces a low-dimensional, typically 2-dimensional, representation of a high-dimensional data set while preserving its topology by creating clusters. Therefore, it is possible to visualize and analyze high-dimensional data more easily (Figure 5) [107,133,147].

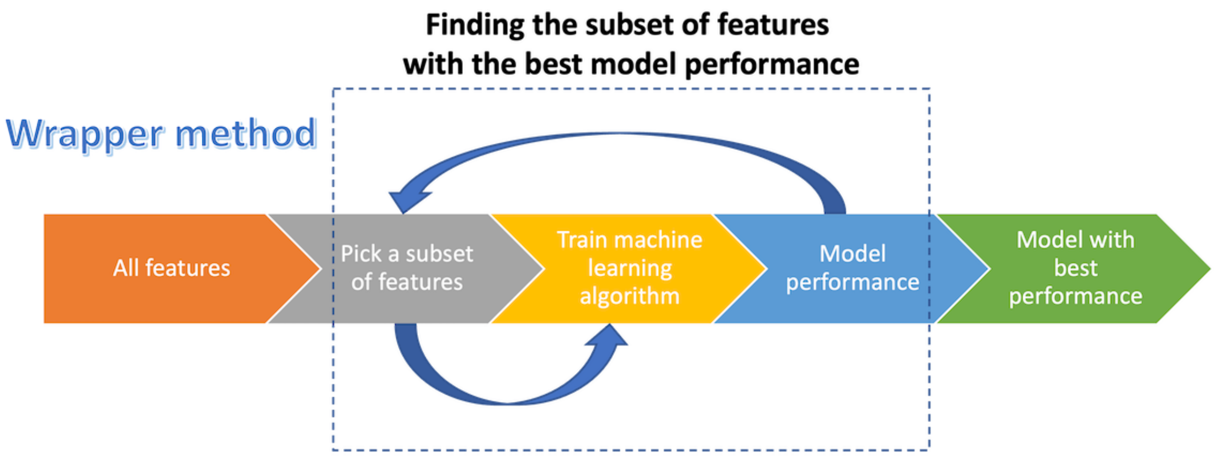
Figure 5. Representation of self-organizing map (SOM) before (left) and after mapping (right; modified from Cho et al [133]).



Wrapper Feature Selection Methods

Wrapper methods try to use a subset of features while training a model. Changes will be made to the feature subset on the basis of the performance of the prior model (Figure 6). Therefore, finding the best features using the wrapper method is a search problem. These methods often have high computing costs [148]. Some of the most common wrapper methods are naive search, sequential forward feature selection, sequential backward feature selection, and generalized sequential search [149]. Some studies used this approach as their feature selection technique [72,75].

Figure 6. Steps of a wrapper feature selection method.



Filter Feature Selection Methods

In general, filter methods are used as a preprocessing step without regard to any ML algorithms. Statistic tests are used instead to select features based on their correlation with dependent variables (Figure 7). The filter feature selection methods used in the literature are mentioned in subsequent sections.

Figure 7. Steps of a filter feature selection method.

Chi-Square Test

This test checks for independence between categorical features and the target variable. Features with high chi-square scores are selected, implying a strong association with the target variable, which may be valuable for the model (n=3, 3%) [56,135,150].

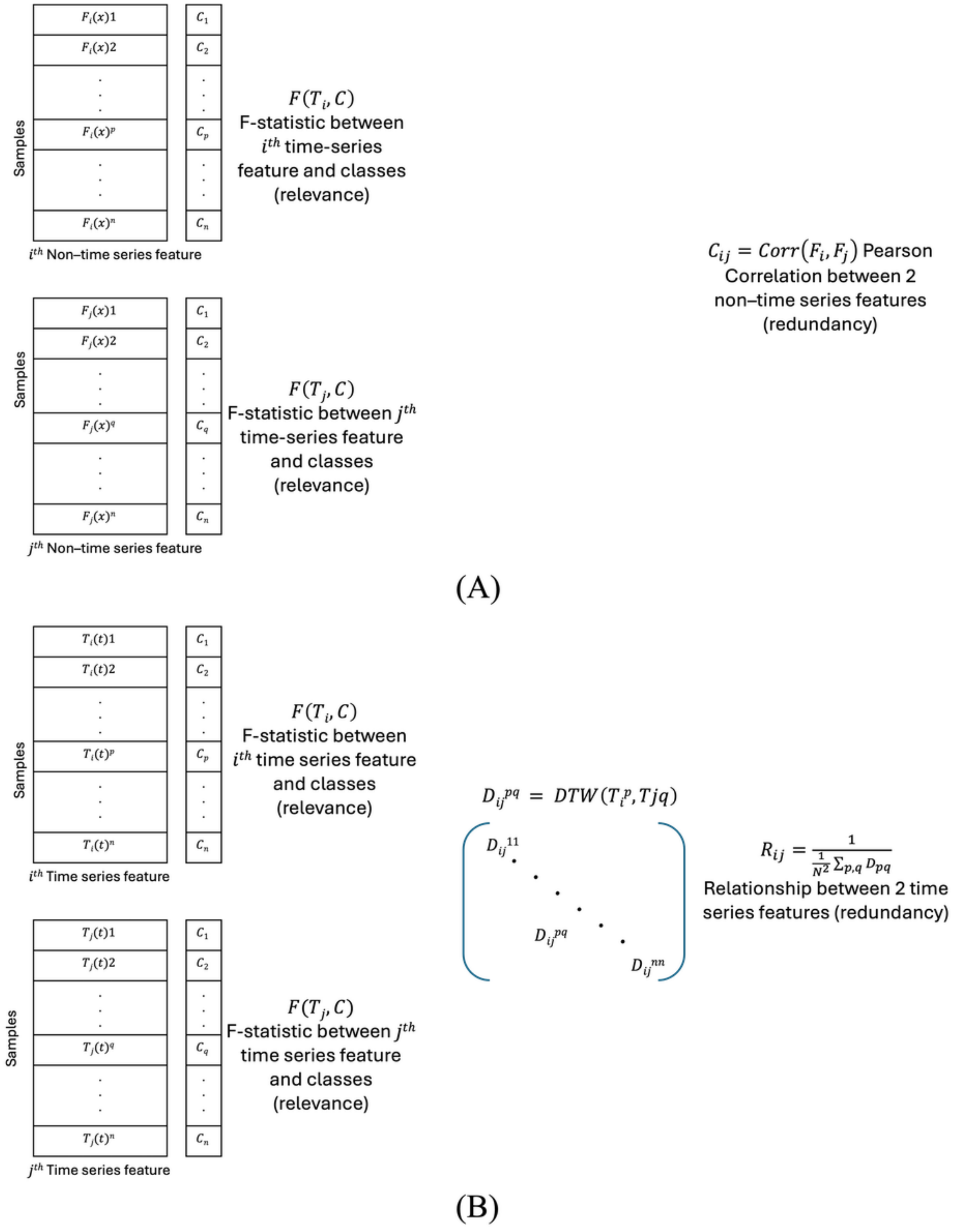
Pearson Correlation

Pearson linear correlation coefficient (n=2, 2%**)** is a way to quantify how closely 2 sets of data are correlated linearly. It indicates how different measures are related to each other by a number between –1 and 1. Therefore, among highly correlated variables, some of them can be removed, as they do not add useful information to ML models [113,151].

Minimum Redundancy Maximum Relevance

The maximum relevance minimum redundancy technique (n=2, 2%**)** chooses characteristics having a high correlation to output (relevance) and a low correlation to one another (redundancy). The *F* statistic is used to determine the correlation between features and the output, while the Pearson correlation coefficient (for non–time series features) and DTW (for time-series features) may be used to calculate the correlation between features (Figure 8). The objective function, which is a function of relevance and redundancy, is then maximized by selecting features one at a time using a greedy search. Mutual information difference and mutual information quotient criteria are both frequently used objective functions that depict the difference or quotient between relevance and redundancy [152,153]. Using this feature selection method, Giannakakis et al [130] have ranked ECG measurements in the order of importance as mean HR, LF, NN50, SD of HR, pNN50, LF or HF, RMSSD, HF, and total power.

Figure 8. Calculation of relevance and redundancy for (A) nontime series features and (B) time-series features. DTW: dynamic time warping.



ML Techniques

The ML algorithms used for stress and MD detection have been reviewed in this section. The papers (n=98) used the DL approach or NN (n=38, 39%), LR (n=26, 27%), naive Bayes (NB; n=22, 22%), decision tree (DT; n=23, 23%), boosting (eg, adaptive boosting, extreme gradient boosting [XGBoost], etc; n=22, 22%), random forest (RF; n=37, 38%), discriminant analysis (DA; eg, linear DA and quadratic DA) (n=6, 6%), fuzzy C-means (FCM; n=2, 2%), k-nearest neighbors (KNNs; n=22, 22%), and support vector machines (SVMs; n=48, 49%). Table 2 shows the distribution of articles by ML model (refer to Table S1 in Multimedia Appendix 2 [29-31,33,36,39-43,56-63,65,66,68,69,73,75-77,79,80,84,88,90,91,94,99,100,105,107-112,114-117,119,122-129,137,139,141,142,146,147,154-160] to find which papers have used each ML technique).

Table 2. Number of articles for each machine learning (ML) model (n=98).

ML model	Articles, n (%)
NN[a]	38 (39)
LR[b]	26 (26)
NB[c]	22 (22)
DT[d]	23 (23)
Boosting	22 (22)
AdaBoost[e]	8 (8)
XGBoost[f]	15 (15)
RF[g]	37 (38)
LDA[h] and QDA[i]	6 (6)
Fuzzy	2 (2)
K-means	4 (4)
KNN[j]	22 (22)
SVM[k]	48 (49)
Other	19 (19)

[a]NN: neural network.
[b]LR: logistic regression.
[c]NB: naive Bayes.
[d]DT: decision tree.
[e]AdaBoost: adaptive boosting.
[f]XGBoost: extreme gradient boosting.
[g]RF: random forest.
[h]LDA: linear discriminant analysis.
[i]QDA: quadratic discriminant analysis.
[j]KNN: k-nearest neighbors.
[k]SVM: support vector machine.

LR Technique

LR (n=26, 27%) is a supervised parametric ML technique in which multiple independent variables will be used to detect the occurrence of stress or normal conditions [72,117]. Some studies used the numerical independent variables (eg, HRV time-domain features: RMSSD, HR, and pNN50) [94,155] or categorical data (eg, answers to multiple choice questions) obtained from questionnaires [107,114,115].

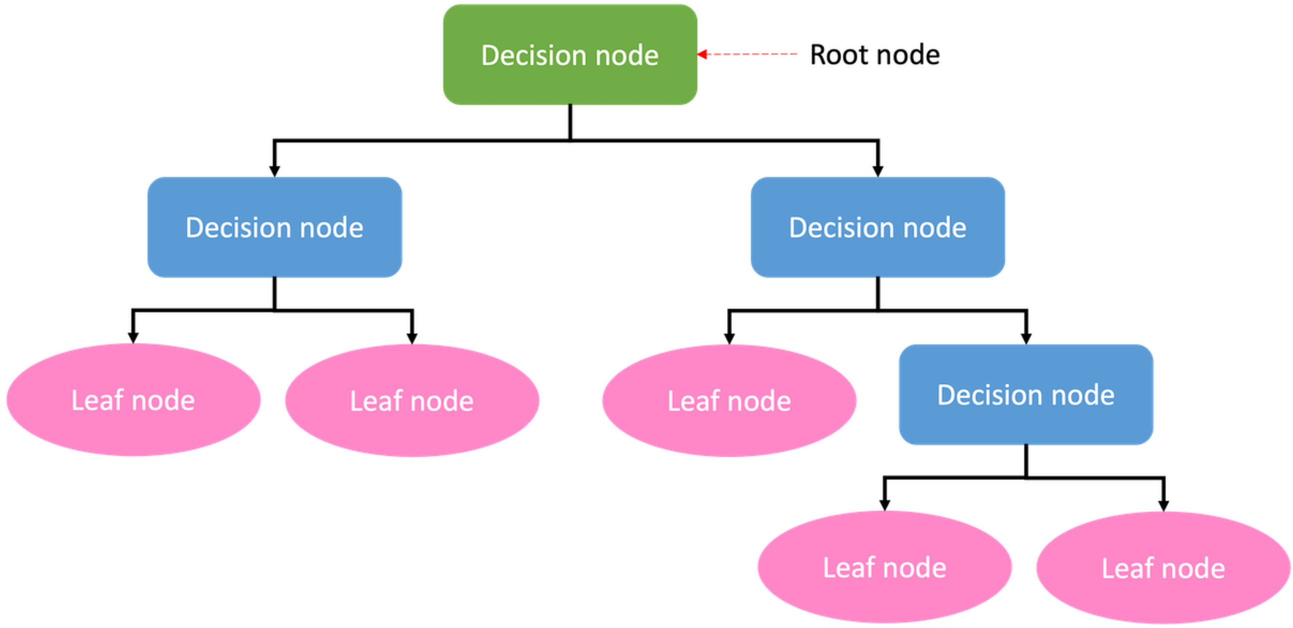
NB Algorithm

NB algorithm (n=22, 22%) is a supervised, generally parametric, classification method that uses the Bayes Theorem as its foundation and has the naive assumption of predictor independence. In other words, the NB classifier assumes that the existence of a given independent variable to predict the dependent variable is independent of the presence of any other independent variable that predicts the dependent variable.

DT Algorithm

The DT (n=23, 23%) is a supervised nonparametric ML algorithm used in classification and regression applications. It comprises a root node, branches, internal nodes, and leaf nodes in a hierarchical, tree-like structure (Figure 9).

Figure 9. Structure of a decision tree.



Boosting Algorithm

Boosting (n=22, 22%) is an ensemble learning for reducing training errors by combining a group of weak learners. When using the Boosting algorithm, models are fitted on random samples of data, and then models are trained repeatedly in a sequence. When each model starts being trained in that sequence, it attempts to make up for the flaws of the one that came before it. The most commonly used Boosting algorithms are adaptive boosting, gradient boosting, and XGBoost.

RF Algorithm

RF (n=37, 38%) is a supervised nonparametric ensemble learning algorithm that uses many DTs built during the training process. The RF algorithm is used for both classification and regression problems. When it comes to classification, the RF's output is the class that most of the DTs choose. For regression purposes, an individual tree's predicted mean or average is returned as the output. Using RFs, we can overcome the tendency of DTs to overfit their training data.

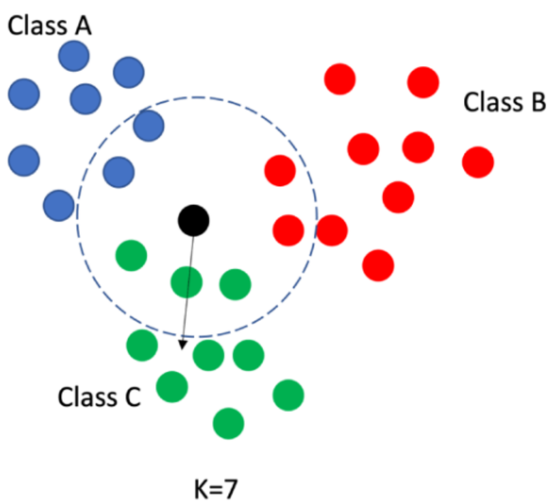
DA Algorithm

DA (n=6, 6%) is a supervised parametric classification algorithm that works with data including a dependent variable and independent variables and is mostly used to classify the observation into a certain group based on the independent variables in the data. Linear DA and quadratic DA are the 2 forms of DA.

KNN Algorithm

KNN (n=22, 22%) is a nonparametric supervised ML algorithm that is used for both classification and regression purposes. In classification, the algorithm determines the label of a new sample not available in the training data by assigning the label of the majority of k-nearest training data points to that new sample (Figure 10). In regression, the output for each sample is the average of the values of KNNs to that sample (not including the sample itself). In this literature, KNN has only been used for classification.

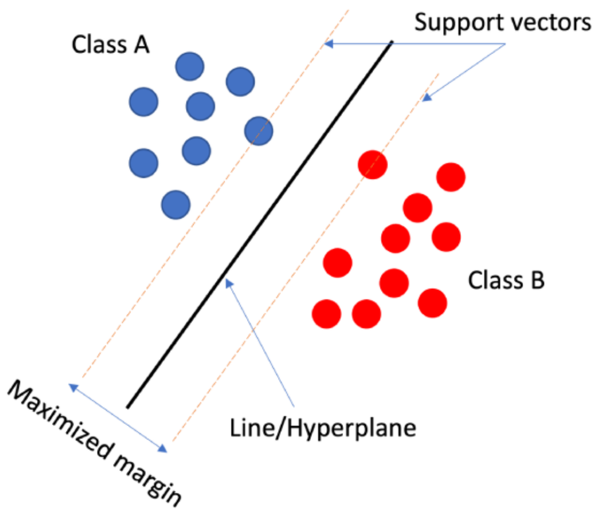
Figure 10. Example of k-nearest neighbor classification with K=7. In this example, the label of “Class C” is assigned to the new (black) datapoint since the majority of the 7-nearest datapoints to the new datapoint are from “Class C.”.



SVM Algorithm

SVM (n=48, 49%) is a parametric supervised ML algorithm used for both classification and regression problems. It can solve both linear and nonlinear problems using nonlinear kernels. For classification, the SVM algorithm finds a line (or a hyperplane for nonlinear kernels) between each pair of classes of the training data in a way that the margin distance of that line or hyperplane to the closest point of each of those 2 classes is maximized (Figure 11). This is repeated for all pairs of classes in the data set. Then, the obtained lines are used as boundaries for the classes. In regression, the SVM tries to find the line or hyperplane that within a very small margin of has the maximum number of data points. That line or hyperplane was used for regression.

Figure 11. Visual representation of support vector machine algorithm.



K-Means Clustering

K-means clustering (n=4, 4%) is an unsupervised ML algorithm that aims to arrange objects into groups based on their similarity. To find those similarities, it calculates the distance of data points into K random cluster centroids and assigns each data point to its closest centroid. The location of each centroid is then updated by the average value of all data points associated with that centroid. This process is repeated until there is no change in the location of the centroids. In ML models for stress detection, K-means clustering has been used in the literature for the personalization of the ML models [58,67], and for labeling the data set [146,156].

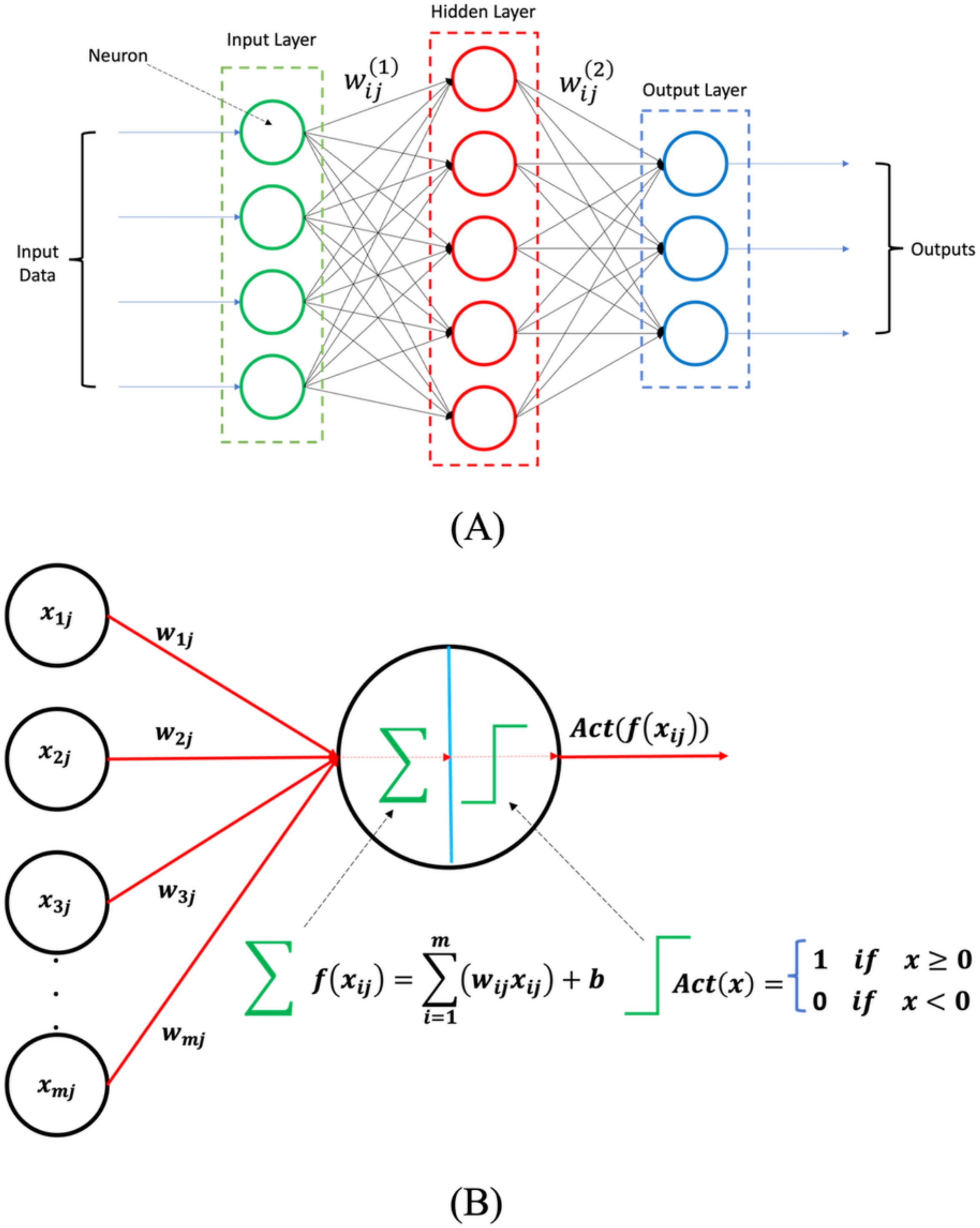
NN Method

DL methods are a subset of ML methods, and NNs are at the heart of the DL algorithms. The NN (n=38, 39%) is a method for implementing ML that uses interconnected nodes or neurons arranged in a layered structure resembling the human brain. The different types of NNs have been explained in subsequent sections.

Artificial Neural Network

It is possible to think of a single perceptron (or neuron) as an abstract LR. In each layer of artificial neural networks (ANNs), a group of multiple perceptron or artificial neurons is used. Figure 12 shows an ANN with 1 layer and its working mechanism.

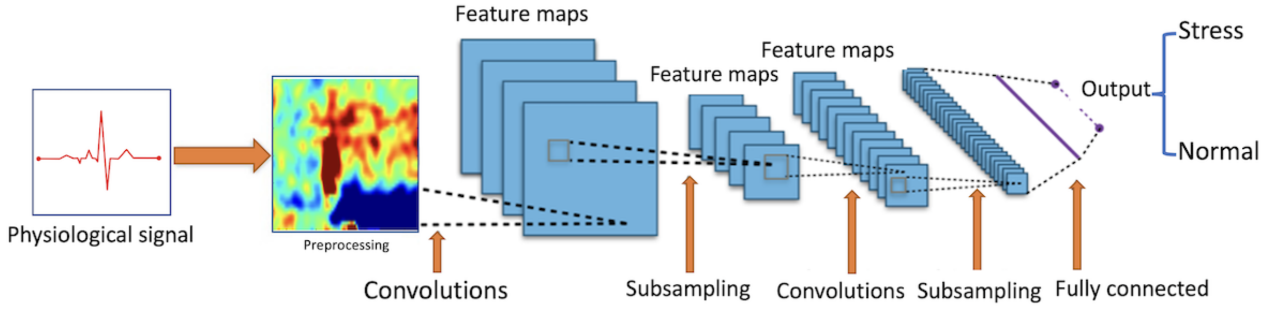
Figure 12. (A) Representation of an artificial neural network with 1 hidden layer. $W_{ij}^{(1)}$ and $W_{ij}^{(2)}$ denote the weights of the links connecting the first layer (input layer) to the hidden layer and the weights of the links connecting the second layer to the next layer (output layer), respectively. (B) Representation of how a single neuron works. First, all the outputs of the previous layer are multiplied by the weights associated with the links connecting them to the jth neuron of the next layer and summed by a bias (summation and bias step). The result is then passed through an activation function (activation step).



CNN Approach

CNNs are a form of NN that is especially adept at handling data structures with a grid-like layout, such as images or objects. Classification and computer vision applications are common uses for CNNs (Figure 13).

Figure 13. Representation of convolutional neural network for a physiological signal.



Recurrent Neural Network Approach

A recurrent neural network (RNN) is a subset of ANNs designed specifically for use with time-series data and other sequence-based data. Long short-term memory (LSTM) networks are the most common type of RNNs. In RNNs, the attention mechanism is a method that simulates cognitive attention in NNs. The purpose of the impact is to encourage the network to give greater attention to the small but significant portions of the input data by enhancing some and reducing others. Since stress may alter a small portion of physiological data (eg, ECG), attention mechanisms can be used to detect stress using RNNs when large data sets are available [141].

Cong et al [102] introduced X-A-BiLSTM, which is a DL model that includes XGBoost (to filter data and handle imbalanced data) and attention Bi-LSTM (LSTM with forward and backward memory and attention mechanism) NN used for stress classification using text data.

Other ML Techniques

The total number of studies in this category include (n=19, 19%) of studies.

Voting Ensemble Classifier

The classification is decided based on weighted voting, which is determined by using a voting ensemble approach. The voting classifier allows for voting in which the final class labels are determined either by the class chosen most frequently by the classification models or by the average of the output probabilities from each classification model. In the literature, this method has been used for PTSD detection [127], stress, and stress-related MDs [84,95,118,156,157].

FCM Clustering

FCM is a clustering approach that assigns every data point to all the clusters with a certain probability instead of assigning each point to only 1 cluster. For instance, a data point that is near the cluster's center will have a high degree of membership, while a data point that is distant from the cluster's center will have a low degree of membership [158]. Since depression and anxiety are not discrete measures, some studies have used FCM as an alternative to other clustering techniques for the detection of these MDs [114,116].

Discussion

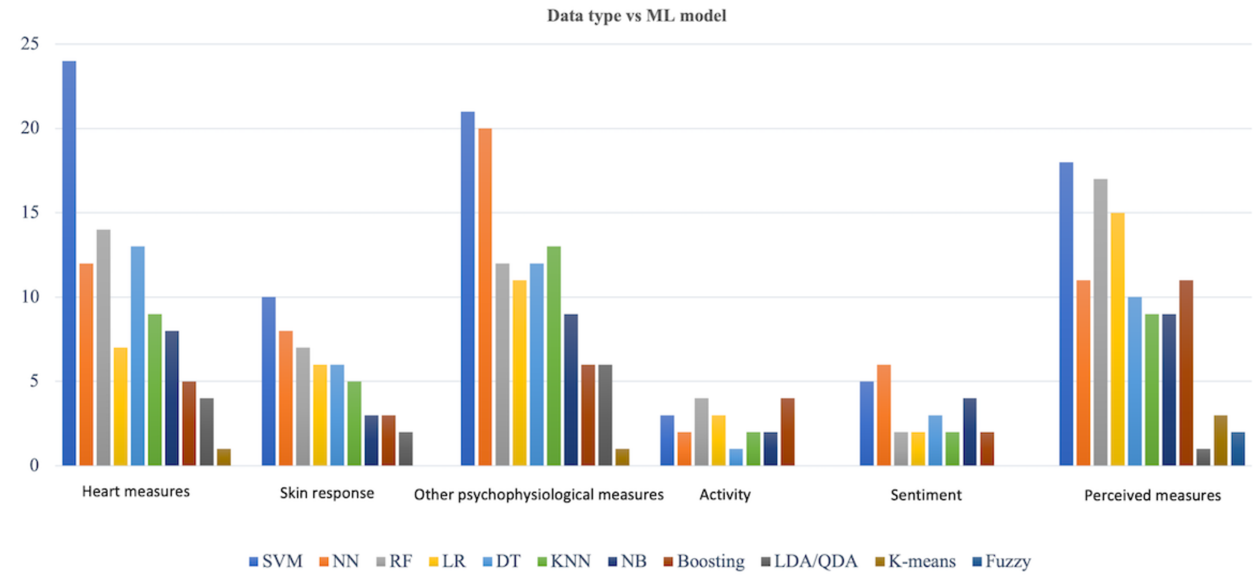
Principal Findings

In this review, the recent ML algorithms; preprocessing techniques; and data (eg, physiological data, questionnaire data, etc) used in the detection, prediction, and monitoring of stress and the most common MDs (ie, depression, anxiety, other stress-related MDs) have been reviewed.

On the basis of this review, it is concluded that among classic ML algorithms (excluding DL approaches), supervised models of SVMs and RF have been used more often and achieved better performance in terms of model accuracy and robustness (measured by parameters such as area under the receiver operating characteristic curve). The accuracy of ML models is a critical indicator of their utility in real-world applications. The review demonstrates that SVM consistently achieves high accuracy across various data types, including HR, HRV, and skin response. For instance, SVM achieved 93% accuracy with HR, PPG, and skin response data in the study by Nath et al [29] and 96% with skin response data in the study by Srividya et al [156]. These results underscore SVM's robustness in handling complex, nonlinear data. RF also shows commendable performance, with an accuracy of 99.88% in the study by Trevisan [159], reflecting its strength in ensemble learning to mitigate overfitting and noise.

Moreover, among the predicting measures for stress and stress-related MDs, HR, HRV, and skin response have been used most often (Figure 14). These measures were the major explaining factors in the ML algorithms to predict stress and stress-related MDs. It is noticeable that DL approaches are becoming more popular as these techniques provide unique specifications that classic ML algorithms cannot provide.

Figure 14. Distribution of machine learning (ML) models used for each type of data. In this figure, skin response and heart measures (including heart rate, heart rate variability, and blood pressure) have been shown separately because of their high use and importance in the literature. Other psychophysiological measures include electroencephalogram, electromyogram, eye tracking, and respiratory signals. Activity includes body movement. Sentiment data include speech and text data. Finally, perceived measures include questionnaires and self-report data. DT: decision tree; KNN: k-nearest neighbors; LDA: linear discriminant analysis; LR: logistic regression; NB: naive Bayes; NN: neural network; QDA: quadratic discriminant analysis; RF: random forest; SVM: support vector machines.



Since stress is a time-dependent event, the relationship between different lags of time can be important for detection of stress. RNNs and CNNs will take into account the relationship between data points in different time series for their decision-making, and they have the potential to enhance the detections. DL models, specifically CNNs and LSTMs, show promising results, with CNNs achieving 92.8% accuracy in HRV and ECG data in the study by Quintero et al [155], indicating their potential in feature-rich physiological data. However, it is worth noting that DL models require substantial data for training, which may limit their applicability in studies with smaller data sets. Attention mechanism in RNNs is a new technique that is becoming popular for finding anomalies in physiological signals. However, based on the review of literature, this mechanism has only been used on text data (not on physiological signals) to detect stress. Therefore, the attention mechanism is the technology that can be further used for physiological signals to detect stress.

Unsupervised ML (and DL algorithms) such as clustering techniques have been used mostly for the preprocessing step to label the data (if labels are not available) and also for finding a representation of the data that achieves the best performance in detection algorithms.

For data preprocessing, feature selection (ie, filter and wrapper methods) and extraction techniques are commonly used. In feature extraction approaches, latent representations of data by transformations such as the output of encoder in autoencoders have been useful to remove data noises and to make the data more compact, making further computations more efficient. PCA and ICA are other most common feature extraction approaches used in the literature.

Among the selected features, statistical indicators of heart measurements such as the mean and SD of HR, along with time and frequency representations of HRV such as RMSSD and total LF and HF power, were most widely used. Heart measurements have also been used more often than other measurements, as they are unobtrusive, noninvasive, affordable, and easier to measure and describe a big portion of stress events. After those measurements, skin response measures have been found to be one of the most important factors in the detection of stress and its related disorders. The time-frequency approaches to analyzing time-series data are becoming more popular in this area as they are proper representations of data for DL approaches that can be more accurate and robust. As an example, for DL algorithms, RNNs with attention mechanisms can help to find portions of data related to stress and its related disorders with higher confidence.

Most of the study models do not interpret the ML models and look at them as black boxes. This limits the contribution to the body of science. Shapley additive explanations is a technique used by some studies to interpret the models such as the evaluation of features to find the most important ones and how in what direction each feature affects the predictions. Shapley additive explanations correlation plot provides insight into the distribution of the features themselves, as well as the relationship between their influence on the model. In other words, it provides the importance of each feature in the prediction of the dependent variable by considering both the main effect and the interaction effect of that feature with other features in the data [31,62,120,159,161,162].

Despite progress in stress detection methodologies, the exploration of personalized models has been limited. Most studies have not gone beyond basic normalization techniques, overlooking the fact that physiological measures are as distinct to individuals as biometric identifiers. A notable exception can be found in a select few studies [67,128,160], which have used more sophisticated personalization techniques, integrating complex data transformations to account for individual variability.

Strengths of the Review

In undertaking this scoping review, we have embarked on a rich exploration of the applications of ML in the field of stress detection, articulating a narrative that is both comprehensive and detailed. The review lays out a landscape in which diverse data types are not merely cataloged but deeply analyzed for their roles and interconnections within the broader context of methodological approaches. This provides a robust understanding of the field's current state and its complexities.

This review has documented a comprehensive assessment of various physiological measurement techniques, including HRV, EEG, ECG, and so on. This assessment is not just a recounting of the types of data used in the literature but a thoughtful consideration of how each contributes to a multifaceted understanding of stress indicators. It is an acknowledgment that the signals of stress are as complex as the condition itself, necessitating a rich palette of investigative tools.

The review also examines a range of advanced preprocessing techniques such as maximum relevance minimum redundancy, self-organizing map, synthetic minority oversampling technique, and PCA. This examination sheds light on how different studies leverage these methods to refine the quality of the data fed into ML models, thereby potentially enhancing the models' accuracy and reliability in detecting stress. It is an illustration of how sophisticated data treatment can lead to more nuanced insights, even if our methodology did not directly use these techniques.

Limitations

Our scoping review acknowledges its inherent constraints, including a possible selection bias due to potential omissions of pertinent studies. It serves as a contemporary cross-section of the rapidly evolving domains of ML and MH, underscoring the imperative for periodic scholarly review to sustain its relevance and precision. While we survey a broad spectrum of ML techniques applied to stress detection, we do not extensively assess their efficacy, suggesting a fertile ground for future empirical investigations to assess these methods across diverse data cohorts and settings. In addition, while we address the preprocessing techniques and their impact on model performance, our discussion does not delve into detailed technical analysis. Finally, the crucial issue of model interpretability is touched upon but not explored in depth, presenting an opportunity for further scholarly explorations.

Conclusions and Future Directions

Overview

The pivotal insights from this review underscore the potential of ML to redefine the approach to MH care, particularly in the diagnosis and management of stress-related conditions and MDs. As we have discerned, there is an expansive field ripe for further exploration, with research gaps suggesting a number of promising directions. Guided by these insights, we can now chart a course for future research that not only expands the boundaries of our scientific understanding but also translates into tangible improvements in clinical practice.

Real-Time and Naturalistic ML Applications

The scarcity of real-time studies in naturalistic settings has highlighted the importance of developing ML models that accurately reflect and respond to the complexities of real life. Future research must prioritize the creation of algorithms capable of operating amidst the unpredictability of daily life, providing immediate insights and adaptable interventions. These models hold the potential to transform practice by offering tools that can preemptively identify stress and MD symptoms, enabling clinicians to intervene before conditions worsen.

Temporal Data and DL

Our review illuminates the untapped potential of time-series data in capturing the evolution of stress and MDs. DL techniques, specifically designed to interpret complex, sequential data, could lead to breakthroughs in how we understand and predict MH trajectories. For practice, this means more sophisticated diagnostic tools that can provide a nuanced picture of a patient's MH over time, enabling personalized treatment plans that are responsive to the patient's changing condition.

Personalization in ML Models

The need for individualized care for MH cannot be overstated. The heterogeneity of stress responses and MD symptoms calls for personalized ML models tailored to individual physiological and behavioral patterns. Future research should focus on leveraging multitask learning to refine algorithms that adapt to individual baselines, enhancing the personalization of care. For clinicians, this means access to tools that can more accurately reflect and respond to the unique needs of each patient, reducing the risk of misdiagnosis and improving treatment efficacy.

Predictive analytics can be instrumental in identifying key factors that contribute to misdiagnosis and delayed help seeking. Future studies should look to build on this knowledge to inform the creation of interventions that encourage timely and accurate diagnosis. In practice, this could lead to the development of targeted screening tools that assist clinicians in recognizing at-risk individuals more effectively. The integration of clinical expertise with ML innovation is crucial for the development of tools that are both advanced and clinically relevant. Collaboration between health care professionals, patients, and artificial intelligence developers will be essential in creating user-centered tools that address real-world needs. This collaborative approach will likely result in the development of artificial intelligence applications that are more intuitive and effective in clinical settings.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search list for literature review.
[XLSX File (Microsoft Excel File), 2269 KB - mental_v11i1e53714_app1.xlsx]

Multimedia Appendix 2
Comparison between machine learning models.
[XLSX File (Microsoft Excel File), 24 KB - mental_v11i1e53714_app2.xlsx]

Multimedia Appendix 3
PRISMA-ScR Checklist.
[PDF File (Adobe PDF File), 307 KB - mental_v11i1e53714_app3.pdf]

References

1. GBD results. Institute for Health Metrics and Evaluation. URL: https://vizhub.healthdata.org/gbd-results?params=gbd-api-2019-permalink/b9a68deae92b3b8772ae33689f5c7fce [accessed 2024-03-27]
2. Mental disorders. World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/mental-disorders [accessed 2024-08-27]
3. Key substance use and mental health indicators in the United States: results from the 2019 National Survey on Drug Use and health. Security Research Hub, U.S. Department of Health and Human Services. 2020. URL: https://digitalcommons.fiu.edu/srhreports/health/health/32 [accessed 2024-04-29]
4. Mojtabai R, Olfson M. National trends in mental health care for US adolescents. JAMA Psychiatry 2020 Jul 01;77(7):703-714 [FREE Full text] [doi: 10.1001/jamapsychiatry.2020.0279] [Medline: 32211824]
5. Mental illness. National Institute of Mental Health. URL: https://www.nimh.nih.gov/health/statistics/mental-illness [accessed 2024-08-27]
6. Mitchell K. Anxiety disorders: types, causes, symptoms, diagnosis, treatment. WebMD LLC. URL: https://www.webmd.com/anxiety-panic/guide/anxiety-disorders [accessed 2024-08-27]
7. Anxiety disorders - symptoms and causes. Mayo Clinic. URL: https://www.mayoclinic.org/diseases-conditions/anxiety/symptoms-causes/syc-20350961 [accessed 2024-08-27]
8. Anxiety disorders. National Institute of Mental Health. URL: https://www.nimh.nih.gov/health/topics/anxiety-disorders [accessed 2024-08-27]
9. Depression. National Institute of Mental Health. URL: https://www.nimh.nih.gov/health/topics/depression [accessed 2024-08-27]
10. Depression (major depressive disorder) - symptoms and causes. Mayo Clinic. URL: https://www.mayoclinic.org/diseases-conditions/depression/symptoms-causes/syc-20356007 [accessed 2024-08-27]
11. Bienertova-Vasku J, Lenart P, Scheringer M. Eustress and distress: neither good nor bad, but rather the same? Bioessays 2020 Jul 17;42(7):e1900238. [doi: 10.1002/bies.201900238] [Medline: 32302008]
12. Stress. The Centre for Addiction and Mental Health. URL: https://www.camh.ca/en/health-info/mental-illness-and-addiction-index/stress [accessed 2024-06-27]
13. Cooper J. The link between stress and depression. WebMD. URL: https://www.webmd.com/depression/features/stress-depression [accessed 2024-08-27]
14. I'm so stressed out! fact sheet. National Institute of Mental Health. URL: https://www.nimh.nih.gov/health/publications/so-stressed-out-fact-sheet [accessed 2024-08-27]
15. Cohen S, Gianaros PJ, Manuck SB. A stage model of stress and disease. Perspect Psychol Sci 2016 Jul 29;11(4):456-463 [FREE Full text] [doi: 10.1177/1745691616646305] [Medline: 27474134]
16. Won E, Kim Y. Stress, the autonomic nervous system, and the immune-kynurenine pathway in the Etiology of depression. Curr Neuropharmacol 2016 Aug 26;14(7):665-673 [FREE Full text] [doi: 10.2174/1570159x14666151208113006] [Medline: 27640517]
17. McCorry LK. Physiology of the autonomic nervous system. Am J Pharm Educ 2007 Aug 15;71(4):78 [FREE Full text] [doi: 10.5688/aj710478] [Medline: 17786266]
18. Pongratz G, Straub RH. The sympathetic nervous response in inflammation. Arthritis Res Ther 2014 Dec 12;16(6):504 [FREE Full text] [doi: 10.1186/s13075-014-0504-2] [Medline: 25789375]
19. Schmidt P, Reiss A, Dürichen R, Laerhoven KV. Wearable-based affect recognition-a review. Sensors (Basel) 2019 Sep 20;19(19):4079 [FREE Full text] [doi: 10.3390/s19194079] [Medline: 31547220]
20. Kim H, Cheon E, Bai D, Lee YH, Koo B. Stress and heart rate variability: a meta-analysis and review of the literature. Psychiatry Investig 2018 Mar;15(3):235-245. [doi: 10.30773/pi.2017.08.17] [Medline: 29486547]
21. Machado AV, Pereira MG, Souza GG, Xavier M, Aguiar C, de Oliveira L, et al. Association between distinct coping styles and heart rate variability changes to an acute psychosocial stress task. Sci Rep 2021 Dec 15;11(1):24025 [FREE Full text] [doi: 10.1038/s41598-021-03386-6] [Medline: 34911998]

22. Adjei T, von Rosenberg W, Nakamura T, Chanwimalueang T, Mandic DP. The ClassA framework: HRV based assessment of SNS and PNS dynamics without LF-HF controversies. Front Physiol 2019;10:505 [FREE Full text] [doi: 10.3389/fphys.2019.00505] [Medline: 31133868]
23. Yoo HH, Yune SJ, Im SJ, Kam BS, Lee SY. Heart rate variability-measured stress and academic achievement in medical students. Med Princ Pract 2021;30(2):193-200 [FREE Full text] [doi: 10.1159/000513781] [Medline: 33326983]
24. Chen J, Abbod M, Shieh J. Pain and stress detection using wearable sensors and devices-a review. Sensors (Basel) 2021 Feb 03;21(4):1030 [FREE Full text] [doi: 10.3390/s21041030] [Medline: 33546235]
25. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science 2015 Jul 17;349(6245):255-260. [doi: 10.1126/science.aaa8415] [Medline: 26185243]
26. Luo J, Wu M, Gopukumar D, Zhao Y. Big data application in biomedical research and health care: a literature review. Biomed Insights 2016 Jan 19;8:BII.S31559. [doi: 10.4137/bii.s31559]
27. Shatte AB, Hutchinson DM, Teague SJ. Machine learning in mental health: a scoping review of methods and applications. Psychol Med 2019 Jul;49(9):1426-1448. [doi: 10.1017/S0033291719000151] [Medline: 30744717]
28. Kofman O, Meiran N, Greenberg E, Balas M, Cohen H. Enhanced performance on executive functions associated with examination stress: Evidence from task-switching and Stroop paradigms. Cogn Emot 2006 Aug;20(5):577-595. [doi: 10.1080/02699930500270913]
29. Nath RK, Thapliyal H, Caban-Holt A, Mohanty SP. Machine learning based solutions for real-time stress monitoring. IEEE Consumer Electron Mag 2020 Sep 1;9(5):34-41. [doi: 10.1109/mce.2020.2993427]
30. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
31. Sadeghi M, McDonald AD, Sasangohar F. Posttraumatic stress disorder hyperarousal event detection using smartwatch physiological and activity data. arXiv Preprint posted online September 29, 2021 [FREE Full text] [doi: 10.1371/journal.pone.0267749]
32. Mallol-Ragolta A, Dhamija S, Boult TE. A multimodal approach for predicting changes in PTSD symptom severity. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. 2018 Presented at: ICMI '18; October 16-20, 2018; Boulder, CO p. 324-333 URL: https://dl.acm.org/doi/10.1145/3242969.3242981 [doi: 10.1145/3242969.3242981]
33. Huang J, Luo X, Peng X. A novel classification method for a driver's cognitive stress level by transferring interbeat intervals of the ECG signal to pictures. Sensors (Basel) 2020 Feb 29;20(5):1340 [FREE Full text] [doi: 10.3390/s20051340] [Medline: 32121440]
34. Durán Acevedo CM, Carrillo Gómez JK, Albarracín Rojas CA. Academic stress detection on university students during COVID-19 outbreak by using an electronic nose and the galvanic skin response. Biomed Signal Process Control 2021 Jul;68:102756 [FREE Full text] [doi: 10.1016/j.bspc.2021.102756] [Medline: 36570516]
35. Wong JC, Wang J, Fu EY, Leong HV, Ngai G. Activity recognition and stress detection via wristband. In: Proceedings of the 17th International Conference on Advances in Mobile Computing & Multimedia. 2019 Presented at: MoMM' 19; December 2-4, 2019; Munich, Germany p. 102-106 URL: https://dl.acm.org/doi/10.1145/3365921.3365950 [doi: 10.1145/3365921.3365950]
36. Šalkevicius J, Damaševičius R, Maskeliunas R, Laukienė I. Anxiety level recognition for virtual reality therapy system using physiological signals. Electronics 2019 Sep 16;8(9):1039. [doi: 10.3390/electronics8091039]
37. Wongkoblap A, Vadillo MA, Curcin V. Classifying depressed users with multiple instance learning from social network data. In: Proceedings of the 2018 IEEE International Conference on Healthcare Informatics. 2018 Presented at: ICHI '18; June 4-7, 2018; New York, NY p. 436 URL: https://ieeexplore.ieee.org/document/8419421 [doi: 10.1109/ichi.2018.00088]
38. Chandra V, Priyarup A, Sethia D. Comparative study of physiological signals from Empatica E4 wristband for stress classification. In: Proceedings of the 5th International Conference on Advances in Computing and Data Sciences. 2021 Presented at: ICACDS '21; April 23-24, 2021; Nashik, India p. 218-229 URL: https://link.springer.com/chapter/10.1007/978-3-030-88244-0_21 [doi: 10.1007/978-3-030-88244-0_21]
39. Baek J, Chung K. Context deep neural network model for predicting depression risk using multiple regression. IEEE Access 2020;8:18171-18181. [doi: 10.1109/access.2020.2968393]
40. Bichindaritz I, Breen C, Cole E, Keshan N, Parimi P. Feature selection and machine learning based multilevel stress detection from ECG signals. In: Proceedings of the 5th KES International Conference on Innovation in Medicine and Healthcare. 2017 Presented at: KES-InMed '17; June 20-22, 2022; Rhodes, Greece p. 202-217 URL: https://link.springer.com/chapter/10.1007/978-3-319-59397-5_22 [doi: 10.1007/978-3-319-59397-5_22]
41. Maritsch M, Bérubé C, Kraus M, Lehmann V, Züger T, Feuerriegel S, et al. Improving heart rate variability measurements from consumer smartwatches with machine learning. In: Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers. 2019 Presented at: UbiComp/ISWC '19; September 9-13, 2019; London, UK p. 934-938 URL: https://dl.acm.org/doi/10.1145/3341162.3346276 [doi: 10.1145/3341162.3346276]
42. Asha N, Ehtesum-Ul-Islam, Khan R. Low-cost heart rate sensor and mental stress detection using machine learning. In: Proceedings of the 2021 5th International Conference on Trends in Electronics and Informatics. 2021 Presented at: ICOEI

'21; June 3-5, 2021; Tirunelveli, India p. 1369-1374 URL: https://ieeexplore.ieee.org/document/9452873 [doi: 10.1109/icoei51242.2021.9452873]
43. Katarya R, Maan S. Predicting mental health disorders using machine learning for employees in technical and non-technical companies. In: Proceedings of the 2020 IEEE International Conference on Advances and Developments in Electrical and Electronics Engineering. 2020 Presented at: ICADEE '20; December 10-11, 2020; Coimbatore, India p. 1-5 URL: https://ieeexplore.ieee.org/document/9368923 [doi: 10.1109/icadee51157.2020.9368923]
44. Tate AE, McCabe RC, Larsson H, Lundström S, Lichtenstein P, Kuja-Halkola R. Predicting mental health problems in adolescence using machine learning techniques. PLoS One 2020;15(4):e0230389 [FREE Full text] [doi: 10.1371/journal.pone.0230389] [Medline: 32251439]
45. Wshah S, Skalka C, Price M. Predicting posttraumatic stress disorder risk: a machine learning approach. JMIR Ment Health 2019 Jul 22;6(7):e13946 [FREE Full text] [doi: 10.2196/13946] [Medline: 31333201]
46. van Eeden WA, Luo C, van Hemert AM, Carlier IV, Penninx BW, Wardenaar KJ, et al. Predicting the 9-year course of mood and anxiety disorders with automated machine learning: a comparison between auto-sklearn, naïve Bayes classifier, and traditional logistic regression. Psychiatry Res 2021 May;299:113823. [doi: 10.1016/j.psychres.2021.113823]
47. Alharthi H. Predicting the level of generalized anxiety disorder of the coronavirus pandemic among college age students using artificial intelligence technology. In: Proceedings of the 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science. 2020 Presented at: DCABES '20; October 16-19, 2020; Xuzhou, China p. 218-221 URL: https://ieeexplore.ieee.org/document/9277785 [doi: 10.1109/dcabes50732.2020.00064]
48. He J, Li K, Liao X, Zhang P, Jiang N. Real-time detection of acute cognitive stress using a convolutional neural network from electrocardiographic signal. IEEE Access 2019;7:42710-42717. [doi: 10.1109/access.2019.2907076]
49. Sah RK, Ghasemzadeh H. Stress classification and personalization: getting the most out of the least. arXiv Preprint posted online July 12, 2021 [FREE Full text]
50. Zainudin Z, Hasan S, Shamsuddin S, Argawal S. Stress detection using machine learning and deep learning. J Phys Conf Ser 2021 Aug 01;1997(1):012019. [doi: 10.1088/1742-6596/1997/1/012019]
51. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
52. Dimitriev DA, Saperova EV, Dimitriev AD. State anxiety and nonlinear dynamics of heart rate variability in students. PLoS One 2016 Jan 25;11(1):e0146131 [FREE Full text] [doi: 10.1371/journal.pone.0146131] [Medline: 26807793]
53. Heart rate variability (HRV):"what it is and how you can track it". Cleveland Clinic. URL: https://my.clevelandclinic.org/health/symptoms/21773-heart-rate-variability-hrv [accessed 2024-08-27]
54. Shaffer F, Ginsberg JP. An overview of heart rate variability metrics and norms. Front Public Health 2017;5:258 [FREE Full text] [doi: 10.3389/fpubh.2017.00258] [Medline: 29034226]
55. What is HRV? Medicore. URL: http://www.medi-core.co.kr/en/technology/hrv.html [accessed 2024-08-27]
56. Choi Y, Jeon Y, Wang L, Kim K. A biological signal-based stress monitoring framework for children using wearable devices. Sensors (Basel) 2017 Aug 23;17(9):1936 [FREE Full text] [doi: 10.3390/s17091936] [Medline: 28832507]
57. Sakri O, Godin C, Vila G, Labyt E, Charbonnier S, Campagne A. A multi-user multi-task model for stress monitoring from wearable sensors. In: Proceedings of the 21st International Conference on Information Fusion. 2018 Presented at: FUSION '18; July 10-13, 2018; Cambridge, UK p. 761-766 URL: https://ieeexplore.ieee.org/document/8455378 [doi: 10.23919/icif.2018.8455378]
58. Elgendi M, Menon C. Machine learning ranks ECG as an optimal wearable biosignal for assessing driving stress. IEEE Access 2020;8:34362-34374. [doi: 10.1109/access.2020.2974933]
59. Wang R, Jia W, Mao Z, Sclabassi RJ, Sun M. Cuff-free blood pressure estimation using pulse transit time and heart rate. Int Conf Signal Process Proc 2014 Oct;2014:115-118 [FREE Full text] [doi: 10.1109/ICOSP.2014.7014980] [Medline: 26213717]
60. Stress and high blood pressure: what's the connection? Mayo Clinic. URL: https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/in-depth/stress-and-high-blood-pressure/art-20044190 [accessed 2024-08-27]
61. Managing stress to control high blood pressure. American Heart Association. URL: https://www.heart.org/en/health-topics/high-blood-pressure/changes-you-can-make-to-manage-high-blood-pressure/managing-stress-to-control-high-blood-pressure [accessed 2024-08-27]
62. Schultebraucks K, Sijbrandij M, Galatzer-Levy I, Mouthaan J, Olff M, van Zuiden M. Forecasting individual risk for long-term posttraumatic stress disorder in emergency medical settings using biomedical data: a machine learning multicenter cohort study. Neurobiol Stress 2021 May;14:100297 [FREE Full text] [doi: 10.1016/j.ynstr.2021.100297] [Medline: 33553513]
63. Masood K, Alghamdi MA. Modeling mental stress using a deep learning framework. IEEE Access 2019;7:68446-68454. [doi: 10.1109/access.2019.2917718]
64. Shah D, Wang GY, Doborjeh M, Doborjeh Z, Kasabov N. Deep learning of EEG data in the NeuCube brain-inspired spiking neural network architecture for a better understanding of depression. In: Proceedings of the 26th International Conference on Neural Information Processing. 2019 Presented at: ICONIP '19; December 12-15, 2019; Sydney, Australia p. 195-206 URL: https://link.springer.com/chapter/10.1007/978-3-030-36718-3_17 [doi: 10.1007/978-3-030-36718-3_17]

65. Akella A, Singh AK, Leong D, Lal S, Newton P, Clifton-Bligh R, et al. Classifying multi-level stress responses from brain cortical EEG in nurses and non-health professionals using machine learning auto encoder. IEEE J Transl Eng Health Med 2021;9:1-9. [doi: 10.1109/jtehm.2021.3077760]
66. Arsalan A, Majid M, Anwar SM, Bagci U. Classification of perceived human stress using physiological signals. In: Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2019 Presented at: EMBC '19; July 23-27, 2019; Berlin, Germany p. 1247-1250 URL: https://ieeexplore.ieee.org/document/8856377 [doi: 10.1109/embc.2019.8856377]
67. Gonzalez-Carabarin L, Castellanos-Alvarado E, Castro-Garcia P, Garcia-Ramirez M. Machine Learning for personalised stress detection: inter-individual variability of EEG-ECG markers for acute-stress response. Comput Methods Programs Biomed 2021 Sep;209:106314. [doi: 10.1016/j.cmpb.2021.106314] [Medline: 34433128]
68. Subhani AR, Mumtaz W, Saad MN, Kamel N, Malik AS. Machine learning framework for the detection of mental stress at multiple levels. IEEE Access 2017;5:13545-13556. [doi: 10.1109/access.2017.2723622]
69. Nagar P, Sethia D. Brain mapping based stress identification using portable EEG based device. In: Proceedings of the 11th International Conference on Communication Systems & Networks. 2019 Presented at: COMSNETS '19; January 7-11, 2019; Bengaluru, India p. 601-606 URL: https://ieeexplore.ieee.org/document/8711009 [doi: 10.1109/comsnets.2019.8711009]
70. Arsalan A, Majid M, Butt AR, Anwar SM. Classification of perceived mental stress using a commercially available EEG headband. IEEE J Biomed Health Inform 2019 Nov;23(6):2257-2264. [doi: 10.1109/jbhi.2019.2926407]
71. McEwen BS, Gianaros PJ. Central role of the brain in stress and adaptation: links to socioeconomic status, health, and disease. Ann N Y Acad Sci 2010 Feb 18;1186(1):190-222 [FREE Full text] [doi: 10.1111/j.1749-6632.2009.05331.x] [Medline: 20201874]
72. Arsalan A, Majid M, Anwar SM. Electroencephalography based machine learning framework for anxiety classification. In: Proceedings of the 2nd International Conference on Intelligent Technologies and Applications. 2019 Presented at: INTAP '19; November 6-8, 2019; Bahawalpur, Pakistan p. 187-197 URL: https://link.springer.com/chapter/10.1007/978-981-15-5232-8_17 [doi: 10.1007/978-981-15-5232-8_17]
73. Stoeve M, Wirth M, Farlock R, Antunovic A, Müller V, Eskofier BM. Eye tracking-based stress classification of athletes in virtual reality. Proc ACM Comput Graph Interact Tech 2022 May 17;5(2):1-17 [FREE Full text] [doi: 10.1145/3530796]
74. Skaramagkas V, Giannakakis G, Ktistakis E, Manousos D, Karatzanis I, Tachos N, et al. Review of eye tracking metrics involved in emotional and cognitive processes. IEEE Rev Biomed Eng 2023;16:260-277. [doi: 10.1109/rbme.2021.3066072]
75. Giannakakis G, Pediaditis M, Manousos D, Kazantzaki E, Chiarugi F, Simos P, et al. Stress and anxiety detection using facial cues from videos. Biomed Signal Process Control 2017 Jan;31:89-101. [doi: 10.1016/j.bspc.2016.06.020]
76. Korda AI, Giannakakis G, Ventouras E, Asvestas PA, Smyrnis N, Marias K, et al. Recognition of blinks activity patterns during stress conditions using CNN and Markovian analysis. Signals 2021 Jan 23;2(1):55-71. [doi: 10.3390/signals2010006]
77. Bauer EA, Wilson KA, MacNamara A. Cognitive and affective psychophysiology. In: Asmundson GJ, editor. Comprehensive Clinical Psychology. Oxford, UK: Elsevier Academic Press; 2022:49-61.
78. Dawson ME, Schell AM, Filion DL. The electrodermal system. In: Cacioppo JT, Tassinary LG, Berntson GG, editors. Handbook of Psychophysiology. Cambridge, MA: Cambridge University Press; 2017:217-243.
79. Philippot P, Chapelle G, Blairy S. Respiratory feedback in the generation of emotion. Cogn Emot 2002 Aug;16(5):605-627. [doi: 10.1080/02699930143000392]
80. Suess WM, Alexander AB, Smith DD, Sweeney HW, Marion RJ. The effects of psychological stress on respiration: a preliminary study of anxiety and hyperventilation. Psychophysiology 1980 Nov;17(6):535-540. [doi: 10.1111/j.1469-8986.1980.tb02293.x] [Medline: 7443919]
81. Cohen HD, Goodenough DR, Witkin HA, Oltman P, Gould H, Shulman E. The effects of stress on components of the respiration cycle. Psychophysiology 1975 Jul;12(4):377-380. [doi: 10.1111/j.1469-8986.1975.tb00005.x] [Medline: 1162000]
82. Seo W, Kim N, Kim S, Lee C, Park S. Deep ECG-respiration network (DeepER Net) for recognizing mental stress. Sensors (Basel) 2019 Jul 09;19(13):3021 [FREE Full text] [doi: 10.3390/s19133021] [Medline: 31324001]
83. Kong F, Wen W, Liu G, Xiong R, Yang X. Autonomic nervous pattern analysis of trait anxiety. Biomed Signal Process Control 2022 Jan;71:103129. [doi: 10.1016/j.bspc.2021.103129]
84. Cipresso P, Colombo D, Riva G. Computational psychometrics using psychophysiological measures for the assessment of acute mental stress. Sensors (Basel) 2019 Feb 14;19(4):781 [FREE Full text] [doi: 10.3390/s19040781] [Medline: 30769812]
85. Electromyography (EMG). Johns Hopkins Medicine. URL: https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/electromyography-emg [accessed 2024-08-27]
86. Al-Jumaily AA, Matin N, Hoshyar AN. Machine learning based biosignals mental stress detection. In: Proceedings of the 6th International Conference on Soft Computing in Data Science. 2021 Presented at: SCDS '21; November 2-3, 2021; Virtual Event p. 28-41 URL: https://link.springer.com/chapter/10.1007/978-981-16-7334-4_3 [doi: 10.1007/978-981-16-7334-4_3]
87. Dhaouadi S, Ben Khelifa MM. A multimodal physiological-based stress recognition: deep learning models' evaluation in gamers' monitoring application. In: Proceedings of the 5th International Conference on Advanced Technologies for Signal and Image Processing. 2020 Presented at: ATSIP '20; September 2-5, 2020; Sousse, Tunisia p. 1-6 URL: https://ieeexplore.ieee.org/document/9231666 [doi: 10.1109/atsip49331.2020.9231666]

88. Liang WC, Yuan J, Sun DC, Lin MH. Changes in physiological parameters induced by indoor simulated driving: effect of lower body exercise at mid-term break. Sensors 2007;9(9):6913-6933 [FREE Full text]
89. Sano A, Picard RW. Stress recognition using wearable sensors and mobile phones. In: Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. 2013 Presented at: ACII '13; September 2-5, 2013; Geneva, Switzerland p. 671-676 URL: https://ieeexplore.ieee.org/document/6681508 [doi: 10.1109/acii.2013.117]
90. Ghandeharioun A, Fedor S, Sangermano L, Ionescu D, Alpert J, Dale C, et al. Objective assessment of depressive symptoms with machine learning and wearable sensors data. In: Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction. 2017 Presented at: ACII '17; October 23-26, 2017; San Antonio, TX p. 325-332 URL: https://ieeexplore.ieee.org/document/8273620 [doi: 10.1109/acii.2017.8273620]
91. Tazawa Y, Liang K, Yoshimura M, Kitazawa M, Kaise Y, Takamiya A, et al. Evaluating depression with multimodal wristband-type wearable device: screening and assessing patient severity utilizing machine-learning. Heliyon 2020 Feb;6(2):e03274 [FREE Full text] [doi: 10.1016/j.heliyon.2020.e03274] [Medline: 32055728]
92. Yoshiuchi K, Inada S, Nakahara R, Akabayashi A, Park H, Park S, et al. Stressful life events and habitual physical activity in older adults: 1-year accelerometer data from the Nakanojo Study. Ment Health Phys Act 2010 Jun;3(1):23-25. [doi: 10.1016/j.mhpa.2010.02.001]
93. Schmidt P, Reiss A, Duerichen R, Marberger C, Van Laerhoven K. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. 2018 Presented at: ICMI '18; October 16-20, 2018; Boulder, CO p. 400-408 URL: https://dl.acm.org/doi/10.1145/3242969.3242985 [doi: 10.1145/3242969.3242985]
94. Dai R, Lu C, Yun L, Lenze E, Avidan M, Kannampallil T. Comparing stress prediction models using smartwatch physiological signals and participant self-reports. Comput Methods Programs Biomed 2021 Sep;208:106207. [doi: 10.1016/j.cmpb.2021.106207] [Medline: 34161847]
95. Siirtola P, Röning J. Comparison of regression and classification models for user-independent and personal stress detection. Sensors (Basel) 2020 Aug 07;20(16):4402 [FREE Full text] [doi: 10.3390/s20164402] [Medline: 32784547]
96. Speech signal analysis with applications in biomedicine and the life sciences. Frontiers. URL: https://www.frontiersin.org/research-topics/3217/speech-signal-analysis-with-applications-in-biomedicine-and-the-life-sciences [accessed 2024-08-27]
97. Nkurikiyeyezu K, Yokokubo A, Lopez G. The effect of person-specific biometrics in improving generic stress predictive models. arXiv Preprint posted online October 4, 2019 [FREE Full text] [doi: 10.48550/arXiv.1910.01770]
98. Garcia-Ceja E, Thambawita V, Hicks SA, Jha D, Jakobsen P, Hammer HL, et al. HTAD: a home-tasks activities dataset with wrist-accelerometer and audio features. In: Proceedings of the 27th International Conference on MultiMedia Modeling. 2021 Presented at: MMM '21; June 22-24, 2021; Prague, Czech Republic p. 196-205 URL: https://link.springer.com/chapter/10.1007/978-3-030-67835-7_17 [doi: 10.31219/osf.io/j723c]
99. Zhao S, Li Q, Li C, Li Y, Lu K. A CNN-based method for depression detecting form audio. In: Proceedings of the 2nd International Conference on Digital Health and Medical Analytics. 2020 Presented at: DHA '20; July 25, 2020; Beijing, China p. 1-10 URL: https://link.springer.com/chapter/10.1007/978-981-16-3631-8_1 [doi: 10.1007/978-981-16-3631-8_1]
100. de Melo WC, Granger E, Hadid A. Depression detection based on deep distribution learning. In: Proceedings of the 2019 IEEE International Conference on Image Processing. 2019 Presented at: ICIP '19; September 22-25, 2019; Taipei, Taiwan p. 4544-4548 URL: https://ieeexplore.ieee.org/document/8803467 [doi: 10.1109/icip.2019.8803467]
101. Awasthi K, Nanda P, Suma KV. Performance analysis of machine learning techniques for classification of stress levels using PPG signals. In: Proceedings of the 2020 IEEE International Conference on Electronics, Computing and Communication Technologies. 2020 Presented at: CONECCT '20; July 2-4, 2020; Bangalore, India p. 1-6 URL: https://ieeexplore.ieee.org/document/9198481 [doi: 10.1109/conecct50063.2020.9198481]
102. Cong Q, Feng Z, Li F, Xiang Y, Rao G, Tao C. X-A-BiLSTM: a deep learning approach for depression detection in imbalanced data. In: Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine. 2018 Presented at: BIBM '18; December 3-6, 2018; Madrid, Spain p. 1624-1627 URL: https://ieeexplore.ieee.org/document/8621230 [doi: 10.1109/bibm.2018.8621230]
103. Ding Y, Chen X, Fu Q, Zhong S. A depression recognition method for college students using deep integrated support vector algorithm. IEEE Access 2020;8:75616-75629. [doi: 10.1109/access.2020.2987523]
104. Islam MR, Kabir MA, Ahmed A, Kamal AR, Wang H, Ulhaq A. Depression detection from social network data using machine learning techniques. Health Inf Sci Syst 2018 Dec;6(1):8. [doi: 10.1007/s13755-018-0046-0] [Medline: 30186594]
105. Liao S, zhang Q, Gan R. Construction of real-time mental health early warning system based on machine learning. J Phys Conf Ser 2021 Feb 01;1812(1):012032. [doi: 10.1088/1742-6596/1812/1/012032]
106. Pandey M, Jha B, Thakur R. An exploratory analysis pertaining to stress detection in adolescents. In: Proceedings of the 2019 Conference on Soft Computing: Theories and Applications. 2019 Presented at: SoCTA '19; December 27-29, 2019; Patna, India p. 413-421 URL: https://link.springer.com/chapter/10.1007/978-981-15-4032-5_38 [doi: 10.1007/978-981-15-4032-5_38]
107. Simms T, Ramstedt C, Rich M, Richards M, Martinez T, Giraud-Carrier C. Detecting cognitive distortions through machine learning text analytics. In: Proceedings of the 2017 IEEE International Conference on Healthcare Informatics. 2017 Presented

at: ICHI '17; August 23-26, 2017; Park City, UT p. 508-512 URL: https://ieeexplore.ieee.org/abstract/document/8031202 [doi: 10.1109/ichi.2017.39]
108. Wang X, Chen S, Li T, Li T, Zhou Y, Zheng J, et al. Assessing depression risk in Chinese microblogs: a corpus and machine learning methods. In: Proceedings of the 2019 IEEE International Conference on Healthcare Informatics. 2019 Presented at: ICHI '19; June 10-13, 2019; Xi'an, China p. 1-5 URL: https://ieeexplore.ieee.org/document/8904506 [doi: 10.1109/ichi.2019.8904506]
109. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5-TR). New York, NY: American Psychological Association; 2022.
110. Salma UM, Ann KA. Active learning from an imbalanced dataset. In: Jain V, Juneja S, Juneja A, Kannan R, editors. Handbook of Machine Learning for Computational Optimization. Boca Raton, FL: CRC Press; 2021.
111. Mushtaq MS, Mellouk A. QoE-based power efficient LTE downlink scheduler. In: Mushtaq MS, Mellouk A, editors. Quality of Experience Paradigm in Multimedia Services: Application to OTT Video Streaming and VoIP Services. San Diego, CA: Elsevier Academic Press; 2017:91-125.
112. Vanishri A, Prajwal V, Murali K, Arunkumar BV, Padma SK, Shyam V. A boosted machine learning approach for detection of depression. In: Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence. 2018 Presented at: SSCI '18; November 18-21, 2018; Bangalore, India p. 41-47 URL: https://ieeexplore.ieee.org/document/8628945 [doi: 10.1109/ssci.2018.8628945]
113. Casaccia S, Romeo L, Calvaresi A, Morresi N, Monteriu A, Frontoni E, et al. Measurement of users' well-being through domotic sensors and machine learning algorithms. IEEE Sensors J 2020 Jul 15;20(14):8029-8038. [doi: 10.1109/jsen.2020.2981209]
114. Iliou T, Konstantopoulou G, Ntekouli M, Lymperopoulou C, Assimakopoulos K, Galiatsatos D, et al. ILIOU machine learning preprocessing method for depression type prediction. Evol Syst 2017 Oct 25;10(1):29-39. [doi: 10.1007/s12530-017-9205-9]
115. Joseph R, Udupa S, Jangale S, Kotkar K, Pawar P. Employee attrition using machine learning and depression analysis. In: Proceedings of the 5th International Conference on Intelligent Computing and Control Systems. 2021 Presented at: ICICCS '21; May 6-8, 2021; Madurai, India p. 1000-1005 URL: https://ieeexplore.ieee.org/document/9432259 [doi: 10.1109/iciccs51141.2021.9432259]
116. Khalil RM, Al-Jumaily A. Machine learning based prediction of depression among type 2 diabetic patients. In: Proceedings of the 12th International Conference on Intelligent Systems and Knowledge Engineering. 2017 Presented at: ISKE '17; November 24-26, 2017; Nanjing, China p. 1-5 URL: https://ieeexplore.ieee.org/document/8258766 [doi: 10.1109/iske.2017.8258766]
117. Kumar A. Machine learning for psychological disorder prediction in Indians during COVID-19 nationwide lockdown. Intell Decis Technol 2021 Mar 24;15(1):161-172. [doi: 10.3233/idt-200061]
118. Kumar P, Chauhan R, Stephan T, Shankar A, Thakur S. A machine learning implementation for mental health care. Application: smart watch for depression detection. In: Proceedings of the 11th International Conference on Cloud Computing, Data Science & Engineering. 2021 Presented at: Confluence '21; January 28-29, 2021; Noida, India p. 568-574. [doi: 10.1109/confluence51648.2021.9377199]
119. Saif Alghawli A, Taloba AI. An enhanced ant colony optimization mechanism for the classification of depressive disorders. Comput Intell Neurosci 2022;2022:1332664 [FREE Full text] [doi: 10.1155/2022/1332664] [Medline: 35800708]
120. Nemesure MD, Heinz MV, Huang R, Jacobson NC. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. Sci Rep 2021 Jan 21;11(1):1980 [FREE Full text] [doi: 10.1038/s41598-021-81368-4] [Medline: 33479383]
121. Obeid JS, Weeda ER, Matuskowitz AJ, Gagnon K, Crawford T, Carr CM, et al. Automated detection of altered mental status in emergency department clinical notes: a deep learning approach. BMC Med Inform Decis Mak 2019 Aug 19;19(1):164 [FREE Full text] [doi: 10.1186/s12911-019-0894-9] [Medline: 31426779]
122. Reddy US, Thota AV, Dharun A. Machine learning techniques for stress prediction in working employees. In: Proceedings of the 2018 IEEE International Conference on Computational Intelligence and Computing Research. 2018 Presented at: ICCIC '18; December 13-15, 2018; Madurai, India p. 1-4 URL: https://ieeexplore.ieee.org/document/8782395?signout=success [doi: 10.1109/iccic.2018.8782395]
123. Sau A, Bhakta I. Predicting anxiety and depression in elderly patients using machine learning technology. Healthc technol lett 2017 Nov 10;4(6):238-243. [doi: 10.1049/htl.2016.0096]
124. Sharma A, Verbeke WJ. Improving diagnosis of depression with XGBOOST machine learning model and a large biomarkers Dutch dataset (= 11,081). Front Big Data 2020 Apr 30;3:15 [FREE Full text] [doi: 10.3389/fdata.2020.00015] [Medline: 33693389]
125. Yang H, Bath PA. Automatic prediction of depression in older age. In: Proceedings of the 3rd International Conference on Medical and Health Informatics. 2019 Presented at: ICMHI '19; May 17-19, 2019; Xiamen, China p. 36-44 URL: https://dl.acm.org/doi/10.1145/3340037.3340042 [doi: 10.1145/3340037.3340042]

126. Zhang W, Liu H, Silenzio VM, Qiu P, Gong W. Machine learning models for the prediction of postpartum depression: application and comparison based on a cohort study. JMIR Med Inform 2020 Apr 30;8(4):e15516 [FREE Full text] [doi: 10.2196/15516] [Medline: 32352387]
127. Annapureddy P, Hossain MF, Kissane T, Frydrychowicz W, Nitu P, Coelho J, et al. Predicting PTSD severity in veterans from self-reports for early intervention: a machine learning approach. In: Proceedings of the 21st International Conference on Information Reuse and Integration for Data Science. 2020 Presented at: IRI '20; August 11-13, 2020; Las Vegas, NV p. 201-208 URL: https://ieeexplore.ieee.org/document/9191375 [doi: 10.1109/iri49571.2020.00036]
128. Li B, Sano A. Early versus late modality fusion of deep wearable sensor features for personalized prediction of tomorrow's mood, health, and stress. Annu Int Conf IEEE Eng Med Biol Soc 2020 Jul;2020:5896-5899. [doi: 10.1109/EMBC44109.2020.9175463] [Medline: 33019316]
129. Novani NP, Arief L, Anjasmara R, Prihatmanto AS. Heart rate variability frequency domain for detection of mental stress using support vector machine. In: Proceedings of the 2018 International Conference on Information Technology Systems and Innovation. 2018 Presented at: ICITSI '18; October 22-26, 2018; Bandung, Indonesia p. 520-525 URL: https://ieeexplore.ieee.org/document/8695938 [doi: 10.1109/icitsi.2018.8695938]
130. Giannakakis G, Marias K, Tsiknakis M. A stress recognition system using HRV parameters and machine learning techniques. In: Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos. 2019 Presented at: ACIIW '19; September 3-6, 2019; Cambridge, UK p. 269-272 URL: https://ieeexplore.ieee.org/document/8925142 [doi: 10.1109/aciiw.2019.8925142]
131. Can YS, Chalabianloo N, Ekiz D, Ersoy C. Continuous stress detection using wearable sensors in real life: algorithmic programming contest case study. Sensors (Basel) 2019 Apr 18;19(8):1849 [FREE Full text] [doi: 10.3390/s19081849] [Medline: 31003456]
132. Coutts LV, Plans D, Brown AW, Collomosse J. Deep learning with wearable based heart rate variability for prediction of mental and general health. J Biomed Inform 2020 Dec;112:103610 [FREE Full text] [doi: 10.1016/j.jbi.2020.103610] [Medline: 33137470]
133. Cho D, Ham J, Oh J, Park J, Kim S, Lee N, et al. Detection of stress levels from biosignals measured in virtual reality environments using a kernel-based extreme learning machine. Sensors (Basel) 2017 Oct 24;17(10):2435. [doi: 10.3390/s17102435] [Medline: 29064457]
134. Pluntke U, Gerke S, Sridhar A, Weiss J, Michel B. Evaluation and classification of physical and psychological stress in firefighters using heart rate variability. In: Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2019 Presented at: EMBC '19; July 23-27, 2019; Berlin, Germany p. 2207-2212 URL: https://ieeexplore.ieee.org/document/8856596 [doi: 10.1109/embc.2019.8856596]
135. Lima R, Osório D, Gamboa H. Heart rate variability and electrodermal activity biosignal processing: predicting the autonomous nervous system response in mental stress. In: Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies. 2019 Presented at: BIOSTEC '19; February 22-24, 2019; Prague, Czech Republic p. 328-351 URL: https://link.springer.com/chapter/10.1007/978-3-030-46970-2_16 [doi: 10.1007/978-3-030-46970-2_16]
136. Montesinos V, Dell'Agnola F, Arza A, Aminifar A, Atienza D. Multi-modal acute stress recognition using off-the-shelf wearable devices. In: Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2019 Presented at: EMBC '19; July 23-27, 2019; Berlin, Germany p. 2196-2201 URL: https://ieeexplore.ieee.org/document/8857130 [doi: 10.1109/embc.2019.8857130]
137. Blum S, Jacobsen NS, Bleichner MG, Debener S. A riemannian modification of artifact subspace reconstruction for EEG artifact handling. Front Hum Neurosci 2019;13:141 [FREE Full text] [doi: 10.3389/fnhum.2019.00141] [Medline: 31105543]
138. Van de Schoot R. Latent growth mixture models to estimate PTSD trajectories. Eur J Psychotraumatol 2015 Mar 02;6(1):27503 [FREE Full text] [doi: 10.3402/ejpt.v6.27503] [Medline: 25735412]
139. Ram N, Grimm KJ. Growth mixture modeling: a method for identifying differences in longitudinal change among unobserved groups. Int J Behav Dev 2009 Oct 26;33(6):565-576 [FREE Full text] [doi: 10.1177/0165025409343765] [Medline: 23885133]
140. van Breada W, Pastor J, Hoogendoorn M, Ruwaard J, Asselbergs J, Riper H. Exploring and comparing machine learning approaches for predicting mood over time. In: Proceedings of the 2016 Conference on Innovation in Medicine and Healthcare. 2016 Presented at: InMed '16; June 15-17, 2016; Tenerife, Spain p. 37-47 URL: https://link.springer.com/chapter/10.1007/978-3-319-39687-3_4 [doi: 10.1007/978-3-319-39687-3_4]
141. Zhang P, Li F, Du L, Zhao R, Chen X, Yang T, et al. Psychological stress detection according to ECG using a deep learning model with attention mechanism. Appl Sci 2021 Mar 23;11(6):2848. [doi: 10.3390/app11062848]
142. Portilla R, Heintz B. Understanding dynamic time warping. The Databricks Blog. URL: https://www.databricks.com/blog/2019/04/30/understanding-dynamic-time-warping.html [accessed 2024-08-27]
143. Gui Q, Jin Z, Xu W. Exploring missing data prediction in medical monitoring: a performance analysis approach. In: Proceedings of the 2014 IEEE Signal Processing in Medicine and Biology Symposium. 2014 Presented at: SPMB '14; December 13, 2014; Philadelphia, PA p. 1-6 URL: https://ieeexplore.ieee.org/document/7002968 [doi: 10.1109/spmb.2014.7002968]

144. McDonald AD, Sasangohar F, Jatav A, Rao AH. Continuous monitoring and detection of post-traumatic stress disorder (PTSD) triggers among veterans: a supervised machine learning approach. IISE Trans Healthc Syst Eng 2019 Jun 11;9(3):201-211. [doi: 10.1080/24725579.2019.1583703]
145. Adheena MA, Sindhu N, Jerritta S. Physiological detection of anxiety. In: Proceedings of the 2018 International Conference on Circuits and Systems in Digital Enterprise Technology. 2018 Presented at: ICCSDET '18; December 21-22, 2018; Kottayam, India p. 1-5 URL: https://ieeexplore.ieee.org/document/8821162 [doi: 10.1109/iccsdet.2018.8821162]
146. Gerych W, Agu E, Rundensteiner E. Classifying depression in imbalanced datasets using an autoencoder- based anomaly detection approach. In: Proceedings of the IEEE 13th International Conference on Semantic Computing. 2019 Presented at: ICSC; January 30-February 1, 2019; Newport Beach, CA p. 124-127 URL: https://ieeexplore.ieee.org/document/8665535 [doi: 10.1109/icosc.2019.8665535]
147. Huysmans D, Smets E, Raedt W, Hoof C, Bogaerts K, Diest I, et al. Unsupervised learning for mental stress detection - exploration of self-organizing maps. In: Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies. 2018 Presented at: BIOSTEC '18; January 19-21, 2018; Funchal, Portugal p. 26-35 URL: https://www.scitepress.org/papers/2018/65411/65411.pdf [doi: 10.5220/0006541100260035]
148. Jain R, Xu W. Artificial Intelligence based wrapper for high dimensional feature selection. BMC Bioinformatics 2023 Oct 18;24(1):392. [doi: 10.1186/s12859-023-05502-x] [Medline: 37853338]
149. Braga-Neto U. Fundamentals of Pattern Recognition and Machine Learning. Cham, Switzerland: Springer; 2020.
150. Zhang Y, Wang S, Hermann A, Joly R, Pathak J. Development and validation of a machine learning algorithm for predicting the risk of postpartum depression among pregnant women. J Affect Disord 2021 Jan 15;279:1-8 [FREE Full text] [doi: 10.1016/j.jad.2020.09.113] [Medline: 33035748]
151. Deka D, Deka B. Characterization of heart rate variability signal for distinction of meditative and pre-meditative states. Biomed Signal Process Control 2021 Apr;66:102414. [doi: 10.1016/j.bspc.2021.102414]
152. Radovic M, Ghalwash M, Filipovic N, Obradovic Z. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. BMC Bioinformatics 2017 Jan 03;18(1):9 [FREE Full text] [doi: 10.1186/s12859-016-1423-9] [Medline: 28049413]
153. Xia L, Malik AS, Subhani AR. A physiological signal-based method for early mental-stress detection. Biomed Signal Process Control 2018 Sep;46:18-32. [doi: 10.1016/j.bspc.2018.06.004]
154. Wilhelm FH, Grossman P, Roth WT. Assessment of heart rate variability during alterations in stress: complex demodulation vs. spectral analysis. Biomed Sci Instrum 2005;41:346-351. [Medline: 15850130]
155. Quintero L, Papapetrou P, Muñoz JE, Fors U. Implementation of mobile-based real-time heart rate variability detection for personalized healthcare. In: Proceedings of the 2019 International Conference on Data Mining Workshops. 2019 Presented at: ICDMW '19; November 8-11, 2019; Beijing, China p. 838-846 URL: https://ieeexplore.ieee.org/document/8955523 [doi: 10.1109/icdmw.2019.00123]
156. Srividya M, Mohanavalli S, Bhalaji N. Behavioral modeling for mental health using machine learning algorithms. J Med Syst 2018 Apr 03;42(5):88. [doi: 10.1007/s10916-018-0934-5] [Medline: 29610979]
157. Alharahsheh YE, Abdullah MA. Predicting individuals mental health status in Kenya using machine learning methods. In: Proceedings of the 12th International Conference on Information and Communication Systems. 2021 Presented at: ICICS '21; May 24-26, 2021; Valencia, Spain p. 94-98 URL: https://ieeexplore.ieee.org/document/9464608 [doi: 10.1109/icics52457.2021.9464608]
158. Fuzzy C-means clustering. The MathWorks. URL: https://www.mathworks.com/help/fuzzy/fuzzy-c-means-clustering.html [accessed 2024-08-27]
159. Trevisan V. Using SHAP values to explain how your machine learning model works. Medium. URL: https://towardsdatascience.com/using-shap-values-to-explain-how-your-machine-learning-model-works-732b3f40e137 [accessed 2024-08-27]
160. Saeed A, Ozcelebi T, Lukkien J, van Erp JB, Trajanovski S. Model adaptation and personalization for physiological stress detection. In: Proceedings of the IEEE 5th International Conference on Data Science and Advanced Analytics. 2018 Presented at: DSAA '18; October 1-3, 2018; Turin, Italy p. 209-216 URL: https://ieeexplore.ieee.org/document/8631447 [doi: 10.1109/dsaa.2018.00031]
161. Vieira EU. You are underutilizing SHAP values — feature groups and correlations. Medium. URL: https://towardsdatascience.com/you-are-underutilizing-shap-values-feature-groups-and-correlations-8df1b136e2c2 [accessed 2024-08-27]
162. O'Sullivan C. Analysing interactions with SHAP. Medium. URL: https://towardsdatascience.com/analysing-interactions-with-shap-8c4a2bc11c2a [accessed 2024-08-27]

Abbreviations

ANN: artificial neural network
ASR: artifact subspace reconstruction
BP: blood pressure
CNN: convolutional neural network

DA: discriminant analysis
DL: deep learning
DT: decision tree
DTW: dynamic time warping
ECG: electrocardiogram
EEG: electroencephalogram
EMG: electromyogram
FCM: fuzzy C-means
HF: high frequency
HR: heart rate
HRV: heart rate variability
ICA: independent component analysis
KNN: k-nearest neighbor
LF: low frequency
LR: logistic regression
LSTM: long short-term memory
MD: mental health disorder
ML: machine learning
NB: naive Bayes
NN: neural network
PCA: principal component analysis
PPG: photoplethysmogram
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
PSD: power spectral density
PTSD: posttraumatic stress disorder
RF: random forest
RMSSD: mean square of successive RR interval differences
RNN: recurrent neural network
SVM: support vector machine
XGBoost: extreme gradient boosting

Edited by J Torous; submitted 16.10.23; peer-reviewed by J Keebler, I Nieto Alvarez; comments to author 27.11.23; revised version received 01.05.24; accepted 17.05.24; published 21.08.24.

Please cite as:
Razavi M, Ziyadidegan S, Mahmoudzadeh A, Kazeminasab S, Baharlouei E, Janfaza V, Jahromi R, Sasangohar F
Machine Learning, Deep Learning, and Data Preprocessing Techniques for Detecting, Predicting, and Monitoring Stress and Stress-Related Mental Disorders: Scoping Review
JMIR Ment Health 2024;11:e53714
URL: https://mental.jmir.org/2024/1/e53714
doi:10.2196/53714
PMID:

Review

Efficacy of eHealth Versus In-Person Cognitive Behavioral Therapy for Insomnia: Systematic Review and Meta-Analysis of Equivalence

Sofie Møgelberg Knutzen[1], MSc; Dinne Skjærlund Christensen[1], PhD; Patrick Cairns[1], MSc; Malene Flensborg Damholdt[1], PhD; Ali Amidi[1], PhD; Robert Zachariae[1,2], DMSc

[1]Department of Psychology and Behavioral Sciences, Aarhus University, Aarhus, Denmark

[2]Department of Oncology, Aarhus University Hospital, Aarhus, Denmark

Corresponding Author:
Robert Zachariae, DMSc
Department of Psychology and Behavioral Sciences
Aarhus University
Bartholins Alle 11
Bld. 1350
Aarhus, DK8000C
Denmark
Phone: 45 24235356
Email: bzach@rm.dk

Abstract

Background: Insomnia is a prevalent condition with significant health, societal, and economic impacts. Cognitive behavioral therapy for insomnia (CBTI) is recommended as the first-line treatment. With limited accessibility to in-person–delivered CBTI (ipCBTI), electronically delivered eHealth CBTI (eCBTI), ranging from telephone- and videoconference-delivered interventions to fully automated web-based programs and mobile apps, has emerged as an alternative. However, the relative efficacy of eCBTI compared to ipCBTI has not been conclusively determined.

Objective: This study aims to test the comparability of eCBTI and ipCBTI through a systematic review and meta-analysis of equivalence based on randomized controlled trials directly comparing the 2 delivery formats.

Methods: A comprehensive search across multiple databases was conducted, leading to the identification and analysis of 15 unique randomized head-to-head comparisons of ipCBTI and eCBTI. Data on sleep and nonsleep outcomes were extracted and subjected to both conventional meta-analytical methods and equivalence testing based on predetermined equivalence margins derived from previously suggested minimal important differences. Supplementary Bayesian analyses were conducted to determine the strength of the available evidence.

Results: The meta-analysis included 15 studies with a total of 1083 participants. Conventional comparisons generally favored ipCBTI. However, the effect sizes were small, and the 2 delivery formats were statistically significantly equivalent (P<.05) for most sleep and nonsleep outcomes. Additional within-group analyses showed that both formats led to statistically significant improvements (P<.05) in insomnia severity; sleep quality; and secondary outcomes such as fatigue, anxiety, and depression. Heterogeneity analyses highlighted the role of treatment duration and dropout rates as potential moderators of the differences in treatment efficacy.

Conclusions: eCBTI and ipCBTI were found to be statistically significantly equivalent for treating insomnia for most examined outcomes, indicating eCBTI as a clinically relevant alternative to ipCBTI. This supports the expansion of eCBTI as a viable option to increase accessibility to effective insomnia treatment. Nonetheless, further research is needed to address the limitations noted, including the high risk of bias in some studies and the potential impact of treatment duration and dropout rates on efficacy.

Trial Registration: PROSPERO CRD42023390811; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=390811

(JMIR Ment Health 2024;11:e58217) doi:10.2196/58217

KEYWORDS

sleep disturbance; digital; telehealth; face-to-face; head-to-head comparison; CBTI; cognitive behavioral therapy for insomnia; mobile phone

Introduction

Background

Insomnia, characterized by difficulties initiating or maintaining sleep, which are perceived as distressing and result in significant impairment of daytime functioning, is a common concern in the general population [1]. It is estimated that approximately 20% of the population experience episodic symptoms of insomnia, resulting in negative consequences for daytime functioning, for example, fatigue, with approximately 10% fulfilling the diagnostic criteria for an insomnia disorder [1]. The association between insomnia and adverse physical and mental health outcomes has been thoroughly documented, with numerous prospective studies showing increased risk of developing cardiovascular disease [2,3], infectious diseases such as the common cold or pneumonia [4,5], all-cause dementia [6,7], mental disorders such as depression and anxiety [8], and social withdrawal and loneliness [9,10]. In addition, not only short but also long sleep duration, both possible indicators of sleep disturbances, have been associated with increased mortality [11,12]. Beyond the personal health implications, insomnia is associated with societal costs through increased health care use, higher levels of work absenteeism, diminished work-related productivity, reduced learning capacity, and poorer academic performance [13-15]. This underlines the extensive societal and economic burdens posed by untreated sleep disturbances.

While hypnotic medications are commonly used to treat insomnia, they are not recommended for long-term use due to the risk of developing tolerance and dependence [16] as well as a wide range of adverse consequences, including daytime drowsiness, impaired cognitive function, increased risk of accidents or falls, and rebound insomnia upon discontinuation [17,18]. Instead, the major sleep medicine and research organizations recommend cognitive behavioral therapy for insomnia (CBTI) as the first-line treatment for insomnia [19-21]. CBTI usually involves a combination of two or more of the following five components [17]: (1) *sleep restriction therapy*, which aims at promoting more efficient and consolidated sleep patterns by first reducing the time spent awake in bed and then gradually allowing the person to increase time in bed [22]; (2) *stimulus-control therapy*, which aims to strengthen the connection between the bed and sleep by associating the bed and bedroom with sleep rather than wakefulness [23]; (3) *relaxation techniques*, which aim to reduce stress, anxiety, and tension that may interfere with falling asleep or staying asleep [24]; (4) *cognitive therapy*, which targets negative thought patterns and maladaptive beliefs about sleep [25]; and (5) *sleep hygiene education*, which focuses on establishing healthy habits and optimizing the sleep environment to promote better sleep [26]. Several meta-analyses have supported the efficacy of CBTI, demonstrating both short-term [16,27] and long-term effects [28] on insomnia, not only as the primary problem but also as a comorbid condition, for example, in patients with chronic pain [29] and survivors of cancer [30]. Compared with pharmacotherapy, CBTI has been found to be at least as effective in reducing insomnia symptoms and generally demonstrates more durable effects [31].

Nonetheless, substantial challenges remain in extending assistance to those affected. Individuals with insomnia rarely receive guideline-compliant treatment, hindered by various obstacles. These include insufficient numbers of trained CBTI providers, low rates of referral by primary care physicians, and geographical and physical barriers that deter patients from receiving help [32-34]. To address these challenges, several alternative eHealth delivery formats of CBTI have been developed and evaluated [32]. These alternatives include telephone- and videoconference-delivered CBTI and fully automated web-based programs and mobile apps.

Recently published meta-analyses have revealed statistically significant and clinically meaningful effects of eHealth CBTI (eCBTI) on various measures, including insomnia severity; self-reported sleep quality; and sleep diary–based outcomes such as sleep onset latency (SOL), wake after sleep onset (WASO), total sleep time (TST), and sleep efficiency [35,36]. This efficacy extends not only to individuals with insomnia as their primary concern [37,38] but also to those with comorbid insomnia, for example, survivors of cancer [39]. However, the results of recent systematic reviews and network meta-analyses comparing various delivery formats of CBTI suggest that in-person–delivered CBTI (ipCBTI) is generally superior to eCBTI, more so for insomnia severity than for sleep diary outcomes [40,41]. In contrast, a network meta-analysis investigating a Food and Drug Administration–authorized prescription eCBTI compared to traditional ipCBTI found that eCBTI was the most efficacious regarding insomnia severity [42].

The inconclusive results of the existing meta-analyses could be due to their reliance on both direct and indirect comparisons and variations in treatment length, dosage, content, and control group types across studies of both formats, which may compromise comparability [41]. To date, no meta-analysis has focused exclusively on randomized controlled trials, conducting direct head-to-head comparisons of eCBTI and ipCBTI, and it thus remains unclear how well the 2 delivery formats compare in terms of efficacy.

In addition, when examining the equivalence or nonequivalence of 2 interventions with meta-analysis, the conventional nonsuperiority null hypothesis test procedure is insufficient. Here, a nonsignificant result merely indicates a failure to reject the null hypothesis of no difference, which cannot conclusively determine nonequivalence or equivalence [43]. To truly test whether treatments are equivalent, we must reject the null hypothesis of nonequivalence, which means that differences in effect sizes (ESs) are as large as or larger than a predetermined equivalence margin [44]. One will usually choose equivalence margins based on previously determined minimal important differences (MIDs), referring to the minimal difference in an outcome of interest that can be viewed as clinically meaningful [45].

Objectives

Given the prevalence of insomnia and the need for diverse treatment approaches, establishing the equivalence or nonequivalence of digital and traditional CBTI delivery formats is crucial. The possible equivalence or nonequivalence of eCBTI and ipCBTI has not yet been subjected to meta-analysis. The aim of this study was, therefore, to test the comparability of eCBTI and ipCBTI with a systematic review and meta-analysis of equivalence based on randomized controlled trials directly comparing the 2 delivery formats.

Methods

This study was registered with PROSPERO (registration CRD42023390811) and conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) statement (Multimedia Appendix 1) [46].

Search Strategy

The electronic databases of CINAHL, Cochrane, Embase, PsycINFO, and PubMed were searched for publications from the earliest time available until January 5, 2024. Keywords related to insomnia (eg, sleep disturbance OR sleep disorder) were combined with keywords related to CBTI (eg, cognitive behavioral OR CBT) and keywords pertaining to eHealth (eg, telehealth OR digital). The search strings were constructed in collaboration with a skilled librarian. Table S1 in Multimedia Appendix 2 comprises detailed search strings for each database. The electronic database searches were supplemented with backward searches of reference lists of included studies. No separate protocol in addition to the one registered with PROSPERO was prepared. The main methodological changes to the original registered protocol were as follows: (1) the search date was changed from 1991 to the earliest time available due to the inclusion of additional electronic delivery formats, for example, telephone-based interventions, and (2) we also extracted data on secondary nonsleep outcomes of fatigue, anxiety, and depression.

Inclusion and Exclusion Criteria

On the basis of the Population, Intervention, Comparison, and Outcome (PICO) approach [47], the inclusion criteria in Textbox 1 were used.

Textbox 1. Inclusion criteria.

Population

- Adults and adolescents (aged ≥12 years) with (1) self-reported poor sleep quality or symptoms of insomnia assessed with relevant instruments, for example, the Insomnia Severity Index (ISI) [48] or the Pittsburgh Sleep Quality Index (PSQI) [49], or (2) an insomnia diagnosis established by a structured clinical interview. Studies of children aged <12 years and studies focusing on other medical sleep disorders (eg, sleep apnea and narcolepsy) were excluded. No exclusions were made based on comorbid disorders.

Intervention

- Electronically delivered eHealth cognitive behavioral therapy for insomnia (eCBTI), defined as cognitive behavioral therapy for insomnia (CBTI) delivered remotely or using digital means without in-person contact, for example, CBTI delivered via telephone or video, web-based CBTI, or smartphone-based CBTI. CBTI was defined as any combination of ≥2 of the standard CBTI components, that is, sleep restriction therapy, stimulus-control therapy, relaxation, cognitive therapy, and sleep hygiene education. Other eHealth interventions aimed at treating insomnia, for example, mindfulness-based interventions, were excluded, as were stand-alone CBTI components.

Comparison

- eCBTI had to be directly compared with in-person–delivered CBTI (ipCBTI), defined as any combination of ≥2 standard CBTI components delivered in person, either individually or in a group format. Other in-person–delivered interventions aimed at treating insomnia, including stand-alone CBTI components, were excluded.

Outcomes

- Studies should report means with SDs or SEs; change scores; effect sizes (eg, Cohen *d*) or data that could be converted into an effect size for at least 1 relevant sleep outcome, that is, insomnia severity or clinically significant sleep disturbance assessed with relevant scales such as the ISI [48] and the PSQI [49]; structured clinical interviews; or a relevant sleep parameter assessed with a sleep diary, actigraphy, or polysomnography. Only randomized controlled trials published in English in peer-reviewed journals were included. Case studies, open trials, and other nonrandomized controlled trials were excluded, together with studies with sample sizes <10.

Study Selection and Data Extraction

Identified references were imported into the web-based software program Covidence (Veritas Health Innovation) [50]. After duplicate removal, title and abstract screening was performed, followed by full-text screening. One author (SMK) conducted the final search, with 3 authors (SMK, DSC, and PC) conducting the screening process independently. Discrepancies were resolved through discussions and, in case of disagreement, by including a fourth author (RZ). The primary outcome was total sleep disturbance calculated as the combined, that is, averaged, results for insomnia severity and sleep quality assessed with validated scales, for example, the Insomnia Severity Index (ISI), the Pittsburgh Sleep Quality Index (PSQI), or similar scales. Secondary sleep outcomes were insomnia severity measured with the ISI; sleep quality measured with the PSQI; and the sleep diary or actigraphy-based outcomes of SOL, WASO, TST, and sleep efficiency calculated as TST relative to time in bed. In addition, we extracted data on the secondary nonsleep outcomes of fatigue, anxiety, and depression, as well as for study characteristics that could potentially explain (moderate) any variations in the differences between ipCBTI and eCBTI, including mean sample age, the proportion of women in the sample, study dropout rates, the type and degree of therapist contact, the number of treatment sessions, treatment length, and the type and number of CBTI components in each condition. A total of 3 authors (SMK, DSC, and PC) extracted data from the included studies independently, and discrepancies were resolved through discussion and by including a fourth author (RZ).

Risk of Bias Assessment

The revised Cochrane Risk of Bias tool [51] was used to evaluate the risk of bias in the included studies. Five sources of bias were assessed: (1) bias arising from the randomization process, (2) bias due to deviations from intended interventions, (3) bias due to missing outcome data, (4) bias in the measurement of the outcome, and (5) bias in the selection of the reported result. All studies were evaluated for each of the 5 potential sources of bias and rated as either low risk, high risk, or some concerns on the primary outcome of sleep disturbance. In addition, an overall assessment of the risk of bias was conducted for each study. As the number of dropouts in studies investigating eCBTI is generally high, with mean attrition rates ranging from 22% to 25% [36,52], it was decided to use a less conservative criterion in domain 3. We thus considered the availability of data from ≥90% of the participants at the postintervention assessment time-point sufficient. The assessments were conducted independently by 3 authors (SMK, DSC, and PC). Disagreements were solved by negotiation.

Data Analysis

Hedges' *g*, a variation of Cohen *d* correcting for a possible bias due to a small sample size [53], was used as the standardized ES. All ES calculations were based on differences between the ipCBTI and eCBTI intervention groups in changes (means and SDs) from preintervention to postintervention time points and from preintervention to follow-up time points, standardized by change score SDs. If the relevant data were not reported, we contacted the authors, requesting them to provide this information. We also analyzed the mean differences across the different sleep-related outcomes, that is, mean differences in ISI and PSQI scores; percentages for sleep efficiency; and minutes for SOL, WASO, and TST. ESs were pooled using the inverse variance method, taking the precision of each study into account. A random-effects model was used in all analyses, with positive ESs indicating ipCBTI being more efficacious than eCBTI. If studies reported results for >1 measure per outcome, for example, insomnia severity or sleep quality, we chose the most commonly used outcome measure, that is, the ISI for insomnia severity and the PSQI for sleep quality, so that only 1 result per study was used in each data synthesis, ensuring the independence of results.

Differences between ipCBTI and eCBTI were first analyzed using a conventional random-effects test of superiority for results at both postintervention and follow-up. The pooled ESs were then subjected to analyses of equivalence [44], testing whether the CI fell within an equivalence interval based on the clinical significance thresholds (or MIDs) previously suggested for the various sleep outcome measures. Thus, the MIDs were 0.5 SD for the standardized mean differences of sleep disturbance, insomnia severity, sleep quality, sleep efficiency, SOL, WASO, and TST, as suggested in a previous study [54]. The MIDs for the mean differences were 4.4 points on the PSQI [55], 5% for sleep efficiency, 10 minutes for SOL, and 15 minutes for WASO and TST [54]. The 6-point MID previously suggested for the ISI [56] was based on an analysis of within-subject improvement, that is, minimal important change. We therefore used 0.5 × the average SD of 4.2 (2.6 points) for ISI at baseline in patients with insomnia reported in the original validation paper [48]. This SD corresponds well with the average SD of 4.1 found for ISI scores across studies at baseline in this review. The equivalence interval of SD 0.25 for depression was chosen based on the MID previously suggested [57]. As no specific MIDs were available for the measures of fatigue and anxiety, 0.5 SD was chosen as the equivalence interval for these measures [58]. The equivalence test is based on two 1-sided tests, with the 2 interventions considered to be statistically significantly equivalent if the largest of the 2 *P* values is <.05 [44].

Heterogeneity was explored by calculating the I^2 statistic [59,60]. In addition, we calculated the 95% prediction interval, that is, the interval in which 95% of future observations from the same family of studies are expected to fall [61]. Possible reasons for heterogeneity of the differences between ipCBTI and eCBTI were explored with moderator analyses comparing the ESs of studies according to the following study characteristics: mean sample age, the proportion of women in the sample (%), overall study dropout (%), the difference in dropout between eCBTI and ipCBTI (%), therapist contact in the eCBTI condition (reference: fully automated, ie, no direct or indirect therapist contact), the number of treatment sessions, and treatment duration (weeks). In addition, we explored the possible role of the number of CBTI components used in ipCBTI and eCBTI, respectively. Both categorical and continuous moderators were analyzed with meta-regression when K (the number of studies in the analysis) was ≥10.

When K was ≥10, the possibility of publication bias was evaluated with funnel plots and the method developed by Egger et al [62]. If the results were suggestive of publication bias, we planned to calculate an adjusted ES using the Duval and Tweedie trim and fill method [63]. The calculations were conducted with Comprehensive Meta-Analysis (version 4; Biostat, Inc) [64] and various formulas in Microsoft Excel.

Finally, to assess the potential efficacy of each condition, we calculated the pooled within-group differences for each condition at postintervention and follow-up for all outcomes.

Supplementary Bayesian Analyses

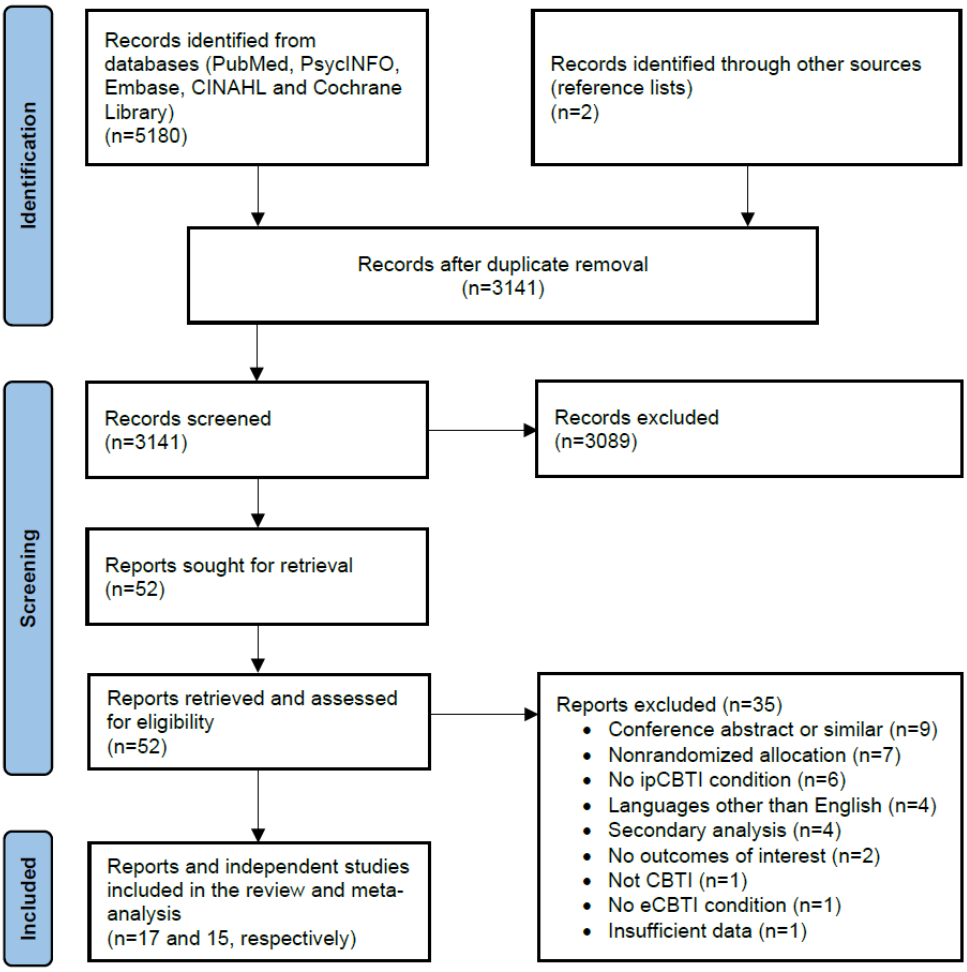
To aid the interpretation of the results, we conducted a supplementary Bayesian model-averaged meta-analysis [65] of the overall comparisons of ipCBTI and eCBTI at postintervention and follow-up. The procedure examined the results of four models: (1) the fixed-effect null hypothesis, that is, the difference between ipCBTI and eCBTI is nonzero (fH_0), (2) the fixed-effect alternative hypothesis, that is, the difference is zero (fH_1), (3) the random-effects null hypothesis (rH_0), and (4) the random-effects alternative hypothesis (rH_1). The Bayesian model-averaged analysis thus avoids selecting either a fixed- or random-effects model and addresses 2 questions in light of the observed data: What is the plausibility that the overall effect is 0, that is, equivalent? and Is there a between-study variability in the ESs? Bayesian methods enable direct probability statements about the hypotheses themselves and avoid other issues associated with null hypothesis significance testing, such as the overreliance on relatively arbitrary *P* value thresholds and the dichotomization of results into "significant" and "nonsignificant" [66]. We chose an uninformed prior probability, that is, 25%, of the 4 models and 2000 iterations. Concerning parameter distributions, we chose previously recommended defaults [65], using a 0-centered Cauchy prior with a scale of 0.707 for the ES and an empirically informed prior distribution of nonzero between-study deviation estimates from 705 meta-analyses [67]. This distribution has been approximated by an inverse-gamma (1, 0.15) prior on the SD (Tau) [65]. The Bayesian analyses were conducted with the computer software JASP (version 16; University of Amsterdam) [68]. All data included in this review are available in tables and figures in the manuscript or Multimedia Appendix 2.

Results

Study Selection

A total of 5180 records were identified via databases, and 2 additional records were identified via reference lists. After 2039 (39.36%) duplicates were removed, 3141 (60.64%) references were screened by title and abstract. Full-text screening was carried out for 1.66% (52/3141) records, and after assessing eligibility, we identified 32% (17/52) full-text reports of 15 unique randomized head-to-head comparisons of ipCBTI and eCBTI. The results of the study selection process are shown in Figure 1. A list of studies excluded after full text screening with reasons for exclusion is provided in Table S2 in Multimedia Appendix 2.

Figure 1. Study selection flowchart. CBTI: cognitive behavioral therapy for insomnia; eCBTI: eHealth cognitive behavioral therapy for insomnia; ipCBTI: in-person–delivered cognitive behavioral therapy for insomnia.



Study Characteristics

The characteristics of the included studies are summarized in Table 1. Of the 15 studies, most studies were conducted in the United States (K=5, 33%), followed by Canada (K=3, 20%) and the Netherlands (K=2, 13%). The rest of the studies (K=5, 33%) were all conducted in different countries. A total of 1083 participants were included in the 15 studies and randomized to eCBTI and ipCBTI treatment conditions. A little more than half, 666/1083 (61.5%), were women. The mean age of the total sample was 40.6 (SD 11.3) years, with mean sample ages ranging from 15.5 to 55.1 years. Of the 15 studies, most focused on insomnia as the primary problem (K=11, 73%), and 4 studies (27%) focused on patients with comorbid insomnia, that is, people maintaining sobriety from alcohol (K=1, 25%), survivors of breast cancer (K=1, 25%), and patients with posttraumatic stress disorder (K=2, 50%). All studies had insomnia as an inclusion criterion, based on diagnostic criteria; validated questionnaires; or quantitative criteria such as SOL, WASO, or early morning awakenings of ≥30 minutes occurring at least 3 nights a week, or sleep efficiency <85%.

Table 1. Study characteristics.

Study, year	Country	Participant characteristics, insomnia type, and sample type	Demographic characteristics: mean age, (SD); women, n/N (%)	eCBTI[a] and ipCBTI[b] participants, n/N (%)	eCBTI treatment format, delivery mode, therapist contact, and the number of sessions	ipCBTI treatment format and the number of sessions	Sleep outcomes	Secondary outcomes	Time to postintervention and time to follow-up (weeks)	eCBTI and ipCBTI study dropout (%)[c]; treatment dropout (%)[d]	eCBTI and ipCBTI components, n	ROB-2[e] (low risk, some concerns, or high risk)
Currie et al [69], 2004	Canada	Adults, comorbid insomnia (people maintaining sobriety from alcohol), and clinical sample	43.3 (10.9); 18/60 (30)	28/57 (49) and 29/57 (51)	Individual, self-help and telephone, synchronous, and 5 sessions	Individual and 5 sessions	Insomnia severity, SQ[f], and sleep diary (SOL[g], WASO[h], TST[i], and sleep efficiency)	Depression	0 and 26	50% and 35%; NR[j] and NR	5 and 5	Some concerns
Bastien et al [70], 2004	Canada	Adults, primary insomnia, and community sample	42.8 (9.7); 18/29 (62)	14/29 (48) and 15/29 (52)	Individual, telephone, synchronous, and 8 sessions	Individual and 8 sessions	Insomnia severity and sleep diary (SOL, WASO, TST, TIB[k], and sleep efficiency)	Anxiety and depression	0 and 26	36% and 53%; 0% and 0%	4 and 4	High risk
Savard et al [71], 2014, and Savard et al [72], 2016	Canada	Adults, comorbid insomnia (breast cancer), and clinical sample	53.9, (8.8); 161/161 (100)	80/161 (49.7) and 81/161 (50.3)	Individual, video-based self-help with phone support, synchronous, and 6 sessions	Individual and 6 sessions	Insomnia severity and sleep diary (SOL, WASO, TST, and sleep efficiency)	Fatigue, anxiety, and depression	0 and 52	39% and 26%; NR and NR	4 and 4	High risk
Blom et al [73], 2015	Sweden	Adults, primary insomnia, and community sample	54.4 (13.8); 23/48 (48)	24/48 (50) and 24/48 (50)	Individual, web based with written feedback, asynchronous, and 8 sessions	Group and 8 sessions	Insomnia severity and sleep diary (SOL, TST, and sleep efficiency)	Depression	0 and 26	46% and 46%; 29% and 17%	5 and 5	Some concerns

Study, year	Country	Participant characteristics, insomnia type, and sample type	Demographic characteristics: mean age, (SD); women, n/N (%)	eCBTI[a] and ipCBTI[b] participants, n/N (%)	eCBTI treatment format, delivery mode, therapist contact, and the number of sessions	ipCBTI treatment format and the number of sessions	Sleep outcomes	Secondary outcomes	Time to postintervention and time to follow-up (weeks)	eCBTI and ipCBTI study dropout (%)[c]; treatment dropout (%)[d]	eCBTI and ipCBTI components, n	ROB-2[e] (low risk, some concerns, or high risk)
de Bruin et al [74], 2015, and de Bruin et al [75], 2018	The Netherlands	Adolescents, primary insomnia, and community sample	15.4 (1.6); 59/77 (77)	39/77 (51) and 38/77 (49)	Individual, web based with written feedback and chat, mixed, and 6 sessions	Group and 6 sessions	Insomnia severity and sleep diary and actigraphy (SOL, WASO, TST, TIB, and sleep efficiency)	Anxiety and depression	0 and 9-52	18% and 5%; 0 and 0	5 and 5	Some concerns
Lancee et al [76], 2016	The Netherlands	Adults, primary insomnia, and community sample	39.9 (13.6); 48/60 (80)	30/60 (50) and 30/60 (50)	Individual, web based with email feedback, asynchronous, and 6 sessions	Individual and 6 sessions	Insomnia severity and sleep diary (TST and sleep efficiency)	Anxiety and depression	4 and 26	30% and 13%; 23% and 7%	5 and 5	High risk
Taylor et al [77], 2017	United States	Adults, primary insomnia, and sample of army personnel	32.7 (7.4); 13/67 (19)	34/67 (51) and 33/67 (49)	Individual, web based (automated), none, and 6 sessions	Individual and 6 sessions	Insomnia severity and sleep diary and actigraphy (SOL, WASO, TST, and sleep efficiency)	—[l]	0 and N/A[m]	41% and 58%; 21% and 12%	5 and 5	High risk
Laurel Franklin et al [78], 2017	United States	Adults, comorbid insomnia (PTSD[n]), and clinical sample	53.8 (12.0); 0/18 (0)	11/18 (61) and 7/18 (39)	Individual, telephone, synchronous, and 6 sessions	Individual and 6 sessions	SQ	—	0 and 13	46% and 14%; 18% and 0%	5 and 5	High risk
Gieselmann and Pietrowsky [79], 2019	Germany	Adults, primary insomnia, and community sample	39.5, (13.1); 26/50 (52)	23/50 (46) and 27/50 (54)	Individual, chat, synchronous, and 3 sessions	Individual and 3 sessions	SQ and sleep diary and actigraphy (SOL, TST, and sleep efficiency)	Fatigue, anxiety, and depression	0 and 9	4% and 34%; NR and NR	2 and 2	Some concerns

Study, year	Country	Participant characteristics, insomnia type, and sample type	Demographic characteristics: mean age, (SD); women, n/N (%)	eCBTI[a] and ipCBTI[b] participants, n/N (%)	eCBTI treatment format, delivery mode, therapist contact, and the number of sessions	ipCBTI treatment format and the number of sessions	Sleep outcomes	Secondary outcomes	Time to postintervention and time to follow-up (weeks)	eCBTI and ipCBTI study dropout (%)[c]; treatment dropout (%)[d]	eCBTI and ipCBTI components, n	ROB-2[e] (low risk, some concerns, or high risk)
Gehrman et al [80], 2020	United States	Adults, comorbid insomnia (PTSD), and clinical sample	55.1 (12.2); 9/95 (10)	49/96 (51) and 47/96 (49)	Group, video conferencing, synchronous, and 6 sessions	Group and 6 sessions	Insomnia severity and SQ	—	2 and 13	63% and 47%; 29% and 26%	5 and 5	Some concerns
Arnedt et al [81], 2021	United States	Adults, primary insomnia, and Community sample	47.2 (16.3); 46/65 (71)	33/65 (51) and 32/65 (49)	Individual, video conferencing, synchronous, and 6 sessions	Individual and 6 sessions	Insomnia severity and sleep diary (SOL, WASO, TST, and sleep efficiency)	Fatigue, anxiety, and depression	0 and 13	NR and NR; 6% and 3%	5 and 5	Some concerns
Gehrman et al [82], 2021	United States	Adults, primary insomnia, and community sample	33.4 (10.3); 26/41 (63)	21/41 (51) and 20/41 (49)	Individual, video conferencing, synchronous, and 8 sessions	Individual and 8 sessions	Insomnia severity	Fatigue, anxiety, and depression	3 and 13	24% and 30%; 19% and 30%	5 and 5	Some concerns
Kallestad et al [83], 2021	Norway	Adults, primary insomnia, and clinical sample	41.3 (11.6); 76/101 (75.2)	49/101 (48.5) and 52/101 (51.5)	Individual, web based (automated), none, and 6 sessions	Individual and 8 sessions	Insomnia severity and sleep diary (SOL, WASO, TST, and sleep efficiency)	Fatigue	0 and 26	16% and 8%; 12% and 0%	4 and 4	Some concerns
Wong et al [84], 2021	Hong Kong	Adolescents or older participants, primary insomnia, and community sample	37.6 (15.3); 90/140 (64.3)	70/140 (50) and 70/140 (50)	Individual, web-based self-help, none, and 4 sessions	Group and 1 session (workshop)	Insomnia severity	Anxiety and depression	4 and 12	26% and 30%; NR and 49%	5 and 5	Some concerns

Study, year	Country	Participant characteristics, insomnia type, and sample type	Demographic characteristics: mean age, (SD); women, n/N (%)	eCBTI[a] and ipCBTI[b] participants, n/N (%)	eCBTI treatment format, delivery mode, therapist contact, and the number of sessions	ipCBTI treatment format and the number of sessions	Sleep outcomes	Secondary outcomes	Time to postintervention and time to follow-up (weeks)	eCBTI and ipCBTI study dropout (%)[c]; treatment dropout (%)[d]	eCBTI and ipCBTI components, n	ROB-2[e] (low risk, some concerns, or high risk)
Chan et al [85], 2022	China	Youth, primary insomnia, and community sample	20.2 (2.4); 61/90 (69)	45/90 (50) and 45/90 (50)	Individual, email self-help, none, and 8 sessions	Group, and 8 sessions	Insomnia severity, SQ, and sleep diary (SOL, WASO, TST, TIB, and sleep efficiency)	Fatigue, anxiety, and depression	1 and 26	47% and 22%; 38% and 4%	5 and 4	High risk

[a]eCBTI: eHealth cognitive behavioral therapy for insomnia.

[b]ipCBTI: in-person–delivered cognitive behavioral therapy for insomnia.

[c]Study dropout: proportion of participants lost to follow-up at the most distant time point after baseline.

[d]Treatment dropout: proportion of participants who dropped out of treatment (defined as completing <50% of treatment cores or sessions).

[e]ROB 2: Cochrane Risk of Bias.

[f]SQ: sleep quality.

[g]SOL: sleep onset latency.

[h]WASO: wake after sleep onset.

[i]TST: total sleep time.

[j]NR: not reported.

[k]TIB: time in bed.

[l]No data.

[m]N/A: not applicable.

[n]PTSD: posttraumatic stress disorder.

Sleep outcomes reported in the 15 included studies were insomnia severity (K=13, 87%), with the ISI being the most frequently used instrument (K=11, 85%), and sleep quality (K=5, 33%), with the PSQI being used by all studies reporting this outcome. Sleep diaries and actigraphy were used in 11 (73%) and 3 (20%) studies, respectively, assessing sleep parameters such as SOL, WASO, TST, time in bed, and sleep efficiency. A total of 11 (73%) studies assessed depression, with the Hospital Anxiety and Depression Scale (HADS) [86] (K=3, 27%) being the most frequently used, followed by the Beck Depression Inventory (BDI) [87] (K=2, 18%), the Patient Health Questionnaire (PHQ) [88] (K=2, 18%), and the Center for Epidemiologic Studies Depression Scale (CES-D) [89] (K=2, 18%). Of the 15 studies, 9 (60%) studies assessed anxiety, with most (K=4, 44%) using the HADS, followed by the General Anxiety Disorder-7 (GAD-7) [90] (K=2, 22%), and 7 (47%) studies assessed fatigue, with most frequently using the Multidimensional Fatigue Inventory (MFI) [91] (K=4, 57%). A total of 14 studies reported follow-up data, with time to follow-up ranging from 9 to 52 weeks.

In all but 1 (7%) of the 15 studies, eCBTI was delivered individually. In most studies (K=10, 67%), eCBTI involved some degree of interaction with a treatment provider, with real-time, synchronous therapist contact being available in 8 (53%) studies and asynchronous support, for example, via email, being offered in 2 (13%) studies. One (7%) study provided both synchronous and asynchronous therapist contact. In 4 (27%) studies, eCBTI was provided completely without interaction with a treatment provider, for example, in a fully automated format. ipCBTI was primarily delivered individually (K=10, 67%), with 5 (33%) studies using a group format.

Risk of Bias

The risk of bias in the individual studies is summarized in Figure S1 in Multimedia Appendix 2. Of the 15 studies, 10 (67%) were characterized as having some concerns regarding the overall risk of bias, and the remaining 5 (33%) were classified as having a high risk of bias overall. No studies were characterized as having a low risk of bias. The reasons for a study being categorized as having a high risk of bias stemmed primarily from "bias due to missing outcome data" due to the combination of high rates of missing outcome data and failure to include analyses correcting for this, for example, sensitivity analyses. Bias raising "some concerns" primarily stemmed from "bias in the measurement of the outcome" due to the combination of using a self-reported outcome and nonblinding. Only few studies had attempted some element of blinding. Of the 15 studies, 3

(20%) [79,81,84] reported that participants were kept blind to study hypotheses, 1 (7%) reported that treatment providers were kept blind to study hypotheses [79], and 2 (13%) reported that the data analyst was blinded to the allocation status of the participants [83,84]. In addition, bias raising “some concerns” stemmed from “bias in the selection of the reported result” due to inadequate preregistration of the analytical strategy.

Comparing Intervention Characteristics of ipCBTI and eCBTI

More participants had dropped out of eCBTI than out of ipCBTI at both postintervention (104/528, 19.7% vs 72/526, 13.7%) and follow-up (176/509, 34.6% vs 145/509, 28.5%). However, the differences did not reach statistical significance (*P*=.17 and .46). No between-condition differences were found in the mean number of sessions (6.0 vs 6.1; *P*=.91), the duration of the intervention (6.7 weeks vs 6.6 weeks; *P*=.83), or the number of CBTI components (4.5 vs 4.6; *P*=.83).

Within-Group Effects

As presented in Table S3 in Multimedia Appendix 2, statistically significant (*P*<.05) improvements from preintervention to postintervention time points and from preintervention to follow-up time points were observed for both ipCBTI and eCBTI for all self-reported sleep outcomes. At postintervention time points, the ESs (Hedges' *g*) ranged from 0.27 (TST) to 1.97 (ISI) for ipCBTI and from 0.23 (TST) to 1.36 (ISI) for eCBTI. Similarly, at follow-up, the ESs ranged from 0.43 (TST) to 1.88 (ISI) for ipCBTI and from 0.39 (TST) to 1.41 (total sleep disturbance) for eCBTI. For the few actigraphy-based sleep outcomes at postintervention (K=3, 20%), only the results for SOL in the ipCBTI condition (Hedges' *g*=0.53; mean difference=–11.5 minutes) reached statistical significance. At postintervention time points, in the ipCBTI condition, the ISI total score was, on average, improved by 9.0 points; the PSQI global score by 4.4 points; diary-based sleep efficiency by 12.1%; and diary-based SOL, WASO, and TST by –20.9, –23.5, and +21.3 minutes, respectively. The comparable results for eCBTI were 7.1 points, 3.5 points, 10.3%, –19.6 minutes, –19.5 minutes, and +16.3 minutes, respectively. As also seen in Table S3 in Multimedia Appendix 2, the within-participant improvements in fatigue, anxiety, and depression were all statistically significant and similar for both delivery formats at postintervention. Similar results were found for the secondary nonsleep outcomes at follow-up (data not shown).

Conventional Superiority Meta-Analysis

As presented in Table 2, when analyzed with conventional superiority meta-analysis, the pooled differences between ipCBTI and eCBTI reached statistical significance in 11 (34%) out of 32 comparisons. The effects generally favored ipCBTI but were small, for example, corresponding to a mean difference of 1.8 points on the ISI and 1.9% in sleep efficiency. The pooled difference for total sleep disturbance corresponded to a small ES (Hedges' *g*=0.32). The forest plots are shown in Figures 2-4 [69-85] and Figures S2-S11 in Multimedia Appendix 2. Concerning the secondary nonsleep outcomes of fatigue, anxiety, and depression, no differences in the conventional superiority analyses reached statistical significance (Table S4 in Multimedia Appendix 2).

Figure 2. Forest plot of postintervention differences (Hedges' *g*) between the effects of eHealth cognitive behavioral therapy for insomnia (eCBTI) and in-person–delivered cognitive behavioral therapy for insomnia (ipCBTI) on total sleep disturbance (red lines denote the equivalence margin).

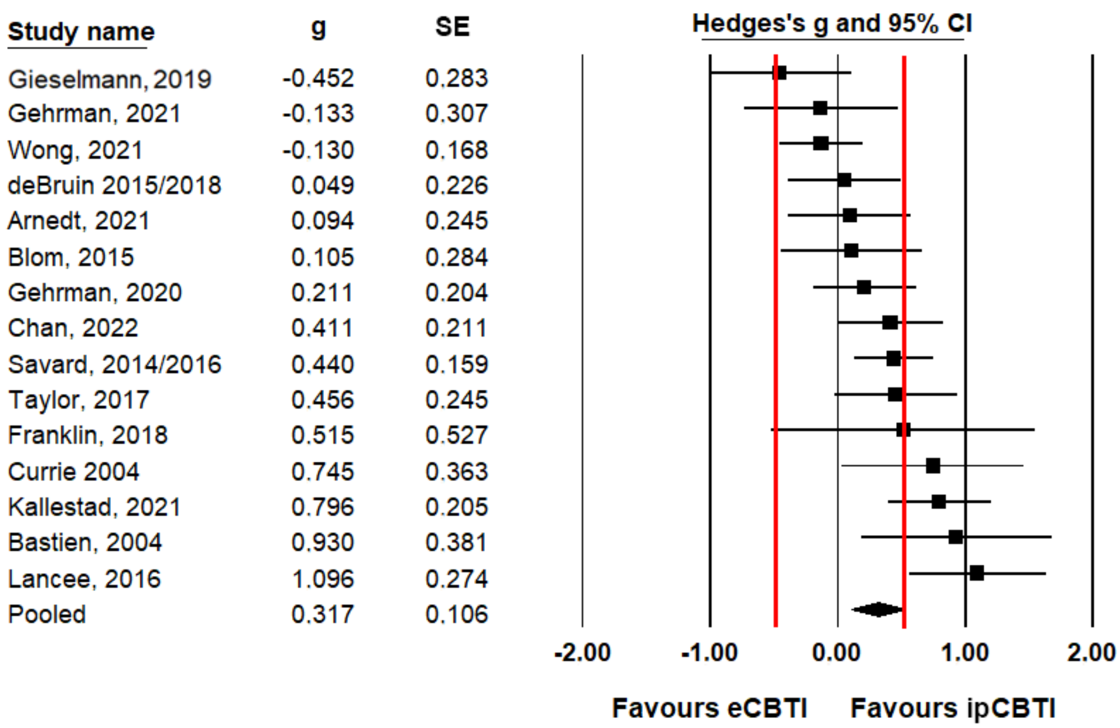


Figure 3. Forest plot of postintervention mean differences (%) between the effects of eHealth cognitive behavioral therapy for insomnia (eCBTI) and in-person–delivered cognitive behavioral therapy for insomnia (ipCBTI) on sleep efficiency (red lines denote the equivalence margin).

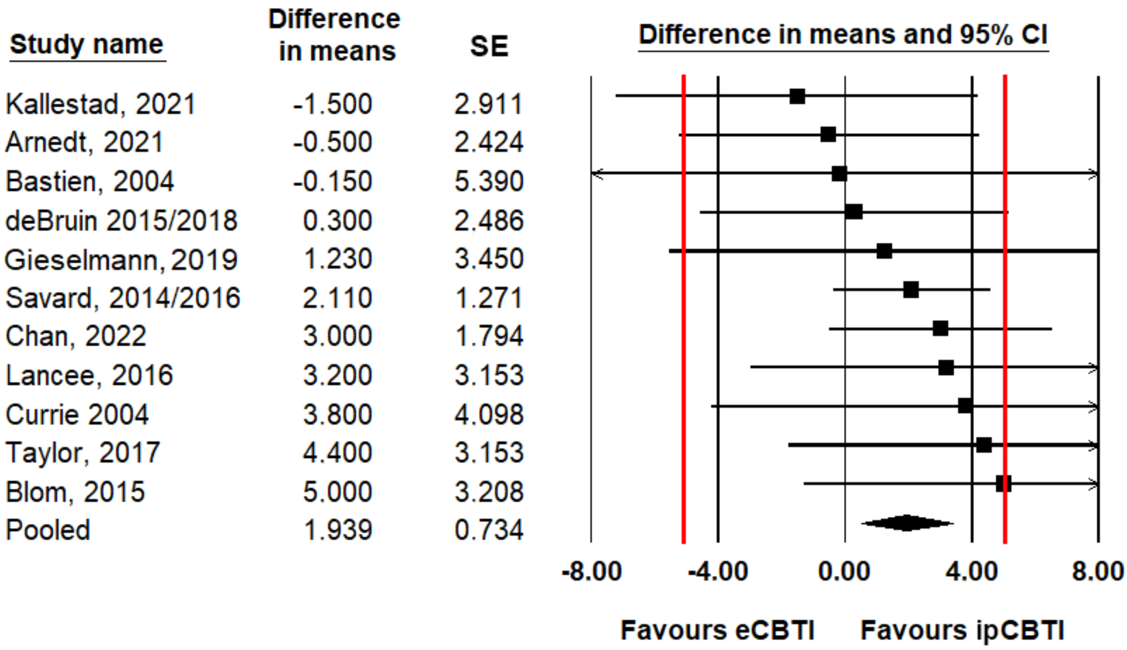


Figure 4. Forest plot of postintervention mean differences (min) between the effects of eHealth cognitive behavioral therapy for insomnia (eCBT) and in-person–delivered cognitive behavioral therapy for insomnia (ipCBTI) on total sleep time (red lines denote the equivalence margin).

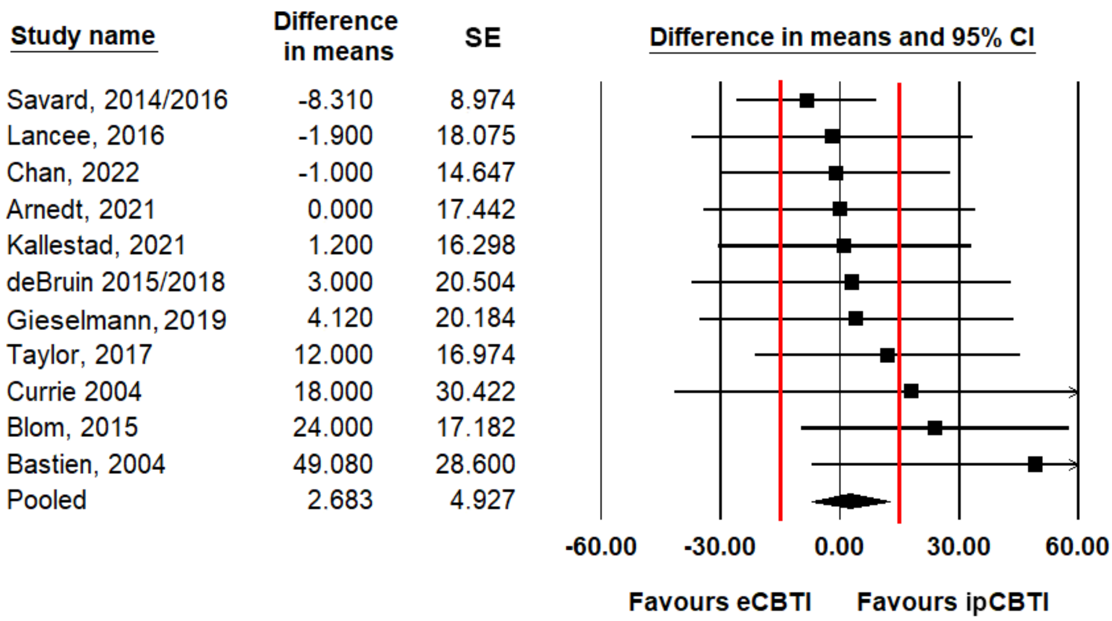


Table 2. Results of a meta-analysis of studies directly comparing the efficacy of in-person–delivered cognitive behavioral therapy for insomnia (ipCBTI) and digitally delivered eHealth CBTI (eCBTI), including tests of nonzero differences and tests of statistical equivalence.

ipCBTI vs eCBTI			Heterogeneity		Pooled effect			Equivalence[a]	
Outcome	K[b]	N[c]	I^2	T^2	Effect[d] (95% CI)	*P* value[e]	95% PI[f,g]	MID[h]	*P* value
Self-reported sleep outcomes at postintervention									
Total sleep disturbance (Hedges' *g*)	15	1068	63.5	0.10	0.32 (0.11 to 0.53)	.003[i]	–0.41 to 1.04	0.5 SD[j]	.04[i]
ISI[k], mean difference (points)	11	897	70.0	2.35	–1.8 (–2.9 to –0.7)	.002[i]	–5.48 to 1.92	2.6[l]	.07
ISI (Hedges' *g*)	11	897	63.7	0.09	0.37 (0.15 to 0.60)	.001[i]	–0.35 to 1.10	0.5 SD[j]	0.13
PSQI[m], mean difference (points)	5	279	63.1	1.48	–0.9 (–2.3 to 0.6)	.23	–5.35 to 3.64	4.4[n]	<.001[i]
PSQI (Hedges' *g*)	5	279	62.0	0.13	0.26 (–0.15 to 0.67)	.22	–1.06 to 1.58	0.5 SD[j]	.12
Sleep efficiency[o] diary[p], mean difference (%)	11	779	00.0	0.00	1.9 (0.5 to 3.4)	.01[i]	—[q]	5%[r]	<.001[i]
Sleep efficiency diary (Hedges' *g*)	11	779	00.0	0.00	0.17 (0.03 to 0.31)	.02[i]	—	0.5 SD[j]	<.001[i]
SOL[s] diary, mean difference (min)	10	719	00.0	0.00	–2.6 (–6.5 to 1.2)	.18	—	10 min[t]	<.001[i]
SOL diary (Hedges' *g*)	10	719	2.0	0.00	0.07 (–0.08 to 0.21)	.39	–0.13 to 0.26	0.5 SD	<.001[i]
WASO[u] diary, mean difference (min)	8	621	14.7	7.80	–2.5 (–7.5 to 2.6)	.34	–11.79 to 6.82	15 min[v]	<.001[i]
WASO (Hedges' *g*)	8	621	12.7	0.01	0.09 (–0.08 to 0.26)	.30	–0.21 to 0.39	0.5 SD[w]	<.001[i]
TST[x], mean difference (min)	11	779	00.0	0.00	–2.7 (–12.3 to 7.0)	.59	—	15 min[y]	.01[i]
TST (Hedges' *g*)	11	779	00.0	0.00	0.05 (–0.09 to 0.19)	.52	—	0.5 SD[w]	<.001[i]
Self-reported sleep outcomes at follow-up									
Total sleep disturbance (Hedges' *g*)	14	988	58.1	0.08	0.24 (0.04 to 0.45)	.02[i]	–0.42 to 0.90	0.5 SD[j]	.01[i]
ISI, mean difference (points)	10	817	65.7	1.98	–1.3 (–2.4 to –0.2)	.02[i]	–4.81 to 2.20	2.6[l]	<.001[i]
ISI (Hedges' *g*)	10	817	61.4	0.08	0.27 (0.04 to 0.50)	.02[i]	–0.44 to 0.98	0.5 SD[j]	.03[i]
PSQI, mean difference (points)	5	279	55.8	1.20	–0.8 (–2.1 to 0.6)	.28	–4.87 to 3.37	4.4[n]	<.001[i]
PSQI (Hedges' *g*)	5	279	53.8	0.09	0.21 (–0.16 to 0.58)	.27	–0.92 to 1.34	0.5 SD[j]	.06
Sleep efficiency diary, mean difference (%)	10	700	52.0	7.81	2.8 (0.3 to 5.4)	.03[i]	–4.26 to 9.92	5%[r]	.047[i]
Sleep efficiency diary (Hedges' *g*)	10	700	51.3	0.06	0.25 (0.03 to 0.47)	.03[i]	–0.38 to 0.88	0.5 SD[j]	.01[i]
SOL diary, mean difference (min)	9	640	9.0	4.46	–4.8 (–9.3 to –0.2)	.04[i]	–12.14 to 2.65	10 min[t]	.01[i]
SOL diary (Hedges' *g*)	9	640	22.0	0.02	0.14 (–0.04 to 0.32)	.12	–0.23 to 0.52	0.5 SD	<.001[i]

ipCBTI vs eCBTI			Heterogeneity		Pooled effect			Equivalence[a]	
Outcome	K[b]	N[c]	I^2	T^2	Effect[d] (95% CI)	*P* value[e]	95% PI[f,g]	MID[h]	*P* value
WASO diary, mean difference (min)	7	542	37.1	24.2	0.6 (–5.9 to 7.0)	.87	–14.66 to 15.77	15 min[v]	<.001[i]
WASO diary (Hedges' *g*)	7	542	35.4	0.03	–0.02 (–0.24 to 0.20)	.86	–0.55 to 0.51	0.5 SD[w]	<.001[i]
TST, mean difference (min)	10	700	38.3	193.2	6.3 (–8.2 to 20.9)	.39	–29.98 to 42.66	15 min[y]	.12
TST (Hedges' *g*)	10	700	36.9	0.03	0.08 (–0.11 to 0.28)	.40	–0.40 to 0.57	0.5 SD[w]	<.001[i]
Actigraphy-based sleep outcomes at postintervention									
Sleep efficiency actigraphy, mean difference (%)	3	194	00.0	0.00	–0.8 (–2.9 to 1.3)	.47	—	5%[r]	<.001[i]
Sleep efficiency actigraphy (Hedges' *g*)	3	194	00.0	0.00	–0.09 (–0.37 to 0.19)	.53	–1.91 to 1.73	0.5 SD[j]	.002[i]
SOL actigraphy, mean difference (min)	3	194	00.0	0.00	–2.6 (–8.9 to 3.6)	.41	–43.07 to 37.83	10 min[t]	.01[i]
SOL actigraphy (Hedges' *g*)	3	194	00.0	0.00	0.14 (–0.14 to 0.42)	.32	–1.68 to 1.96	0.5 SD[j]	.01[i]
TST actigraphy, mean difference (min)	3	194	2.8	7.04	–16.9 (–34.6 to 0.8)	.06	–136.2 to 102.4	15 min[y]	.42
TST actigraphy (Hedges' *g*)	3	194	31.2	0.03	–0.24 (–0.58 to 0.10)	.17	–3.33 to 2.87	0.5 SD[j]	.09

[a]The test of equivalence tests whether the CI falls within an equivalence interval. The equivalence test is based on the largest *P* value from two 1-sided tests [44].

[b]K: number of studies.

[c]N: total number of participants.

[d]Analyses were conducted for outcomes with K≥3 for mean differences (%, min) and standardized mean difference (SMD; adjusted for small sample bias; Hedges' *g*) [53], with positive values of Hedges' *g* indicating a difference of effects in favor of ipCBTI compared with eCBTI.

[e]2-tailed *P* values.

[f]PI: prediction interval.

[g]95% prediction interval is the interval in which 95% of future observations from the same family of studies will fall [61].

[h]MID: minimal important difference (or clinical significance threshold) [54].

[i]Statistically significant *P* values (*P*<.05) indicate equivalence.

[j]SMD=0.50, as suggested by Edinger et al [54].

[k]ISI: Insomnia Severity Index.

[l]2.6 point difference on the Insomnia Severity Index (ISI), corresponding to 0.5 × SD found in the original validation study (SD 4.2) [48] (average ISI baseline SD across studies in this review=4.1).

[m]PSQI: Pittsburgh Sleep Quality Index.

[n]4.4 point difference on the Pittsburgh Sleep Quality Index (PSQI), as suggested by Longo et al [55].

[o]Sleep efficiency (%; total sleep time/time in bed × 100).

[p]Diary: sleep parameters based on sleep diaries, for example, the Consensus Sleep Diary [92].

[q]Not applicable.

[r]5% difference, as suggested by Edinger et al [54].

[s]SOL: sleep onset latency (min).

[t]10-minute difference in sleep onset latency, as suggested by Edinger et al [54].

[u]WASO: wake after sleep onset (min).

[v]15-minute difference in wake after sleep onset, as suggested by Edinger et al [54].

[w]When no minimal important differences are available, we chose 0.5 SD, as suggested by Norman et al [58].

[x]TST: total sleep time (min).

[y]15 minutes difference in total sleep time, as suggested by Edinger et al [54].

Equivalence Meta-Analysis

As shown in Table 2, the 95% CI for total sleep disturbance was included in the prespecified equivalence interval for this outcome, and, based on the largest p-value of two one-sided tests, the null hypothesis of nonequivalence was rejected (*P*=.04). As shown in Table 2, when examining the various sleep outcomes, ipCBTI and eCBTI were statistically significantly equivalent for 25 (78%) out of 32 calculations. Furthermore, ipCBTI and eCBTI emerged as statistically significantly equivalent for all 3 secondary nonsleep outcomes at both time points (Table S4 in Multimedia Appendix 2).

Publication Bias and Outliers

Inspecting the funnel plot and Egger's test for total sleep disturbance, which included data from all included studies (15/15, 100%), did not indicate publication bias (Egger test, *P*=.56; refer to the funnel plot in Figure S12 in Multimedia Appendix 2). Considering ESs, larger or smaller than 2 SDs beyond the pooled ES, revealed no outliers.

Heterogeneity and Moderator Analyses

As seen in Table 2, heterogeneity analyses suggested that varying proportions of the variance in postintervention effects stem from between-study differences beyond random error. The I^2 values were highest for the questionnaire-based sleep outcomes (62%-70%) and generally lower for the sleep diary and actigraphy-based outcomes (0%-31.2%). The data also suggested relatively high levels of heterogeneity for outcomes at follow-up. As shown in Table 3, when exploring possible explanations for the heterogeneity with meta-regression, 2 (18%) of the 11 analyzed moderators reached statistical significance. Differences between the proportions of dropouts in the eCBTI and ipCBTI moderated the between-group effects, with higher dropout rates in eCBTI compared with ipCBTI being associated with larger differences in favor of ipCBTI at both postintervention and follow-up, explaining 50% and 74% of the variation, respectively. Longer overall treatment duration was associated with larger differences in favor of ipCBTI compared with eCBTI at both time points. No statistically significant effects were found for the remaining moderators analyzed.

Table 3. Results of moderator analyses based on standardized mean differences (Hedges' *g*) in total self-reported sleep disturbance outcomes between in-person–delivered cognitive behavioral therapy for insomnia (ipCBTI) and eHealth cognitive behavioral therapy for insomnia (eCBTI) at postintervention and follow-up time points.

Moderator and time point	K[a]	Slope[b] (95% CI)	*P* value[c]	R^2
Mean sample age				
Postintervention	14	0.01 (–0.01 to 0.02)	.55	0.04
Follow-up	12	–0.01 (–0.02 to 0.01)	.34	0.18
Percentage of women				
Postintervention	15	0.00 (–0.01 to 0.01)	.71	0.03
Follow-up	13	0.00 (–0.00 to 0.01)	.40	0.07
Comorbid insomnia (reference: Insomnia as primary problem)				
Postintervention	15	0.16 (–0.31 to 0.62)	.51	0.04
Follow-up	15	0.12 (–0.31 to 0.56)	.58	0.06
Study dropout (%)				
Postintervention	13	0.00 (–0.02 to 0.03)	.78	0.00
Follow-up	13	0.00 (–0.01 to 0.01)	.68	0.10
eCBTI-ipCBTI dropout difference (%)[d]				
Postintervention	13	0.02 (0.00 to 0.03)	*.02* [e]	0.50
Follow-up	13	0.02 (0.01 to 0.03)	*.004*	0.74
eCBTI therapist contact (reference: none)				
Postintervention	15	–0.07 (–0.51 to 0.36)	.74	0.01
Follow-up	14	0.16 (–0.28 to 0.59)	.48	0.11
Number of treatment sessions				
Postintervention	15	0.10 (–0.03 to 0.24)	.14	0.26
Follow-up	14	0.11 (–0.03 to 0.25)	.13	0.32
Treatment duration (weeks)				
Postintervention	15	0.18 (0.09 to 0.26)	*<.001*	0.93
Follow-up	14	0.16 (0.08 to 0.24)	*<.001*	0.94
Number of eCBTI components				
Postintervention	15	0.11 (–0.15 to 0.37)	.41	0.03
Follow-up	14	0.18 (–0.06 to 0.42)	.14	0.15
Number of ipCBTI components				
Postintervention	15	0.09 (–0.16 to 0.35)	.47	0.02
Follow-up	14	0.17 (–0.07 to 0.41)	.18	0.11
Time to follow-up (weeks)				
Follow-up	14	0.01 (–0.00 to 0.02)	.07	0.34

[a]K: number of studies in the analysis.

[b]Meta-regression (maximum likelihood method), conducted when K≥10.

[c]2-tailed *P* value.

[d]Difference in dropout (%) between conditions (eHealth cognitive behavioral therapy for insomnia [eCBTI] dropout minus in-person–delivered cognitive behavioral therapy for insomnia [ipCBTI] dropout). Positive values correspond to a higher dropout rate in eCBTI than in ipCBTI. Combined self-reported sleep quality outcomes include measures of insomnia severity (Insomnia Severity Index) and sleep quality (Pittsburgh Sleep Quality Index).

[e]Significant values (*P*<.05) are italicized.

Results of Supplementary Bayesian Analyses

As presented in Table S5 in Multimedia Appendix 2, the Bayesian meta-analyses favored the alternative hypothesis of equivalence, that is, a zero difference between ipCBTI and eCBTI, for 4 (67%) out of 6 sleep outcomes at postintervention time points. The Bayes factors (BFs), that is, the probabilities of the alternative hypotheses relative to the null hypotheses, ranged from 1.7 (PSQI) to 9.9 (TST), indicating that a zero difference between ipCBTI and eCBTI is 1.7 to 9.9 times more likely than a nonzero difference. The level of evidence [93] ranged from anecdotal (BF=1-3) for PSQI to moderate (BF=3-10) for SOL, WASO, and TST. A 0 and nonzero difference for sleep efficiency appeared equally likely (BF=1.2). Insomnia severity assessed with the ISI was the only outcome for which the current evidence clearly favored a nonzero difference, with this result being 10.5 times more likely than the null hypothesis. Concerning heterogeneity, the data provided strong (BF=13.0) and anecdotal evidence (BF=2.3) for heterogeneous ISI and PSQI ESs, respectively. Nonheterogeneity was slightly more likely for the remaining outcomes (BF=1.9-3.5).

Discussion

Sleep Outcomes

When pooling the results of the 15 unique randomized trials directly comparing eCBTIs with ipCBTIs using conventional meta-analysis, the observed differences generally favored ipCBTI. Specifically, the postintervention results revealed statistically significant advantages for ipCBTI across several dimensions, including overall sleep disturbance (encompassing both insomnia severity and sleep quality), insomnia severity assessed independently, and sleep efficiency. While ipCBTI was statistically significantly superior to eCBTI for these outcomes, the magnitudes of these differences were modest, corresponding to small ESs (Hedges' *g*) and small mean, nonstandardized differences. For example, concerning the latter, the pooled mean differences in favor of ipCBTI for insomnia severity and sleep efficiency were only 1.8 points (on the ISI) and 1.9%, respectively. Furthermore, for total sleep disturbance and sleep efficiency, the CIs fell within the suggested equivalence margins of 0.5 SD and 5%, respectively [54].

Regarding the remaining self-reported sleep outcomes at postintervention, none yielded statistically significantly superior results in favor of either delivery type. Furthermore, with the exception of Hedges' *g* for the PSQI, all remaining analyses showed the 2 delivery types to be statistically significantly equivalent; that is, the CIs of the pooled effect parameter fell within the suggested equivalence margin for that parameter. The same general pattern was observed for the results obtained at the (on average) 21-week follow-up. On the basis of the available data, eCBTI and ipCBTI were statistically significantly equivalent for almost all self-reported sleep outcomes, except for the ES for sleep quality assessed with the PSQI and the mean difference in minutes for TST.

While equivalence indicates similar efficacy, if one only examines the between-group differences, it cannot be determined whether the equivalence stems from similar small or similar large improvements in both conditions. Therefore, we also calculated the within-group effects for each delivery format. The results revealed that both ipCBTI and eCBTI were associated with statistically significant within-condition improvements in all self-reported outcomes at both postintervention and follow-up. The largest effects were seen in both conditions for total sleep disturbance, insomnia severity, sleep quality, and sleep efficiency. Small-to-medium effects were observed for the remaining self-reported sleep outcomes. Therefore, it may be concluded that both delivery formats appear efficacious, displaying improvements at postintervention corresponding to 9- and 7-point reductions on the ISI, 12% and 10% improvements in sleep efficiency, 21- and 20-minute reductions in SOL, 24- and 20-minute reductions in WASO, and 21- to 16-minute increase in TST for ipCBTI and eCBTI, respectively. These effects are all clinically relevant and well beyond the suggested MIDs and minimal important changes for these outcomes, that is, a 6-point within-person change on the ISI [56] and 10, 15, and 15 minutes for SOL, WASO, and TST, respectively [54]. In addition, these clinically relevant, positive improvements were sustained over time, supporting previous findings that CBTI, regardless of the delivery format, yields robust long-term effects [28].

Only 3 (20%) of the 15 studies assessed sleep objectively, that is, with actigraphy. Despite the small number of studies, the results for sleep efficiency and SOL showed the 2 delivery formats to be statistically significantly equivalent. In contrast, the results for actigraphy-assessed TST appeared to be in favor of eCBTI, with eCBTI resulting in increased TST and ipCBTI in reduced TST. However, neither the conventional nor the equivalence analyses reached statistical significance. When examining the within-group effects, statistically significant improvements were seen for actigraphy-based SOL in the ipCBTI group. The remaining effects for TST and sleep efficiency failed to reach statistical significance. Such discrepancies between effects on self-reported and objectively assessed sleep outcomes, especially concerning estimates of sleep duration, are a well-recognized issue in sleep research and clinical practice [94].

Nonsleep Outcomes

It is well known that insomnia can lead to various physical and mental symptoms, including increased levels of fatigue [95], and that it is a significant predictor of later onset of mental disorders such as depression and anxiety [8]. Therefore, we also explored the effects on the secondary nonsleep outcomes of fatigue, anxiety, and depression. The 2 delivery formats of CBTI were statistically significantly equivalent in their effects on these symptoms, and both yielded statistically significant medium-to-large within-condition improvements of almost identical magnitude in all 3 outcomes. Thus, our results add to the more general findings that CBTI may not only improve insomnia itself but also associated psychological and physical symptoms such as depression, anxiety, and fatigue [96-98], with our findings indicating that both delivery formats appear equally efficacious in reducing these symptoms.

Heterogeneity and Its Sources

When exploring possible heterogeneity of the postintervention effects, the relatively large I^2 statistics observed for both insomnia severity and sleep quality and the combined outcome of total sleep disturbance indicate that a considerable proportion of the variance is due to systematic differences between the study characteristics. In contrast, the differences in effects on sleep diary outcomes displayed little or no signs of heterogeneity. When we investigated possible sources of heterogeneity for the combined total sleep disturbance outcome, 2 study characteristics emerged as statistically significant moderators at both postintervention and follow-up.

First, higher dropout rates in eCBTI than in ipCBTI were significantly associated with larger differences in favor of ipCBTI in effect on total sleep disturbance. At postintervention time points, on average, 30% more participants in eCBTI had dropped out compared to ipCBTI. While we do not know the reasons for the higher dropout rates in eCBTI, this factor, which explained between 50% and 74% of the variance in between-condition differences in effect, could represent an important, potentially modifiable factor that needs to be addressed if the efficacy of eCBTI is to be further increased. While the research on adherence to electronically adapted interventions for insomnia is still limited, studies in this [99] and other clinical populations suggest that common factors influencing dropout and adherence across such interventions include engagement and motivation; technical issues and usability; and demographic factors such as age, educational level, and digital literacy [100,101]. Second, while there was no difference in the mean duration of the 2 delivery formats, interventions with longer duration favored ipCBTI. The moderating effect of intervention duration persisted when adjusting for study dropout. We have no clear explanation for this finding, but longer treatment duration may allow for increased trust and improved therapeutic alliance in personally delivered CBTI, which, in turn, will increase the effect.

None of the remaining moderators analyzed reached statistical significance, including demographic characteristics such as mean sample age and percentage of women in the sample; study characteristics such as time to follow-up; and treatment characteristics such as therapist contact versus no contact, the number of treatment sessions, and the number of CBTI components. Some of the nonsignificant results could be viewed as surprising. For example, one might have expected larger differences between ipCBTI and fully automated eCBTI than between ipCBTI and eCBTI with some degree of therapist contact. One would also have expected age to play a role, for example, that older sample age would be associated with larger between-condition differences. Possible reasons for nonsignificant findings could be insufficient between-study variation, for example, in sample age, and inadequate statistical power due to the relatively small number of studies. Further research is needed to identify the common and different factors associated with the increased efficacy of the 2 delivery formats.

Limitations

Our findings should be interpreted cautiously for several reasons. First, the interpretability is challenged by between-study heterogeneity, for example, by considerable between-study differences in eCBTI formats, with some eCBTIs delivered with direct therapist contact via telephone or videoconferencing; some delivered on the web with asynchronous therapist contact, for example, through email; and others offered as fully automated programs. While we attempted to explore the possible moderating role of such variations and found no indication of a moderating effect of the degree of therapist contact involved, the relatively small number of studies may have limited our ability to identify the influence of such characteristics.

Second, as demonstrated by the results of the Bayesian analyses, the small number of studies restricts the strength of the evidence. While the currently available evidence favored equivalent effects for 4 (67%) out of 6 outcomes (sleep quality, SOL, WASO, and TST), the level of evidence was weak (ie, anecdotal) to moderate. The evidence for sleep efficiency was inconclusive, and while the level of evidence for a nonzero difference in favor of ipCBTI was characterized as *strong*, the BF was only just above the lower limit (ie, ≥10) [93].

Third, interpreting the differences between eCBTI and ipCBTI as equivalent or nonequivalent clearly depends on the chosen equivalence margins. While we chose the MIDs suggested in the literature, for example, 2.6 points, 4.4 points, 5%, 10 minutes, 15 minutes, and 15 minutes for ISI, PSQI, sleep efficiency, SOL, WASO, and TST, respectively [48,54,55], specific MIDs have not been identified for all the corresponding ESs. While we used the 0.5 SD suggested in the literature [54,58], the clinical relevance of this value has yet to be established for several of the sleep outcomes investigated in this review.

Finally, assessed with the revised Cochrane Risk of Bias tool [51,102], one-third of the studies were characterized as having an overall *high risk* of bias, and the remaining two-thirds were characterized as having *some concerns*. Among the main reasons for these categorizations were high rates of missing outcome data and the use of self-reported outcomes. These issues cannot easily be amended. For example, it is not too surprising that behavioral interventions, in general, and eHealth interventions, in particular, have higher dropout rates than pharmacological trials. In addition, while sleep characteristics such as SOL, WASO, and TST can be assessed with both self-report and objective measures, insomnia is inherently a subjective outcome, which can be evaluated only with self-report. Furthermore, ensuring blinding is another factor that is difficult to obtain with behavioral interventions and not possible when comparing in-person and electronically delivered interventions. Other reasons for the identified risks of bias can be addressed more easily, including the failure to include sensitivity analyses correcting for missing outcome data and insufficient preregistering of analytical plans.

Conclusions

This, to our knowledge, first systematic review and meta-analysis of randomized head-to-head comparisons of eCBTI and ipCBTI suggests that while the effects tended to be in favor of the latter, the mean differences were generally of small magnitudes, with several approaching 0. Furthermore, the 2 CBTI delivery formats were statistically significantly equivalent for most outcomes examined. Statistically significant equivalence means that the CIs of the differences fell within the prespecified equivalence margins, with the latter being based on the minimal clinically relevant differences suggested in the literature for each of the outcomes in question. Importantly, when examining the within-condition effects, both delivery formats yielded large and clinically relevant effects on most outcomes, including the nonsleep outcomes of fatigue, anxiety, and depression. Although the results should be interpreted cautiously due to the currently limited evidence base, they support eCBTI, including fully automated programs, as clinically relevant alternatives to ipCBTI. These results are promising for people with insomnia, given the challenges of meeting population needs with conventional treatment formats.

Authors' Contributions

SMK contributed to conceptualization, formal analysis, investigation, data curation, writing the original draft (*Introduction*, *Methods*, and *Results* sections), reviewing and editing the manuscript, and visualization. DSC contributed to conceptualization, investigation, data curation, and reviewing and editing the manuscript. PC contributed to data curation and reviewing and editing the manuscript. MFD contributed to conceptualization and reviewing and editing the manuscript. AA contributed to conceptualization and reviewing and editing the manuscript. RZ contributed to conceptualization, formal analysis, investigation, writing the original draft (*Methods*, *Results*, and *Discussion* sections), reviewing and editing the manuscript, visualization, and supervision.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Completed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) 2020 checklist.

[PDF File (Adobe PDF File), 737 KB - mental_v11i1e58217_app1.pdf]

Multimedia Appendix 2

Tables S1-S5 provide the detailed search strategy, list of excluded studies, within-group pre-post effects of in-person–delivered cognitive behavioral therapy for insomnia (ipCBTI) and eHealth cognitive behavioral therapy for insomnia (eCBTI), superiority and equivalence analyses of secondary non-sleep outcomes, supplementary Bayesian meta-analyses, and risk of bias assessments. Figures S1-S12 present forest plots of postintervention differences between effects of eCBTI and ipCBTI on primary and secondary sleep outcomes.

[PDF File (Adobe PDF File), 2352 KB - mental_v11i1e58217_app2.pdf]

References

1. Morin CM, Jarrin DC. Epidemiology of insomnia: prevalence, course, risk factors, and public health burden. Sleep Med Clin 2022 Jun;17(2):173-191. [doi: 10.1016/j.jsmc.2022.03.003] [Medline: 35659072]
2. Javaheri S, Redline S. Insomnia and risk of cardiovascular disease. Chest 2017 Aug;152(2):435-444 [FREE Full text] [doi: 10.1016/j.chest.2017.01.026] [Medline: 28153671]
3. Sofi F, Cesari F, Casini A, Macchi C, Abbate R, Gensini GF. Insomnia and risk of cardiovascular disease: a meta-analysis. Eur J Prev Cardiol 2014 Jan;21(1):57-64. [doi: 10.1177/2047487312460020] [Medline: 22942213]
4. Cohen S, Doyle WJ, Alper CM, Janicki-Deverts D, Turner RB. Sleep habits and susceptibility to the common cold. Arch Intern Med 2009 Jan 12;169(1):62-67 [FREE Full text] [doi: 10.1001/archinternmed.2008.505] [Medline: 19139325]
5. Patel SR, Malhotra A, Gao X, Hu FB, Neuman MI, Fawzi WW. A prospective study of sleep duration and pneumonia risk in women. Sleep 2012 Jan 01;35(1):97-101 [FREE Full text] [doi: 10.5665/sleep.1594] [Medline: 22215923]
6. de Almondes KM, Costa MV, Malloy-Diniz LF, Diniz BS. Insomnia and risk of dementia in older adults: systematic review and meta-analysis. J Psychiatr Res 2016 Jun;77:109-115. [doi: 10.1016/j.jpsychires.2016.02.021] [Medline: 27017287]
7. Shi L, Chen SJ, Ma MY, Bao YP, Han Y, Wang YM, et al. Sleep disturbances increase the risk of dementia: a systematic review and meta-analysis. Sleep Med Rev 2018 Aug;40:4-16. [doi: 10.1016/j.smrv.2017.06.010] [Medline: 28890168]
8. Hertenstein E, Feige B, Gmeiner T, Kienzler C, Spiegelhalder K, Johann A, et al. Insomnia as a predictor of mental disorders: a systematic review and meta-analysis. Sleep Med Rev 2019 Feb;43:96-105. [doi: 10.1016/j.smrv.2018.10.006] [Medline: 30537570]
9. Hom MA, Hames JL, Bodell LP, Buchman-Schmitt JM, Chu C, Rogers ML, et al. Investigating insomnia as a cross-sectional and longitudinal predictor of loneliness: findings from six samples. Psychiatry Res 2017 Jul;253:116-128 [FREE Full text] [doi: 10.1016/j.psychres.2017.03.046] [Medline: 28364589]
10. Ben Simon E, Walker MP. Sleep loss causes social withdrawal and loneliness. Nat Commun 2018 Aug 14;9(1):3146 [FREE Full text] [doi: 10.1038/s41467-018-05377-0] [Medline: 30108218]

11. Cappuccio FP, D'Elia L, Strazzullo P, Miller MA. Sleep duration and all-cause mortality: a systematic review and meta-analysis of prospective studies. Sleep 2010 May;33(5):585-592 [FREE Full text] [doi: 10.1093/sleep/33.5.585] [Medline: 20469800]
12. Liu TZ, Xu C, Rota M, Cai H, Zhang C, Shi MJ, et al. Sleep duration and risk of all-cause mortality: a flexible, non-linear, meta-regression of 40 prospective cohort studies. Sleep Med Rev 2017 Apr;32:28-36. [doi: 10.1016/j.smrv.2016.02.005] [Medline: 27067616]
13. Daley M, Morin CM, LeBlanc M, Grégoire JP, Savard J. The economic burden of insomnia: direct and indirect costs for individuals with insomnia syndrome, insomnia symptoms, and good sleepers. Sleep 2009 Jan;32(1):55-64 [FREE Full text] [Medline: 19189779]
14. Daley M, Morin CM, LeBlanc M, Grégoire JP, Savard J, Baillargeon L. Insomnia and its relationship to health-care utilization, work absenteeism, productivity and accidents. Sleep Med 2009 Apr;10(4):427-438. [doi: 10.1016/j.sleep.2008.04.005] [Medline: 18753000]
15. Curcio G, Ferrara M, de Gennaro L. Sleep loss, learning capacity and academic performance. Sleep Med Rev 2006 Oct;10(5):323-337. [doi: 10.1016/j.smrv.2005.11.001] [Medline: 16564189]
16. Trauer JM, Qian MY, Doyle JS, Rajaratnam SM, Cunnington D. Cognitive behavioral therapy for chronic insomnia: a systematic review and meta-analysis. Ann Intern Med 2015 Aug 04;163(3):191-204. [doi: 10.7326/M14-2841] [Medline: 26054060]
17. Morin CM, Benca R. Chronic insomnia. The Lancet 2012 Mar;379(9821):1129-1141. [doi: 10.1016/s0140-6736(11)60750-2]
18. Kripke DF. Hypnotic drug risks of mortality, infection, depression, and cancer: but lack of benefit. F1000Res 2016;5:918 [FREE Full text] [doi: 10.12688/f1000research.8729.3] [Medline: 27303633]
19. Edinger JD, Arnedt JT, Bertisch SM, Carney CE, Harrington JJ, Lichstein KL, et al. Behavioral and psychological treatments for chronic insomnia disorder in adults: an American Academy of Sleep Medicine clinical practice guideline. J Clin Sleep Med 2021 Feb 01;17(2):255-262 [FREE Full text] [doi: 10.5664/jcsm.8986] [Medline: 33164742]
20. Riemann D, Baglioni C, Bassetti C, Bjorvatn B, Dolenc Groselj L, Ellis JG, et al. European guideline for the diagnosis and treatment of insomnia. J Sleep Res 2017 Dec;26(6):675-700. [doi: 10.1111/jsr.12594] [Medline: 28875581]
21. Morin CM, Bei B, Bjorvatn B, Poyares D, Spiegelhalder K, Wing YK. World sleep society international sleep medicine guidelines position statement endorsement of "behavioral and psychological treatments for chronic insomnia disorder in adults: an American Academy of sleep medicine clinical practice guidelines". Sleep Med 2023 Sep;109:164-169. [doi: 10.1016/j.sleep.2023.07.001] [Medline: 37454606]
22. Spielman AJ, Saskin P, Thorpy MJ. Treatment of chronic insomnia by restriction of time in bed. Sleep 1987 Jan;10(1):45-56. [doi: 10.1093/sleep/10.1.45]
23. Bootzin R, Perlis M. Stimulus-control therapy. In: Perlis M, Aloia M, Kuhn B, editors. Behavioral Treatments for Sleep Disorders: A Volume in Practical Resources for the Mental Health Professional. London, UK: Academic Press; 2011:21-30.
24. Freedman R, Papsdorf JD. Biofeedback and progressive relaxation treatment of sleep-onset insomnia. Biofeedback Self Regul 1976 Sep;1(3):253-271. [doi: 10.1007/bf01001167]
25. Belanger L, Savard J, Morin CM. Clinical management of insomnia using cognitive therapy. Behav Sleep Med 2006 Aug;4(3):179-198. [doi: 10.1207/s15402010bsm0403_4] [Medline: 16879081]
26. Bootzin RR, Perlis ML. Nonpharmacologic treatments of insomnia. J Clin Psychiatry 1992 Jun;53 Suppl:37-41. [Medline: 1613018]
27. van Straten A, van der Zweerde T, Kleiboer A, Cuijpers P, Morin CM, Lancee J. Cognitive and behavioral therapies in the treatment of insomnia: a meta-analysis. Sleep Med Rev 2018 Apr;38:3-16. [doi: 10.1016/j.smrv.2017.02.001] [Medline: 28392168]
28. van der Zweerde T, Bisdounis L, Kyle SD, Lancee J, van Straten A. Cognitive behavioral therapy for insomnia: a meta-analysis of long-term effects in controlled studies. Sleep Med Rev 2019 Dec;48:101208. [doi: 10.1016/j.smrv.2019.08.002] [Medline: 31491656]
29. Selvanathan J, Pham C, Nagappa M, Peng PW, Englesakis M, Espie CA, et al. Cognitive behavioral therapy for insomnia in patients with chronic pain - a systematic review and meta-analysis of randomized controlled trials. Sleep Med Rev 2021 Dec;60:101460. [doi: 10.1016/j.smrv.2021.101460] [Medline: 33610967]
30. Squires LR, Rash JA, Fawcett J, Garland SN. Systematic review and meta-analysis of cognitive-behavioural therapy for insomnia on subjective and actigraphy-measured sleep and comorbid symptoms in cancer survivors. Sleep Med Rev 2022 Jun;63:101615. [doi: 10.1016/j.smrv.2022.101615] [Medline: 35303692]
31. Zhang Y, Ren R, Yang L, Zhang H, Shi Y, Shi J, et al. Comparative efficacy and acceptability of psychotherapies, pharmacotherapies, and their combination for the treatment of adult insomnia: a systematic review and network meta-analysis. Sleep Med Rev 2022 Oct;65:101687. [doi: 10.1016/j.smrv.2022.101687] [Medline: 36027795]
32. Koffel E, Bramoweth AD, Ulmer CS. Increasing access to and utilization of cognitive behavioral therapy for insomnia (CBT-I): a narrative review. J Gen Intern Med 2018 Jun;33(6):955-962 [FREE Full text] [doi: 10.1007/s11606-018-4390-1] [Medline: 29619651]
33. Conroy DA, Ebben MR. Referral practices for cognitive behavioral therapy for insomnia: a survey study. Behav Neurol 2015;2015:819402 [FREE Full text] [doi: 10.1155/2015/819402] [Medline: 26265887]

34. Roach M, Juday T, Tuly R, Chou JW, Jena AB, Doghramji PP. Challenges and opportunities in insomnia disorder. Int J Neurosci 2021 Nov;131(11):1058-1065. [doi: 10.1080/00207454.2020.1773460] [Medline: 32449423]
35. Cheng SK, Dizon J. Computerised cognitive behavioural therapy for insomnia: a systematic review and meta-analysis. Psychother Psychosom 2012 May 11;81(4):206-216. [doi: 10.1159/000335379] [Medline: 22585048]
36. Zachariae R, Lyby MS, Ritterband LM, O'Toole MS. Efficacy of internet-delivered cognitive-behavioral therapy for insomnia - a systematic review and meta-analysis of randomized controlled trials. Sleep Med Rev 2016 Dec;30:1-10. [doi: 10.1016/j.smrv.2015.10.004] [Medline: 26615572]
37. Ritterband LM, Thorndike FP, Gonder-Frederick LA, Magee JC, Bailey ET, Saylor DK, et al. Efficacy of an internet-based behavioral intervention for adults with insomnia. Arch Gen Psychiatry 2009 Jul 01;66(7):692-698 [FREE Full text] [doi: 10.1001/archgenpsychiatry.2009.66] [Medline: 19581560]
38. Kaldo V, Jernelöv S, Blom K, Ljótsson B, Brodin M, Jörgensen M, et al. Guided internet cognitive behavioral therapy for insomnia compared to a control treatment - a randomized trial. Behav Res Ther 2015 Aug;71:90-100 [FREE Full text] [doi: 10.1016/j.brat.2015.06.001] [Medline: 26091917]
39. Zachariae R, Amidi A, Damholdt MF, Clausen CD, Dahlgaard J, Lord H, et al. Internet-delivered cognitive-behavioral therapy for insomnia in breast cancer survivors: a randomized controlled trial. J Natl Cancer Inst 2018 Aug 01;110(8):880-887 [FREE Full text] [doi: 10.1093/jnci/djx293] [Medline: 29471478]
40. Hasan F, Tu YK, Yang CM, James Gordon C, Wu D, Lee HC, et al. Comparative efficacy of digital cognitive behavioral therapy for insomnia: a systematic review and network meta-analysis. Sleep Med Rev 2022 Feb;61:101567. [doi: 10.1016/j.smrv.2021.101567] [Medline: 34902820]
41. Simon L, Steinmetz L, Feige B, Benz F, Spiegelhalder K, Baumeister H. Comparative efficacy of onsite, digital, and other settings for cognitive behavioral therapy for insomnia: a systematic review and network meta-analysis. Sci Rep 2023 Feb 02;13(1):1929 [FREE Full text] [doi: 10.1038/s41598-023-28853-0] [Medline: 36732610]
42. Forma F, Pratiwadi R, El-Moustaid F, Smith N, Thorndike F, Velez F. Network meta-analysis comparing the effectiveness of a prescription digital therapeutic for chronic insomnia to medications and face-to-face cognitive behavioral therapy in adults. Curr Med Res Opin 2022 Oct;38(10):1727-1738. [doi: 10.1080/03007995.2022.2108616] [Medline: 35938209]
43. Lakens D, Scheel AM, Isager PM. Equivalence testing for psychological research: a tutorial. Adv Method Pract Psychol Sci 2018 Jun 01;1(2):259-269. [doi: 10.1177/2515245918770963]
44. Rogers JL, Howard KI, Vessey JT. Using significance tests to evaluate equivalence between two experimental groups. Psychol Bull 1993;113(3):553-565. [doi: 10.1037//0033-2909.113.3.553]
45. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989 Dec;10(4):407-415. [doi: 10.1016/0197-2456(89)90005-6] [Medline: 2691207]
46. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
47. Straus SE. Evidence-Based Medicine: How to Practice and Teach EBM. New York, NY: Churchill Livingstone; 2005.
48. Bastien CH, Vallières A, Morin CM. Validation of the Insomnia Severity Index as an outcome measure for insomnia research. Sleep Med 2001 Jul;2(4):297-307. [doi: 10.1016/s1389-9457(00)00065-4] [Medline: 11438246]
49. Buysse DJ, Reynolds CF3, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. Psychiatry Res 1989 May;28(2):193-213. [doi: 10.1016/0165-1781(89)90047-4] [Medline: 2748771]
50. Covidence systematic review software. Veritas Health Innovation. URL: https://www.covidence.org/ [accessed 2024-07-09]
51. Sterne JA, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019 Aug 28;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
52. Soh HL, Ho RC, Ho CS, Tam WW. Efficacy of digital cognitive behavioural therapy for insomnia: a meta-analysis of randomised controlled trials. Sleep Med 2020 Nov;75:315-325. [doi: 10.1016/j.sleep.2020.08.020] [Medline: 32950013]
53. Hedges LV, Olkin I. Statistical Methods for Meta-Analysis. Amsterdam, The Netherlands: Elsevier Science; 1985.
54. Edinger JD, Arnedt JT, Bertisch SM, Carney CE, Harrington JJ, Lichstein KL, et al. Behavioral and psychological treatments for chronic insomnia disorder in adults: an American Academy of Sleep Medicine systematic review, meta-analysis, and GRADE assessment. J Clin Sleep Med 2021 Feb 01;17(2):263-298 [FREE Full text] [doi: 10.5664/jcsm.8988] [Medline: 33164741]
55. Longo UG, Berton A, de Salvatore S, Piergentili I, Casciani E, Faldetta A, et al. Minimal clinically important difference and patient acceptable symptom state for the Pittsburgh Sleep Quality Index in patients who underwent rotator cuff tear repair. Int J Environ Res Public Health 2021 Aug 17;18(16):8666 [FREE Full text] [doi: 10.3390/ijerph18168666] [Medline: 34444415]
56. Yang M, Morin CM, Schaefer K, Wallenstein GV. Interpreting score differences in the Insomnia Severity Index: using health-related outcomes to define the minimally important difference. Curr Med Res Opin 2009 Oct 19;25(10):2487-2494. [doi: 10.1185/03007990903167415] [Medline: 19689221]
57. Cuijpers P, Turner EH, Koole SL, van Dijke A, Smit F. What is the threshold for a clinically relevant effect? the case of major depressive disorders. Depress Anxiety 2014 May;31(5):374-378. [doi: 10.1002/da.22249] [Medline: 24677535]

58. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 2003 May;41(5):582-592. [doi: 10.1097/01.MLR.0000062554.74615.4C] [Medline: 12719681]
59. Cooper H, Hedges LV, Valentine JC. The Handbook of Research Synthesis and Meta-Analysis. New York, NY: Russell Sage Foundation; 2009.
60. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 06;327(7414):557-560 [FREE Full text] [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
61. IntHout J, Ioannidis JP, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open 2016 Jul 12;6(7):e010247 [FREE Full text] [doi: 10.1136/bmjopen-2015-010247] [Medline: 27406637]
62. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634 [FREE Full text] [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
63. Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 2000 Jun 24;56(2):455-463. [doi: 10.1111/j.0006-341x.2000.00455.x] [Medline: 10877304]
64. Borenstein M, Hedges LV, Higgins JP, Rothstein H. Comprehensive meta‐analysis software. In: Introduction to Meta‐Analysis. Hoboken, NJ: John Wiley & Sons; 2021.
65. Gronau QF, van Erp S, Heck DW, Cesario J, Jonas KJ, Wagenmakers EJ. A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: the case of felt power. Compr Results Soc Psychol 2017 Jun 28;2(1):123-138. [doi: 10.1080/23743603.2017.1326760]
66. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. Am Stat 2016 Jun 09;70(2):129-133. [doi: 10.1080/00031305.2016.1154108]
67. van Erp S, Verhagen J, Grasman R, Wagenmakers EJ. Estimates of between-study heterogeneity for 705 meta-analyses reported in psychological bulletin from 1990-2013. J Open Psychol Data 2017 Aug 17;5(1):4. [doi: 10.5334/jopd.33]
68. JASP 0.17.1 (Windows 64 Bit). JASP Team. URL: https://jasp-stats.org/previous-versions/ [accessed 2024-07-09]
69. Currie SR, Clark S, Hodgins DC, El-Guebaly N. Randomized controlled trial of brief cognitive-behavioural interventions for insomnia in recovering alcoholics. Addiction 2004 Sep;99(9):1121-1132. [doi: 10.1111/j.1360-0443.2004.00835.x] [Medline: 15317632]
70. Bastien CH, Morin CM, Ouellet MC, Blais FC, Bouchard S. Cognitive-behavioral therapy for insomnia: comparison of individual therapy, group therapy, and telephone consultations. J Consult Clin Psychol 2004 Aug;72(4):653-659. [doi: 10.1037/0022-006X.72.4.653] [Medline: 15301650]
71. Savard J, Ivers H, Savard MH, Morin CM. Is a video-based cognitive behavioral therapy for insomnia as efficacious as a professionally administered treatment in breast cancer? results of a randomized controlled trial. Sleep 2014 Aug 01;37(8):1305-1314 [FREE Full text] [doi: 10.5665/sleep.3918] [Medline: 25083010]
72. Savard J, Ivers H, Savard M, Morin CM. Long-term effects of two formats of cognitive behavioral therapy for insomnia comorbid with breast cancer. Sleep 2016 Apr 01;39(4):813-823 [FREE Full text] [doi: 10.5665/sleep.5634] [Medline: 26715229]
73. Blom K, Tarkian Tillgren H, Wiklund T, Danlycke E, Forssén M, Söderström A, et al. Internet-vs. group-delivered cognitive behavior therapy for insomnia: a randomized controlled non-inferiority trial. Behav Res Ther 2015 Jul;70:47-55 [FREE Full text] [doi: 10.1016/j.brat.2015.05.002] [Medline: 25981329]
74. de Bruin EJ, Bögels SM, Oort FJ, Meijer AM. Efficacy of cognitive behavioral therapy for insomnia in adolescents: a randomized controlled trial with internet therapy, group therapy and a waiting list condition. Sleep 2015 Dec 01;38(12):1913-1926 [FREE Full text] [doi: 10.5665/sleep.5240] [Medline: 26158889]
75. de Bruin EJ, Bögels SM, Oort FJ, Meijer AM. Improvements of adolescent psychopathology after insomnia treatment: results from a randomized controlled trial over 1 year. J Child Psychol Psychiatry 2018 May;59(5):509-522. [doi: 10.1111/jcpp.12834] [Medline: 29052846]
76. Lancee J, van Straten A, Morina N, Kaldo V, Kamphuis JH. Guided online or face-to-face cognitive behavioral treatment for insomnia: a randomized wait-list controlled trial. Sleep 2016 Jan 01;39(1):183-191 [FREE Full text] [doi: 10.5665/sleep.5344] [Medline: 26414893]
77. Taylor DJ, Peterson AL, Pruiksma KE, Young-McCaughan S, Nicholson K, Mintz J. Internet and in-person cognitive behavioral therapy for insomnia in military personnel: a randomized clinical trial. Sleep 2017 Jun 01;40(6). [doi: 10.1093/sleep/zsx075] [Medline: 28472528]
78. Laurel Franklin C, Walton JL, Raines AM, Chambliss JL, Corrigan SA, Cuccurullo LA, et al. Pilot study comparing telephone to in-person delivery of cognitive-behavioural therapy for trauma-related insomnia for rural veterans. J Telemed Telecare 2017 Sep 26;24(9):629-635. [doi: 10.1177/1357633x17732366] [Medline: 28950755]
79. Gieselmann A, Pietrowsky R. The effects of brief chat-based and face-to-face psychotherapy for insomnia: a randomized waiting list controlled trial. Sleep Med 2019 Sep;61:63-72. [doi: 10.1016/j.sleep.2019.03.024] [Medline: 31262551]
80. Gehrman P, Barilla H, Medvedeva E, Bellamy S, O'Brien E, Kuna ST. Randomized trial of telehealth delivery of cognitive-behavioral treatment for insomnia vs. in-person treatment in veterans with PTSD. J Affect Disord Rep 2020 Dec;1:100018. [doi: 10.1016/j.jadr.2020.100018]

81. Arnedt JT, Conroy DA, Mooney A, Furgal A, Sen A, Eisenberg D. Telemedicine versus face-to-face delivery of cognitive behavioral therapy for insomnia: a randomized controlled noninferiority trial. Sleep 2021 Jan 21;44(1):zsaa136. [doi: 10.1093/sleep/zsaa136] [Medline: 32658298]
82. Gehrman P, Gunter P, Findley J, Frasso R, Weljie AM, Kuna ST, et al. Randomized noninferiority trial of telehealth delivery of cognitive behavioral treatment of insomnia compared to in-person care. J Clin Psychiatry 2021 Aug 24;82(5):20m13723. [doi: 10.4088/JCP.20m13723] [Medline: 34428360]
83. Kallestad H, Scott J, Vedaa Ø, Lydersen S, Vethe D, Morken G, et al. Mode of delivery of cognitive behavioral therapy for insomnia: a randomized controlled non-inferiority trial of digital and face-to-face therapy. Sleep 2021 Dec 10;44(12):zsab185 [FREE Full text] [doi: 10.1093/sleep/zsab185] [Medline: 34291808]
84. Wong KY, Chung KF, Au CH. Low-intensity cognitive behavioral therapy for insomnia as the entry of the stepped-care model in the community: a randomized controlled trial. Behav Sleep Med 2021;19(3):378-394. [doi: 10.1080/15402002.2020.1764000] [Medline: 32429708]
85. Chan NY, Lam SP, Zhang J, Chan JW, Yu MM, Suh S, et al. Efficacy of email-delivered versus face-to-face group cognitive behavioral therapy for insomnia in youths: a randomized controlled trial. J Adolesc Health 2022 May;70(5):763-773. [doi: 10.1016/j.jadohealth.2021.11.005] [Medline: 35125265]
86. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand 1983 Jun 23;67(6):361-370. [doi: 10.1111/j.1600-0447.1983.tb09716.x] [Medline: 6880820]
87. Beck AT, Steer RA, Brown GK. BDI-II, Beck Depression Inventory: Manual. New York, NY: Psychological Corporation; 1995.
88. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
89. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. Appl Psychol Meas 2016 Jul 26;1(3):385-401. [doi: 10.1177/014662167700100306]
90. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
91. Smets EM, Garssen B, Bonke B, de Haes JC. The Multidimensional Fatigue Inventory (MFI) psychometric qualities of an instrument to assess fatigue. J Psychosom Res 1995 Apr;39(3):315-325. [doi: 10.1016/0022-3999(94)00125-o] [Medline: 7636775]
92. Carney CE, Buysse DJ, Ancoli-Israel S, Edinger JD, Krystal AD, Lichstein KL, et al. The consensus sleep diary: standardizing prospective sleep self-monitoring. Sleep 2012 Feb 01;35(2):287-302 [FREE Full text] [doi: 10.5665/sleep.1642] [Medline: 22294820]
93. Jeffreys H. Theory of Probability, Third Edition. Oxford, UK: Oxford University Press; 1984.
94. Buysse DJ. Sleep health: can we define it? does it matter? Sleep 2014 Jan 01;37(1):9-17 [FREE Full text] [doi: 10.5665/sleep.3298] [Medline: 24470692]
95. Huang L, Zhu W, Li N, Zhang B, Dai W, Li S, et al. Functions and mechanisms of adenosine and its receptors in sleep regulation. Sleep Med 2024 Mar;115:210-217 [FREE Full text] [doi: 10.1016/j.sleep.2024.02.012] [Medline: 38373361]
96. Ballesio A, Aquino MR, Feige B, Johann AF, Kyle SD, Spiegelhalder K, et al. The effectiveness of behavioural and cognitive behavioural therapies for insomnia on depressive and fatigue symptoms: a systematic review and network meta-analysis. Sleep Med Rev 2018 Feb;37:114-129. [doi: 10.1016/j.smrv.2017.01.006] [Medline: 28619248]
97. Ye YY, Zhang YF, Chen J, Liu J, Li XJ, Liu YZ, et al. Internet-based cognitive behavioral therapy for insomnia (ICBT-i) improves comorbid anxiety and depression-a meta-analysis of randomized controlled trials. PLoS One 2015;10(11):e0142258 [FREE Full text] [doi: 10.1371/journal.pone.0142258] [Medline: 26581107]
98. Tang NK, Lereya ST, Boulton H, Miller MA, Wolke D, Cappuccio FP. Nonpharmacological treatments of insomnia for long-term painful conditions: a systematic review and meta-analysis of patient-reported outcomes in randomized controlled trials. Sleep 2015 Nov 01;38(11):1751-1764 [FREE Full text] [doi: 10.5665/sleep.5158] [Medline: 25902806]
99. Sanchez-Ortuno MM, Pecune F, Coelho J, Micoulaud-Franchi JA, Salles N, Auriacombe M, et al. Predictors of users' adherence to a fully automated digital intervention to manage insomnia complaints. J Am Med Inform Assoc 2023 Nov 17;30(12):1934-1942. [doi: 10.1093/jamia/ocad163] [Medline: 37672004]
100. Pedersen DH, Mansourvar M, Sortsø C, Schmidt T. Predicting dropouts from an electronic health platform for lifestyle interventions: analysis of methods and predictors. J Med Internet Res 2019 Sep 04;21(9):e13617 [FREE Full text] [doi: 10.2196/13617] [Medline: 31486409]
101. Alfonsson S, Olsson E, Hursti T. Motivation and treatment credibility predicts dropout, treatment adherence, and clinical outcomes in an internet-based cognitive behavioral relaxation program: a randomized controlled trial. J Med Internet Res 2016 Mar 08;18(3):e52 [FREE Full text] [doi: 10.2196/jmir.5352] [Medline: 26957354]
102. McGuinness LA, Higgins JP. Risk-of-bias VISualization (robvis): an R package and Shiny web app for visualizing risk-of-bias assessments. Res Synth Methods 2021 Jan;12(1):55-61. [doi: 10.1002/jrsm.1411] [Medline: 32336025]

Abbreviations

BF: Bayes factor
CBTI: cognitive behavioral therapy for insomnia
eCBTI: eHealth cognitive behavioral therapy for insomnia
ES: effect size
ipCBTI: in-person–delivered cognitive behavioral therapy for insomnia
ISI: Insomnia Severity Index
MID: minimal important difference
PICO: Population, Intervention, Comparison, and Outcome
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis
PSQI: Pittsburgh Sleep Quality Index
SOL: sleep onset latency
TST: total sleep time
WASO: wake after sleep onset

Edited by J Torous; submitted 09.03.24; peer-reviewed by E Jafari, M Sakata; comments to author 03.05.24; revised version received 22.05.24; accepted 15.06.24; published 26.08.24.

Please cite as:
Knutzen SM, Christensen DS, Cairns P, Damholdt MF, Amidi A, Zachariae R
Efficacy of eHealth Versus In-Person Cognitive Behavioral Therapy for Insomnia: Systematic Review and Meta-Analysis of Equivalence
JMIR Ment Health 2024;11:e58217
URL: https://mental.jmir.org/2024/1/e58217
doi:10.2196/58217
PMID:

Review

Evaluation of Digital Mental Health Technologies in the United States: Systematic Literature Review and Framework Synthesis

Julianna Catania[1], MPH; Steph Beaver[1], MChem; Rakshitha S Kamath[1], MS, MSL; Emma Worthington[2], MPH; Minyi Lu[3], PhD; Hema Gandhi[3], PhD; Heidi C Waters[3], PhD; Daniel C Malone[4], PhD

[1]Costello Medical, Boston, MA, United States

[2]Costello Medical, Cambridge, United Kingdom

[3]Otsuka Pharmaceutical Development & Commercialization Inc, Princeton, NJ, United States

[4]Department of Pharmacotherapy, Skaggs College of Pharmacy, University of Utah, Salt Lake City, UT, United States

Corresponding Author:
Daniel C Malone, PhD
Department of Pharmacotherapy
Skaggs College of Pharmacy
University of Utah
Room 4922
30 S 2000 East
Salt Lake City, UT, 84112
United States
Phone: 1 801 581 6257
Email: dan.malone@utah.edu

Abstract

Background: Digital mental health technologies (DMHTs) have the potential to enhance mental health care delivery. However, there is little information on how DMHTs are evaluated and what factors influence their use.

Objective: A systematic literature review was conducted to understand how DMHTs are valued in the United States from user, payer, and employer perspectives.

Methods: Articles published after 2017 were identified from MEDLINE, Embase, PsycINFO, Cochrane Library, the Health Technology Assessment Database, and digital and mental health congresses. Each article was evaluated by 2 independent reviewers to identify US studies reporting on factors considered in the evaluation of DMHTs targeting mental health, Alzheimer disease, epilepsy, autism spectrum disorder, or attention-deficit/hyperactivity disorder. Study quality was assessed using the Critical Appraisal Skills Program Qualitative and Cohort Studies Checklists. Studies were coded and indexed using the American Psychiatric Association's Mental Health App Evaluation Framework to extract and synthesize relevant information, and novel themes were added iteratively as identified.

Results: Of the 4353 articles screened, data from 26 unique studies from patient, caregiver, and health care provider perspectives were included. Engagement style was the most reported theme (23/26, 88%), with users valuing DMHT usability, particularly alignment with therapeutic goals through features including anxiety management tools. Key barriers to DMHT use included limited internet access, poor technical literacy, and privacy concerns. Novel findings included the discreetness of DMHTs to avoid stigma.

Conclusions: Usability, cost, accessibility, technical considerations, and alignment with therapeutic goals are important to users, although DMHT valuation varies across individuals. DMHT apps should be developed and selected with specific user needs in mind.

(JMIR Ment Health 2024;11:e57401) doi:10.2196/57401

KEYWORDS

mental health; mobile health; mHealth; digital health; digital therapeutics; systematic review; framework synthesis; mixed methods

Introduction

Background

Digital health comprises a broad range of technologies, including mobile health, health information technology, wearable devices, and personalized medicine, which serve as tools to enhance health care delivery. Recently, several digital mental health (MH) therapeutics, a category of digital MH technologies (DMHTs), have received US Food and Drug Administration (FDA) approval to prevent, manage, or treat a medical disorder or disease based on evidence from superiority trials and compliance with technical guidelines [1,2]. However, most DMHTs, particularly apps, fall outside FDA jurisdiction because they are not intended to diagnose, treat, or prevent disease and because they are "low risk" in that they would not cause harm in the event of malfunction [3]. Due to this lack of regulatory framework, few DMHTs are supported by published efficacy studies. One study found that only 16% of MH apps recommended by college counseling centers were supported by efficacy studies published in peer-reviewed journals [4].

Nonetheless, many health care providers (HCPs) use MH apps in clinical practice. Up to 83% of behavioral health providers in a small study covering the Greater Boston area reported using apps as part of their clinical care, particularly mindfulness apps for patient anxiety management [5]. As many DMHTs are currently widely used in clinical practice without undergoing any formal assessment for quality or relevance, understanding how DMHTs should be assessed based on factors impacting their value from the perspective of key stakeholders, such as patients, caregivers, providers, payers, and employers, could improve the selection of DMHTs for use by patients, thereby increasing care quality and outcomes for those seeking MH support.

Objective

To address identified gaps, a systematic literature review (SLR) was conducted using a published framework to synthesize emerging themes from mixed methods evidence in order to understand how digital health solutions, encompassing both digital therapeutics and direct-to-consumer digital health technologies, are valued, with a focus on MH disorders, Alzheimer disease, epilepsy, autism spectrum disorder (ASD), and attention-deficit/hyperactivity disorder (ADHD) in the United States.

Methods

Overview

The SLR was performed in accordance with a prespecified protocol and reported in line with the Enhancing Transparency in Reporting the Synthesis of Qualitative Research and PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [6,7]. The protocol was not registered.

Search Strategy

Electronic databases, encompassing MEDLINE (including MEDLINE In-Process, MEDLINE Daily, and MEDLINE Epub Ahead of Print); Embase; the Cochrane Library (including Cochrane Database of Systematic Reviews and Cochrane Central Register of Controlled Trials); PsycINFO; and the Health Technology Assessment Database, were selected in alignment with this SLR's target indications and were searched on June 17, 2022. The search terms included combinations of free-text and Medical Subject Heading or Emtree terms related to indications of interest, DMHTs, and relevant outcomes or assessment types (eg, technology assessments and cost; Tables S1-S5 in Multimedia Appendix 1). Searches were limited to studies performed in the United States and to those published from 2017 onward.

Manual hand searches of gray literature, namely, the bibliographies of relevant SLRs identified from the electronic database searches and key conference proceedings (2019-2022), were performed to identify additional studies of relevance (Table S6 in Multimedia Appendix 1). The FDA website was also searched to identify factors involved in the FDA's appraisal of relevant MH apps, which could supplement the factors identified in this SLR (Table S7 in Multimedia Appendix 1).

Study Selection

Studies were included in the SLR if they met prespecified criteria defined using the SPIDER (Sample, Phenomenon of Interest, Design, Evaluation, Research type) framework, which is appropriate for mixed methods research questions. Eligible studies were published in the English language, were set in the United States, and reported quantitative or qualitative outcomes relating to the factors considered in the evaluation of DMHTs. Only studies published in 2017 or later were included because of the rapidly evolving research area. Eligible studies reported on MH, Alzheimer disease, epilepsy, ASD, or ADHD from user, payer, or employer perspectives (Table S8 in Multimedia Appendix 1). While the primary focus of the SLR was MH, neurological conditions were also of interest because their pathologies, symptoms, and treatment strategies can overlap with those of mental illnesses. Alzheimer disease, epilepsy, ASD, and ADHD were selected because they are highly researched and represent diverse types of neurological conditions.

The titles and abstracts of records were assessed for inclusion against these eligibility criteria by 2 independent reviewers, and discrepancies were resolved by consensus, with arbitration by a third reviewer if necessary. Full texts of potentially relevant articles were acquired and screened using the same methodology.

Study Prioritization

Due to the large volume of the evidence identified, additional eligibility criteria were applied to prioritize primary research on theoretical DMHT valuation factors. In line with the thematic framework synthesis objective, *theoretical valuation factors* were defined as user or DMHT attributes that impact interaction with or perception of DMHTs; therefore, studies that reported only efficacy outcomes, such as mental illness symptom improvement, were deprioritized for full-text review. Secondary research was also deprioritized for full-text review. Studies that reviewed a select app against a framework and studies that reported only the outcomes specific to a select app were deprioritized for data extraction. For example, a study reporting the usability of a *specific* app's features would have been deprioritized, while a study reporting what *types* of features increase MH app usability *in general* would not.

Data Extraction

All relevant data were extracted into a prespecified Microsoft Excel grid, and a quality assessment was performed for each study. Studies that reported only qualitative data were assessed with the Critical Appraisal Skills Program Qualitative Studies Checklist. Studies that reported only quantitative data were evaluated with the Critical Appraisal Skills Program Cohort Study Checklist, and studies reporting both qualitative and quantitative data were evaluated with both checklists [8]. Data extractions and quality assessments were performed by a single individual for each study, with the information verified by a second independent individual. Discrepancies were resolved by consensus, with arbitration by a third individual if necessary.

Framework Synthesis

A framework synthesis approach was undertaken to synthesize qualitative and quantitative data identified from the SLR. In line with the "best fit" framework synthesis approach, data were indexed deductively against an existing framework where possible, and novel themes were added inductively as needed [9,10]. The American Psychiatric Association (APA) Mental Health App Evaluation framework was considered the most appropriate framework to address the research objectives of this SLR because its key valuation themes were developed using psychiatrist and patient input, are broadly shared by other evaluation frameworks, are widely acknowledged in the literature, and have been described as durable and adaptable [11-13].

The APA model follows a hierarchical and chronological order whereby the evaluator moves through the framework using prompting questions (eg, "Does the app work offline?"). For this SLR, these questions were either thematically grouped into subthemes or left as prompting questions, as appropriate. The framework was therefore ultimately adapted into 3 levels: themes, subthemes, and more granular valuation criteria. It should be emphasized that this SLR did not aim to formally develop an updated framework to be used in practice by HCPs and their patients but rather was used to form a theoretical basis for understanding DMHT valuation factors, for which novel themes were expected to emerge.

A data-based convergent approach was used to synthesize quantitative and qualitative data [14]. Data were initially indexed deductively against the prespecified themes within the data collection instrument and then further synthesized within Docear [15], a mind-map software used to organize and connect data and concepts. Indexing was performed by 1 reviewer and checked by a second independent reviewer. New themes and subthemes that emerged from the literature through inductive coding were added post hoc to the thematic framework, with all extracted data then considered against both the prespecified and novel themes. The evidence identified for each theme was synthesized narratively, taking into consideration the context and design of each study.

Results

Included Studies

A total of 4974 records were retrieved from the electronic databases. Of the 3374 (67.83%) unique records identified following deduplication across the databases, 2891 (85.68%) were excluded based on the eligibility criteria, and an additional 456 (13.52%) were deprioritized because they were not directly related to the topic of interest for this SLR. Excluded and deprioritized full texts are listed in Tables S9 and S10 in Multimedia Appendix 1, respectively. Therefore, 27 (0.54%) articles were included from the electronic database searches. In addition, 1 article reporting on the same study as an already-included conference abstract was identified during supporting targeted searches and included as a supplementary record, resulting in a total of 28 articles reporting on 26 unique studies (Figure S1 in Multimedia Appendix 1). No relevant FDA appraisals were identified in the supplementary search.

Of the 26 included studies, 8 (31%) were quantitative, 12 (46%) were qualitative, and 6 (23%) used a mixed methods approach. While 5 (19%) studies assessed prospective cohorts, 22 (85%) used a cross-sectional approach, including 1 (4%) study that contained both a prospective cohort and a cross-sectional cohort (Table 1). All studies (26/26, 100%) investigated a user perspective, with none specifically investigating payer or employer perspectives. Only 1 (4%) study, which examined ingestible sensor pills and smart pill dispensers to track adherence, investigated a DMHT that was not an app [16].

Table 1. Summary of included study characteristics and outcomes.

Study (author, year)	Design[a]	Perspective and population[a]	Objectives	Data collection methods[a]
Afra et al [17], 2018	Cross-sectional, quantitative	• Patients with epilepsy who were regular smartphone users recruited from the University of Utah Adult Comprehensive Epilepsy Clinic (N=40)	To develop a drug-device combination product using an app in combination with antiseizure medications as an epilepsy treatment	Custom survey
Beard et al [18], 2019	Cross-sectional, quantitative	• Patients treated at a partial hospitalization program located in a nonprofit, insurance-based psychiatric hospital; diagnoses included MDD[b], BD[c], anxiety, OCD[d], stress-related disorders, and psychotic disorders (N=322)	To characterize general smartphone app and social media use in an acute transdiagnostic psychiatric sample with high smartphone ownership, characterize current engagement and interest in the use of smartphone apps to support MH[e], and test demographic and clinical predictors of smartphone use	Custom survey
Borghouts et al [19], 2022	Cross-sectional, mixed methods	• General users[f]: members of the Center on Deafness Inland Empire, comprised people with lived experience as members of the deaf or hard-of-hearing community (N=10)	To investigate the MH needs of the deaf or hard-of-hearing community and how MH apps might support these needs	Custom survey; focus group
Boster and McCarthy [20], 2018	Cross-sectional, qualitative	• Speech-language pathologists experienced in augmentative and alternative communication using a device in children with ASD[g] recruited through social media and professional listserves (N=8) • Parents (caregivers) of children with ASD recruited through national organizations (N=5)	To gain insight from speech-language pathologists and parents of children with ASD regarding appealing features of augmentative and alternative communication apps	Focus groups; poll questions
Buck et al [21], 2021a	Cross-sectional, quantitative	• Caregivers of young adult family members who experienced early psychosis (onset before age 35) recruited through HCP[h] referrals or ads (N=43)	To assess caregivers' interest in an array of specific potential mHealth[i] functions to guide the development of mHealth for caregivers of young adults with early psychosis	Custom survey
Buck et al [22], 2021b	Cross-sectional, quantitative	• Users: young adults (aged 18-30 years) with a diagnosis of a psychotic disorder or self-reported history of psychotic symptoms recruited through HCP referrals or ads (N=77)	To understand the needs, interests, and preferences of young adults with early psychosis regarding mHealth by surveying interest in mHealth features and delivery modalities and by collecting information about their digital and web-based behaviors	Custom survey
Carpenter-Song et al [23], 2018	Prospective cohort, qualitative	• Patients at a community MH center (N=15)	To examine current practices and orientations toward technology among consumers in 3 mental health settings in the United States	Semistructured interviews
Casarez et al [24], 2019	Cross-sectional, qualitative	• Caregivers: spouses or partners of patients with BD recruited from a local outpatient psychiatry clinic or psychiatric hospital (N=13)	To explore how the well-being of spouses and partners of patients with BD can be improved through mHealth technology	Focus groups; minimally structured, open-ended individual interviews
Connolly et al [25], 2018	Cross-sectional, qualitative	• Patients: US military veterans (aged 18-70 years) who screened positive for PTSD[j], alcohol use disorder, or MDD during the previous year at 9 community-based VA[k] outpatient clinics (N=66)	To examine veterans' attitudes toward smartphone apps and to assess whether openness toward this technology varies by age or rurality	Semistructured interviews informed by the State of the Art Access Model

Study (author, year)	Design[a]	Perspective and population[a]	Objectives	Data collection methods[a]
Cummings et al [26], 2019	Cross-sectional, qualitative	• Parents and grandparents with children or grandchildren (caregivers) enrolled in a public health insurance program who received ≥2 months of ADHD[l] treatment at 4 safety-net clinics (N=37) • Administrators at the same clinics (N=41)	To examine stakeholder perspectives regarding whether mHealth tools can improve MH treatment for low-income youth with ADHD in safety-net settings and what functions would improve treatment	Focus groups (caregivers) and interviews (HCPs and staff), both semistructured and including open-ended questions and targeted probes
Dinkel et al [27], 2021	Cross-sectional, qualitative	• Adult patients (aged ≥19 years) with a current or prior diagnosis of depression recruited during medical visits from 2 integrated primary care clinics (N=17) • HCPs and staff at the same clinics (N=15)	To explore patient and clinic-level perceptions of the use of depression self-management apps within an integrated primary care setting	Semistructured focus groups; semistructured interviews
Forma et al [16], 2022	Cross-sectional, quantitative	• Caregivers of patients with BD, MDD, or schizophrenia who believed their patients had adherence issues to second-generation oral atypical antipsychotic medication (N=184)	To assess caregivers' preferences and willingness to pay for digital (ingestible sensor pill, medication containers with electronic monitoring, mobile apps, and smart pill dispensers) and nondigital (medication diary and simple pill organizer) tools	Custom discrete choice experiment survey
Hoffman et al [5], 2019	Prospective interventional, mixed methods	• HCPs (N=24) in a routine primary care behavioral health setting who reported their own and patients' (sample size not reported) MH app use and feedback; patient conditions included anxiety, stress, depression, and substance use	To test the feasibility of using mHealth apps to augment integrated primary care services, solicit feedback from patients and providers to guide implementation, and develop an MH app toolkit for system-wide dissemination	Custom survey
Huberty et al [28], 2022	Cross-sectional (current Calm (Calm.com, Inc) users) and prospective interventional (nonusers of Calm, HCPs), qualitative	• General users[f]: patients with cancer and survivors of cancer with smartphones, some of whom were current subscribers of Calm, a meditation app (N=17) • HCPs, staff, and not-for-profit partners in cancer care with smartphones (N=10)	To develop a mobile meditation app prototype specifically designed for patients with cancer and survivors of cancer	Custom surveys; focus groups
Kern et al [29], 2018	Cross-sectional, quantitative	• General users[f]: students from a midwestern university with smartphones (N=721)	To investigate the potential usefulness of MH apps and attitudes toward using them	Custom survey
Knapp et al [30], 2021	Prospective cohort, qualitative	• Clinical staff members who provide behavioral health care for children and adolescents with conditions, including ADHD and depression, at a large community service organization in a midwestern state (N=37)	To learn about considerations and perspectives of community behavioral HCPs on incorporating digital tools into their clinical care for children and adolescents	Focus groups
Kornfield et al [31], 2022	Prospective cohort, qualitative	• Users: participants with at least moderate levels of depression or anxiety symptoms on the PHQ-9[m] or GAD-7[n] questionnaires, but without serious mental illnesses (eg, BD, schizophrenia), who were not receiving formal care and recruited upon completing free web-based MH self-screening surveys hosted by Mental Health America (N=28)	To investigate how digital technologies can engage young adults in self-managing their MH outside the formal care system	Web-based asynchronous discussion; synchronous web-based design workshop

Study (author, year)	Design[a]	Perspective and population[a]	Objectives	Data collection methods[a]
Lipschitz et al [32], 2019	Cross-sectional, quantitative	• Users: veterans enrolled in care at the VA Boston Healthcare System diagnosed with an anxiety disorder (including OCD), unipolar depressive disorder, or PTSD and who had at least 1 encounter in the local primary care clinic (N=149)	To assess patients' interest in mHealth interventions for MH, to identify whether provider endorsement would impact interest, to determine reasons for nonuse of mHealth interventions for MH, and to identify what mHealth content or features are of most interest to patients	Custom survey
Mata-Greve et al [33], 2021	Cross-sectional, mixed methods	• General users[f]: essential workers during the COVID-19 pandemic or workers who were unemployed or furloughed because of the COVID-19 pandemic, recruited from a web-based research platform (N=1987)	To document psychological stress, to explore DMHT[o] use in response to COVID-19–related stress, to explore the usability and user burden of DMHTs, and to explore which aspects and features of DMHTs were seen as necessary for managing stress during a pandemic by having participants design their own ideal DMHTs	Survey combining custom and validated measures (System Usability Scale, Use Burden Scale)
Melcher et al [34], 2022 and Melcher and Torous [4], 2020	Cross-sectional, mixed methods	• General users[f]: college students aged 18-25 years, recruited through social media and word of mouth (N=100)	To examine why college students show poor engagement with MH apps and how apps may be adapted to suit this population	Custom survey; interviews
Schueller et al [35], 2018	Cross-sectional, mixed methods	• General users[f]: smartphone owners recruited from a research registry (N=827)	To understand where users search for MH apps, what aspects of MH apps they find appealing, and what factors influence their decisions to use MH apps	Custom survey; focus group interviews
Schueller et al [36], 2021	Cross-sectional, qualitative	• General users[f]: participants who had used an app that allowed them to track their mood, feelings, or mental well-being for ≥2 weeks, recruited from a research registry (N=22)	To understand motivations for and experiences in using mood-tracking apps from people who used them in real-world contexts	Semistructured interviews
Stiles-Shields et al [37], 2017	Cross-sectional, qualitative	• General users[f]: participants recruited from web-based postings; approximately equal numbers of participants were above and below the criteria for a referral for psychotherapy for depression (N=20)	To identify the barriers to the use of a mobile app to deliver treatment for depression and to provide design implications on the basis of identified barriers	Card sorting task
Storm et al [38], 2021	Cross-sectional, qualitative	• Patients with diagnoses of schizophrenia, schizoaffective disorder, BD, or persistent MDD in active treatment at a community MH center (N=17) • Peer support specialists at the same center (N=15)	To identify stakeholders' perspectives on partnering to inform the software development life cycle of a smartphone health app intervention for people with serious mental illness	Semistructured interviews
Torous et al [39], 2018	Cross-sectional, quantitative	• Outpatients attending psychiatric clinics; one clinic primarily treated mood and anxiety disorders, and the other primarily treated psychotic disorders (N=185)	To understand how individuals with mental illness use their mobile phones by exploring their access to mobile phones and their use of MH apps	Custom survey
Zhou and Parmanto [40], 2020	Cross-sectional, mixed methods	• Users: participants with mild or moderate depression with local privacy concerns when using MH apps, recruited from a research registry (N=40)	To determine user preferences among the several privacy protection methods used in current mHealth apps and the reasons behind those preferences	Custom survey; interview

[a]Only information relevant to this systematic literature review is reported in this table.

[b]MDD: major depressive disorder.

[c]BD: bipolar disorder.

[d]OCD: obsessive-compulsive disorder.

[e]MH: mental health.

[f]General users are participants who were not necessarily diagnosed with indications of interest.

[g]ASD: autism spectrum disorder.

[h]HCP: Health care provider.

[i]mHealth: mobile health.

[j]PTSD: posttraumatic stress disorder.

[k]VA: Veterans Affairs.

[l]ADHD: attention-deficit/hyperactivity disorder.

[m]PHQ-9: Personal Health Questionnaire-9.

[n]GAD-7: Generalized Anxiety Disorder-7.

[o]DMHT: digital mental health technology.

Most frequently, studies focused on indications for mood, anxiety, or psychotic disorders (15/26, 58%), with other indications of focus including ADHD (2/26, 8%), ASD (1/26, 4%), and epilepsy (1/26, 4%). No relevant studies focused on Alzheimer disease were identified.

A total of 8 (31%) studies assessed the perspectives toward DMHTs of general population participants who were not necessarily diagnosed with relevant conditions [19,28,29,33-37]. Of these populations, several were identified as having an increased risk of MH conditions, such as patients with cancer [28], college students [29,34], deaf or hard-of-hearing individuals [19], and people who were unemployed or furloughed during the COVID-19 pandemic [33]. In addition, 1 (4%) study included a mix of patients who were above and below the referral criteria for psychotherapy for depression [37].

Thematic Analysis

Overview

Evidence was identified for all 5 themes included in the APA framework: engagement style (23/26, 88%), background and accessibility (16/26, 62%), privacy and security (13/26, 50%), therapeutic goal (12/26, 46%), and clinical foundation (8/26, 31%; Table 2). Five novel criteria were identified and added to the framework post hoc, 1 each under engagement style (forgetting or feeling unmotivated to use DMHTs) and privacy and security (personal image and stigma) and 3 under background and accessibility (willingness to pay, insurance restrictions, and cost savings compared with professional care).

XSL•FO
RenderX

Table 2. Studies reporting on each theme, subtheme, and criterion.

Subtheme	Criteria (study reference)
Engagement style	
Short-term usability	• Ease of use [5,25,34,35] • Available engagement styles [20-22,24,25,30,31,34,37]
Long-term usability	• Alignment of app with needs and priorities [5,16-22,24,26,28-34,36-39] • Forgot or unmotivated to use[a] [5,25,31,37]
Customizability	• No further stratification [20,24,28,31,32,34,35]
Background and accessibility	
Technical	• Offline functionality [19,23,25,27,30,37] • Compatibility with different operating systems[b] • Accessibility [5,19,25,27-29,32,35,38]
Business model[b]	• Funding sources or conflicts of interest[b]
Costs	• Additional or hidden costs [26,37] • Willingness to paya [16,25,34,35] • Insurance restrictions[a] [20] • Cost savings compared with professional care[a] [29] • No further stratification [25-27,34]
Medical claims[b]	• Specific medical claims[b] • Trustworthiness of source[b]
Stability	• Frequency of software updates [35,37]
Privacy and security	
No specific subtheme	• No further stratification [34,36,37]
Data collection and storage	• Ability to opt out of data collection or delete data[b] • Data storage location[b] • Security associated with collection, use, and transmission of sensitive data (including personal health information) [5,19,27,29,32]
Privacy policy	• Transparency and accessibility of privacy policy [34,35,38] • Declaration of data use and purpose [34] • Data sharing with third parties [25] • Systems to respond to potential harms or safety concernsb
Personal health information	• Description of use of personal health information [34] • Personal image and stigmaa [5,25,29,40]
Security measures	• Security systems used [30,35,40]
Clinical foundation	
Impressions of use	• Accuracy and relevancy of app content [25,26] • Alignment in app appearance and its claimed purpose[b]
User feedback	• Evidence of specific benefit from user feedback or user research studies [27,35] • Validation of app usability and feasibility[b]
Clinical validity	• Supporting sources or references for use cases of the app [34,35] • Evidence of specific benefit [27,34-36] • Evidence of effectiveness or efficacy [32,34,37] • Clarity in functional scope[b]
Therapeutic goal	

Subtheme	Criteria (study reference)
Clinically actionable	• Positive change or skill acquisition [5,18,26-28,30,31,34,36] • Ease of sharing and interpretation of data [25-27,30]
Therapeutic alliance	• Possibility for collaboration with an HCPc [5,21,22,27,34,36] • The therapeutic alliance between patient and HCP [5,26]
Data ownership, access, and export[b]	• User ownership of data[b] • Opportunity for sharing of data with electronic medical records and other data tools (Apple HealthKit, Fitbit)[b] • Opportunity for use with a provider and ability to export or transfer data[b]

[a]Novel findings that emerged from this systematic literature review.

[b]These subthemes and criteria were included in the American Psychiatric Association's framework but were not reported on by studies included in this systematic literature review.

[c]HCP: health care provider.

Theme 1: Engagement Style

Engagement style was the most reported theme, with evidence identified from 23 (88%) of the 26 studies. Engagement style encompasses how and why users do or do not interact with DMHTs. The long-term usability subtheme was reported by 96% (22/23) of studies, short-term usability by 12 (52%) studies, and customizability by 7 (30%) studies. Findings from short- and long-term usability subthemes were highly interconnected.

A total of 4 studies reported that ease of use promoted short-term DMHT engagement. In the study by Schueller et al [35], 89.6% of a general population of smartphone users reported ease of use for MH apps as "important" or "very important," and users qualitatively reported dislike of "overwhelming," difficult-to-navigate apps. In addition, users valued apps that were "simplistic" [34], fit into their daily schedules, and were available when needed (eg, during acute symptom experiences) [5,25]. Select supporting qualitative data are presented in Table 3.

Table 3. Select key quotes identified for systematic literature review findings.

Subtheme and criteria: findings	Key quotes
Engagement style	
Short-term usability	
Ease of use	• "I like short exercises. I can use them in different places." [Patient in routine behavioral health care] [5] • "Whenever I have one of those outbursts and frustration, I can just open it up, say 'Okay, what's my first step?'" [Male veteran, aged 26 years] [25]
Available engagement styles: use of animation and visuals	• "They love badges. And decorating their avatars, like getting a new hat...So, they're very motivated to get through their modules when they get to earn something at the end." [Pediatric behavioral health clinician] [30] • "It could become visually distracting—children preferring the animation rather than actually creating genuine, communicative messages." [Caregiver or speech-language pathologist for children with ASD[a]] [20]
Long-term usability	
Alignment of app with needs and priorities: gamification	• "I've seen some kid clients come alive because they're excited because they wanna beat their score. And just helping them like, 'How do you have to communicate? You have to keep talking. You have to keep going.' It's helped with that." [Pediatric behavioral health clinician] [30]
Alignment of app with needs and priorities: anxiety management	• "App features that could help to reduce anxiety, for example, guided meditation, breathing exercises, or positive affirmation [may be] useful." [Community MH[b] center peer support specialist] [38] • "Stuff that's purely motivational...can feel alienating if I'm depressed...but focusing on something specific, like doing a breathing exercise...would be cool." [Patient with anxiety or depression] [31]
Alignment of app with needs and priorities: tracking mood, symptoms, or sleep	• "They [the adolescent] can bring it up on their phone...and we look at just is she daily fluctuating? If so, what happened during that day?" [Pediatric behavioral health clinician] [30] • "I don't know...if he's good or he's getting better or worse or anything like that. Just everything being simple in one place, and just hit a couple of buttons and not have to write anything down will be very good." [Caregiver of a child with ADHD[c]] [26]
Alignment of app with needs and priorities: social media–like features	• "I like hearing other people's stories and what they did, and it kind of helps me feel a little better. And I kind of like bounce off it and do what they did and try these new things that they're doing." [User with anxiety or depression] [31]
Alignment of app with needs and priorities: peer support and chat functions	• "Incorporating lived experiences into a [smartphone] app and organize the [intervention] process to address lived experience because that's what it's all about." [Community MH center peer support specialist] [38] • "So maybe the bipolar individual also has access to the same app and then so they talk to each other...That way when I get home, I know ahead of time, it was an okay day today...Or if it was not a good day ok, so I know that I need to come in a little more reserved." [Spouse of an individual with BD[d]] [24]
Forgot or unmotivated to use[e]	• "For someone who may be severely depressed, or someone who needs help, [writing messages] is almost like hard to do. Because if they're having a hard time motivating or encouraging themselves, they might not feel like this is something they could do." [User with anxiety or depression] [31] • "I notice a good number of patients mentions they did not continue using in home. [...] Maybe because this area is still new for patients?" [Routine behavioral health care staff] [5] • "[My son] has one of those crazy little phones that you can do everything with. I just don't have an interest." [Female veteran aged 57 years] [25]
Background and accessibility	
Technical considerations	
Accessibility: mobility barriers	• "They can't figure out why my hands shake so bad...so trying to use a smartphone [is frustrating]...I don't have a whole lot of feeling in my hands." [Male veteran aged 40 years] [25]
Accessibility: technical literacy	• "I haven't gotten acclimated to a smartphone yet...the technology is kind of difficult to navigate." [Male veteran aged 66 years] [25]
Offline functionality: internet and mobile data access as a barrier to use	• "[A young person's smartphone] is normally one of the first things that get taken away if they do have a bad day. So, this is the thing you can use when you're having a bad day to calm down, but then mom and dad won't let you use it because you had a bad day." [Pediatric behavioral health clinician] [30] • "There have been times I think people have suggested, 'Check this app out, check that app out,' and for the most part I don't think I have...I do only have so much data." [Patient receiving psychiatric care] [23]

Subtheme and criteria: findings	Key quotes
Costs	
Willingness to pay[e]	• "If they don't have the free trial and they want money, I'm not even gonna look at it. I'm not gonna pay for something before I've gotten the chance to see if it's gonna work for me or not." [General user, smartphone owner] [35] • "...[T]hey gave the option to pay $50.00 a year. And I did that, because I liked the idea of what they were trying to do." [General user, smartphone owner] [35]
Privacy and security	
Data collection and storage	
Security associated with collection, use, and transmission of sensitive data (including personal health information)	• "I'm worried about my data." [Patient in routine behavioral health care] [5] • "Any apps that terms and conditions you're forfeiting your information as soon as you click to that to anything so and I'm not worried about getting identity theft." [Patient with current or prior depression diagnosis] [27]
Privacy policy	
Transparency and accessibility of privacy policy	• "To use a smartphone app with a client I would want to make sure it's secure before going any further." [Community MH center peer support specialist] [38]
Personal health information	
Personal image and stigma[e]	• "I worry about my virtual image. I'd feel more comfortable using an app from CHA[f] that is protected in the same way my EMR[g] is protected." [Patient in routine behavioral health care] [5]
Security measures	
Security systems used	• "The app doesn't read as something like, My Personal Diary...it reads as something that you might just pass by if you don't know what its intention is, which can be good for teenagers who are afraid of people looking into their stuff." [Pediatric behavioral health clinician] [30]
Therapeutic goal	
Clinically actionable	
Positive change or skill acquisition: apps that impart skills and encourage positive change, in an easy way	• "Great way to have patients practice exercises between sessions; both provider and patient happy to have concrete tool." [Routine behavioral health care staff] [5] • "I almost wonder, like, if you logged in, what would you like to address today, like, symptom management versus stress...You almost need, like, an emergency toolkit and then you almost need, like, your day-to-day stuff." [HCP[h] in cancer care] [28]
Ease of sharing and interpretation of data: increase of engagement and symptom reporting	• "[I] feel like sometimes I'll give parents follow up things to do while I'm not there, and they'll forget about it throughout the week, but because they're on their phone or whatever so much throughout the week, I feel like we could send them reminders or this is what we need to do before the next week. I think that that would encourage them to be more engaged, at least in the process." [Pediatric behavioral health clinician] [30]
Therapeutic alliance	
Therapeutic alliance between patient and HCP	• "Sometimes I think my training in behavioral medicine allows me to create a different tool with the patient that is more specific to them." [Routine behavioral health care staff] [5]
Clinical foundation	
Clinical validity	

Subtheme and criteria: findings	Key quotes
Evidence of specific benefit: HCP recommendations	• "My doctor tells me to use an app, I'm probably going to use it." [Patient with current or prior depression diagnosis] [27]
Evidence of specific benefit: increased usage if supported by research, academic institution, or reputable professional society	• "I would trust an app supported by my university more than a random app I found online." [General user, college student] [34] • "I think it would be helpful, too, to have like the American Psychiatric Association or something, one of those, the licensure bodies or whatever—if they had official recommendations or backing." [General user, smartphone owner] [35]

[a]ASD: autism spectrum disorder.

[b]MH: mental health.

[c]ADHD: attention-deficit/hyperactivity disorder

[d]BD: bipolar disorder.

[e]Novel criteria identified by this systematic literature review.

[f]CHA: Cambridge Health Alliance.

[g]EMR: electronic medical record.

[h]HCP: health care provider.

Users valued DMHT features that aligned with their needs and priorities, as reflected by findings within the long-term usability subtheme. Across 9 studies, quantitative and qualitative findings demonstrated high interest in anxiety management features such as relaxation tools, breathing exercises, and mindfulness or meditation activities, and 10 studies identified interest in mood, symptom, or sleep tracking (Tables 3 and 4). While most studies (24/26, 92%) focused on MH, patients with epilepsy also reported high interest in features to record seizure dates and types [17]. Importantly, users in 2 studies emphasized the need for developers to tailor DMHTs to the needs and priorities of the target population (Table 3) [28,31]. Relatedly, mixed attitudes were reported toward positive affirmations and words of encouragement, with many users expressing interest but others emphasizing the value of a human component to DMHTs or cautioning against blanket encouragement and automated messages that could feel insincere [19,25,31].

Table 4. Quantitative evidence related to anxiety management and mood, symptoms, or sleep tracking features.

Features, study, perspective, and finding	Patients, n (%)	Likert score, mean (SD)
Anxiety management		
Buck et al [22], 2021b		
Young adults with early psychosis		
Interest in skill practices for managing stress and improving mood	64 (84.2)	3.30 (0.98)[a]
Interest in skill practices for relaxation	57 (76)	3.09 (1.12)[a]
Interest in information about relaxation exercises	59 (77.6)	3.00 (1.16)[a]
Interest in information about healthy sleep practices	56 (73.7)	2.93 (1.15)[a]
Interest in mindfulness or meditation practices	44 (59.4)	2.61 (1.34)[a]
Afra et al [17], 2018		
Patients with epilepsy		
Interest in music to help seizure control	—[b] (75)	—
Interest in relaxing music that may help alleviate stress	— (68)	—
Interest in relaxing imagery that may help alleviate stress	— (40)	—
Interest in drawing or writing while listening to music	— (35)	—
Interest in practicing mindfulness	— (63)	—
Torous et al [39], 2018		
Outpatients attending a private psychiatric clinic		
Comfort level for mindfulness and therapy	—	3.75[c]
Outpatients attending a state psychiatric clinic		
Comfort level for mindfulness and therapy	—	3.17[c]
Beard et al [18], 2019		
Patients in a partial hospitalization program in a psychiatric hospital		
Current use of an MH[d] app with the primary purpose being mindfulness or meditation	— (71)	—
Mata-Greve et al [33], 2021		
Workers furloughed during COVID-19		
Most frequently endorsed mindfulness tools as a feature when provided options to build their own app	687 (67.8)	—
Essential workers employed during COVID-19		
Most frequently endorsed mindfulness tools as a feature when provided options to build their own app	584 (60)	—
Nondistressed essential workers employed or workers furloughed during COVID-19		
Most frequently endorsed mindfulness tools as a feature when provided options to build their own app	305 (61.4)	—
Distressed essential workers employed or workers furloughed during COVID-19		
Most frequently endorsed mindfulness tools as a feature when provided options to build their own app	966 (65.3)	—
Hoffman et al [5], 2019		
Staff in a routine primary care behavioral health setting		
The ability to manage mood, anxiety, or substance use through the use of DMHTs[e] was seen as a benefit of incorporating DMHTs into clinical care	13 (57)	—
Symptom, mood, or sleep trackers		
Kern et al [29], 2018		

Features, study, perspective, and finding	Patients, n (%)	Likert score, mean (SD)
General population of college students		
Willingness to use an MH app to track mood or anxiety	41 (10.3)	—
Afra et al [17], 2018		
Patients with epilepsy		
Interest in a diary to record the date of seizures	— (85)	—
Interest in a digital diary to record the type of seizure	— (73)	—
Interest in digital diary to log the missed dosages of their medications	— (78)	—
Lipschitz et al [32], 2019		
Veterans with anxiety, MDD[f], or PTSD[g]		
Interested in progress monitoring (track mood, stress, anxiety, or PTSD symptoms)	95 (63.8)	—
Subgroup of smartphone owners		
Interested in progress monitoring (track mood, stress, anxiety, or PTSD symptoms)	80 (67.2)	—
Buck et al [22], 2021b		
Young adults with early psychosis		
Interest in a feature to set and track goals	60 (78)	3.10 (1.05)[a]
Interest in a feature to track symptoms over time	70 (90.9)	3.44 (0.90)[a]
Interest in a feature to track changes in progress toward goals	66 (86.9)	3.37 (0.86)[a]
Interest in a feature to track wellness behaviors (eg, steps or activity)	48 (64.9)	2.86 (1.22)[a]
Beard et al [18], 2019		
Patients in a partial hospitalization program in a psychiatric hospital		
Current use of an MH app with the primary purpose being mood tracking	— (10)	—
Willingness to use an MH app daily to monitor condition	262 (81)	—
Subgroup with higher education		
Willingness to use an MH app daily to monitor condition	— (85)	—
Subgroup with lower education		
Willingness to use an MH app daily to monitor condition	— (77)	—
Mata-Greve et al [33], 2021		
Workers furloughed during COVID-19		
Most frequently endorsed symptom tracking (tracking sleep or mood) as a feature when provided options to build their app	605 (59.7)	—
Essential workers employed during COVID-19		
Most frequently endorsed symptom tracking (tracking sleep or mood) as a feature when provided options to build their app	555 (57)	—
Nondistressed essential workers employed or workers furloughed during COVID-19		
Most frequently endorsed symptom tracking (tracking sleep or mood) as a feature when provided options to build their app	270 (54.3)	—
Distressed essential workers employed or workers furloughed during COVID-19		
Most frequently endorsed symptom tracking (tracking sleep or mood) as a feature when provided options to build their own app	890 (60.2)	—
Torous et al [39], 2018		
Outpatients attending a private psychiatric clinic		
Comfort level for in-app symptom surveys	—	3.50[c]
Outpatients attending a state psychiatric clinic		

Features, study, perspective, and finding	Patients, n (%)	Likert score, mean (SD)
Comfort level for in-app symptom surveys	—	3.11[c]
Outpatients attending a private psychiatric clinic		
Comfort level for passive call or text monitoring	—	2.32[c]
Outpatients attending a state psychiatric clinic		
Comfort level for passive call or text monitoring	—	2.39[c]
Outpatients attending a private psychiatric clinic		
Comfort level for passive GPS monitoring	—	2.31[c]
Outpatients attending a state psychiatric clinic		
Comfort level for passive GPS monitoring	—	2.78[c]

[a]A 5-point Likert scale (0-4) was used.

[b]Not available.

[c]A 5-point Likert scale (1-5) was used.

[d]MH: mental health.

[e]DMHT: digital mental health technology.

[f]MDD: major depressive disorder.

[g]PTSD: posttraumatic stress disorder.

Both patients and caregivers expressed interest in psychoeducational content that aligned with their needs and priorities. When surveyed, >60% of veterans with anxiety or major depressive disorder (MDD), patients with epilepsy, young adults with psychosis, and essential and furloughed workers during the COVID-19 pandemic expressed interest in relevant psychoeducational content [17,22,32,33]. In contrast, only 4% of college students in another study reported using an MH app for information about MH, although an MH diagnosis was not required for study participation [29].

Caregivers of young adults with psychosis, caregivers of children with ADHD, and spouses and partners of people with bipolar disorder (BD) were all interested in information related to caring for the individual with the given disorder, such as information on psychological and pharmacological treatments, symptoms and symptom changes, and the MH system [21,24,26]. Comparatively smaller, but still notable, proportions of caregivers of patients with psychosis were interested in caregiver-focused information; for instance, 62% to 69% were interested in relaxation exercises, stress and mood management, and community events for caregivers, while 85% to 90% were interested in the aforementioned patient-focused information [21].

Information delivery–style preference was captured under the short-term usability subtheme. One study in young adults with psychosis and another study with their caregivers revealed that delivering information in a variety of formats was important; when presented with nonmutually exclusive options, >50% of both populations were interested in text content, video content, audio content, and discussion boards [21,22].

Social interaction promoted long-term engagement. Qualitatively, 3 studies found that users valued learning about similar experiences from others via social media–like features, which normalized their experiences and could provide new symptom management strategies (Table 3) [28,31,36]. Similarly, 67% of both young adults with psychosis and deaf or hard-of-hearing survey participants (N=9) reported interest in peer support via chat features [19,22]. However, a comparatively smaller proportion of veterans with anxiety or MDD (48.3% of the full cohort and 51.3% of the smartphone user subgroup) were interested in peer support [32].

Overall, users endorsed social features to support their MH. In the study by Casarez et al [24], spouses and partners of people with BD likewise desired features to communicate with other caregivers and also emphasized that DMHTs could facilitate conversation and understanding with patients, a sentiment echoed by peer support specialists by Storm et al [38] (Table 3). However, one oncology HCP cautioned that similar to support groups, “very strict guidelines of what is said” should be implemented to manage potential risks from shared social media–like content, although little additional context was provided [28].

Spouses and partners of people with BD also suggested both in-app information on accessing professional resources and direct counseling for the patient at times when other support might be inaccessible [24]. More than half of all workers, employed or unemployed during the COVID-19 pandemic, likewise endorsed links to resources, counseling, and crisis support as DMHT features, and 81.6% of young adults with psychosis endorsed a feature to communicate with professional experts [22,33]. Importantly, compared with patients attending public clinics, patients attending private psychiatric clinics expressed a higher comfort level for in-app communication with HCPs, suggesting demographic differences in the valuation of access to professional support through DMHTs [39].

A total of 9 studies reported an interest in DMHT reminders and notifications. Across 3 studies, >70% of patients or caregivers were interested in appointment reminders [17,21,22].

In addition, 73% and 68% of patients with epilepsy reported interest in reminders for medication refills and adherence, respectively [17]. Beyond apps, caregivers of patients with MDD, BD, and schizophrenia preferred an ingestible pill sensor that tracked medication adherence, physical activity, mood, and rest 9.79 (95% CI 4.81-19.9), 7.47 (95% CI 3.81-14.65), and 6.71 (95% CI 3.29-13.69) times more than a nondigital pill organizer, respectively [16]. Qualitatively, patients and caregivers also appreciated reminders, especially if reasonably timed or delivered via text messages [27,31].

Short-term DMHT engagement was also supported by games and graphics, which could communicate information in an accessible way [24], provide tools for stress management [17,33], and be used therapeutically with children [20,30]. However, some HCPs and caregivers expressed concerns that graphics and games may be distracting for certain children (Table 3) [20].

In a novel finding, 3 studies reported forgetfulness or lack of motivation as an influence on DMHT engagement. In some cases, disuse was related to stress, other MH symptoms, or poor technical literacy (Table 3) [5,25,31]. In contrast, "forgetting to use" DMHTs and "lack of motivation" were perceived as relatively small barriers to use in the study by Stiles-Shields et al [37].

The third subtheme under engagement style was customizability, which was generally valued by users; 70.9% of a general population of smartphone users noted customization was an important factor [35]. Similarly, 9.4% of all surveyed veterans and 10.9% of those with smartphones reported disliking a prior DMHT due to a lack of personalization [32]. Users specifically wanted to be able to opt out of irrelevant features, customize audiovisual and design elements, add personal notes to tracked mood data, and provide ongoing feedback to facilitate personalization [20,24,28,31,34].

Theme 2: Background and Accessibility

A total of 16 (62%) studies reported findings related to DMHT background and accessibility, which considers the developer of the DMHT, as well as functionality and accessibility. Of these, 12 (75%) studies reported on the technical considerations subtheme, 9 (56%) on costs, and 2 (13%) on stability.

Under technical considerations, 9 studies assessed diverse accessibility concerns. Broadly, Storm et al [38] emphasized that DMHTs should be developed in consideration of patients' social, cognitive, and environmental needs to avoid overwhelming users. Specifically, 2 studies reported language as a barrier. Deaf or hard-of-hearing participants recommended visual content presentation, such as videos and icons, alongside text and American Sign Language translations where possible [19]. Similarly, when discussing English-only apps, 1 provider stated as follows: "language is a barrier for some [patients]" [5]. Mobility issues related to MH symptoms or other conditions and technical literacy, such as difficulties remembering passwords and navigating smartphones or apps, created accessibility barriers as well (Table 3) [5,25,27,28]. Additional concerns included apps that restricted use based on geographic location [19], user difficulty in finding relevant, useful apps [32], and limited mobile device memory for downloading apps [5,19].

Offline functionality, reported by 6 studies, was also captured under the technical considerations subtheme. A majority (5/9, 56%) of participants included in the study by Borghouts et al [19] expressed concern about their mobile data plans when using their devices. Correspondingly, "availability of Wi-Fi" was noted as a top barrier to the use of apps for depression by Stiles-Shields et al [37], and several veterans in another study reported that home Wi-Fi connectivity facilitated app use by eliminating cellular data fees [25,37]. Quotes from patients and HCPs echoed the concern about apps without offline functionality (Table 3) [23,30].

Data fees were also captured under the costs subtheme, with hidden or additional costs described as a barrier to app use by 2 studies [26,37]. Parents of children with ADHD reported that difficulty paying phone bills could result in their phones being shut off, limiting DMHT use; one MH clinic administrator stated as follows: "We often encounter parents' phones being shut off because they haven't paid their bill...If the app were free or low cost, I imagine it could be very helpful" [26]. In addition to hidden costs, this quote identifies up-front app costs as a barrier. Quantitatively, more than half of a general population of surveyed college students expressed that cost was a top concern for the use of MH apps [34]. Qualitative findings from 2 additional studies likewise identified cost as a barrier to DMHT use [25,27].

Three novel cost attributes were identified by this SLR: willingness to pay, insurance restrictions, and cost savings compared with professional care. Four studies, 3 of which focused on apps, explored willingness to pay for DMHTs from a user perspective. Willingness to pay varied based on user preference; some surveyed college students and smartphone users among general populations valued free apps due to financial restrictions or uncertainty around app effectiveness, although 1 student commented that the quality of free trials might be inferior [34,35]. Some smartphone users also voiced a limit on how much they would be willing to spend for an app subscription (Table 3) [35]. Forma et al [16] found that caregivers were willing to pay US $255.04 (95% CI US $123.21-US $386.86) more per month for a pill with an ingestible sensor that tracked medication adherence, physical activity, and rest and could connect to an app that also collected self-reported mood data. Moreover, the caregivers were willing to pay US $124.50 (95% CI US $48.18-US $200.81) more per month for an app-connected pill organizer alone than for a nondigital pill organizer [16]. In contrast, some veterans expressed total disinterest in paid apps, with 1 user citing poor technical literacy ("don't have the knowledge") in addition to cost as affecting willingness to pay [25].

In another novel finding, a speech-language pathologist working with children with ASD preferred a single app including multiple features over separate apps for particular features due to insurance restrictions: "I agree that teaching Apps should be an in-App feature versus their own app because sometimes insurance doesn't allow us to open the iPads purchased through insurance" [20]. Although no further detail was provided for this finding, it suggests that there may be restrictions on the use of other apps on devices purchased under insurance, which may have implications for DMHT use in formal care settings due to the lack of financial support.

In a third novel cost-related finding, a small number of participants from a general population of students (3.6%) in one study preferred using an MH app to seeing an MH professional due to cost savings [29].

A total of 13% (2/16) of studies reported on the subtheme of app stability and technical difficulties, with crashes and poor display quality decreasing DMHT value [35,37]. Participants in the study by Schueller et al [35] reported that technical difficulties were often an issue for apps developed by medical institutions, which might be effective and safe but less usable than apps from other developers.

Theme 3: Privacy and Security

A total of 13 (50%) out of 26 studies reported findings related to the privacy and security theme, which covered the use and protection of user data by DMHTs. Subthemes were reported relatively equally: data collection and storage (5/13, 38%), personal health information (PHI; 5/13, 38%), privacy policies (4/13, 31%), general privacy (3/13, 23%), and security measures (3/13, 23%).

Quantitative and qualitative findings on general privacy (ie, evidence not categorized under any specific subtheme), the data collection and storage subtheme, and the privacy policies subtheme revealed heterogeneous concerns (Table 3). A total of 74% of a general population of college students reported privacy as a top concern for MH apps, although further details on the specific area of concern were unclear [34]. In the study by Stiles-Shields et al [37], participants were highly concerned with data access but less so with general privacy. Echoing the concerns about data collection and storage, 59.1% of veterans with anxiety or MDD in 1 study were concerned about in-app PHI protection [32]; however, a qualitative study in veterans with posttraumatic stress disorder, alcohol use disorder, or MDD reported that a relatively small number of participants expressed privacy concerns. In the latter study, reasons for the concerns included distrust in Veterans Affairs, belief that digital data are inherently not confidential, and fear of phone hacking [25]. From an HCP perspective, none of the surveyed behavioral health HCPs agreed with the statement “My patients are concerned about data security,” despite multiple patients within the same study reporting privacy concerns [5].

Still, privacy policies were important overall, with 70.5% of smartphone MH app users rating having a privacy policy as “very important” or “important” [35]. Melcher et al [34] found that although users valued data protection, some reported a lack of awareness about data privacy, and others were concerned about obscure privacy policies and PHI use. As noted in the data collection and storage subtheme, veteran concerns about government use of PHI were heterogeneous [25].

A novel valuation factor not included in the APA framework related to user concern with PHI privacy and security regarding MH diagnoses and MH app use is a desire to upkeep their personal image or avoid stigma (Table 3) [5,25,29,40]. For instance, 21.1% of a general college student population preferred MH app use to seeing an MH professional due to anonymity or reduced stigma [29]. One participant in a study of Veterans Affairs health service users described access to professional care via MH apps as convenient because they could avoid disclosing their use of MH services to explain leaving work early for an appointment [25].

In line with the overarching concern about PHI privacy and security, users valued app security measures. Schueller et al [35] reported that 74.2% of users rated data encryption as “important” or “very important.” Users in another study perceived the level of privacy protection as the highest for apps using a combination of a generic app name (ie, not reflecting the indicated MH disorder); easily hidden modules; and secure, user-authenticated web portals for making module changes [40]. Behavioral health clinic staff echoed the importance of discreet MH app names (Table 3) [30].

Theme 4: Therapeutic Goal

There were 12 (46%) studies that reported on the factors relating to the integration of DMHTs with users’ therapeutic goals. The clinical actionability and therapeutic alliance subthemes were reported by 83% (10/12) and 58% (7/12) of studies, respectively.

A total of 9 studies reported the value of clinically actionable insights from apps where the users could acquire and practice new skills to make positive changes in their lives (Table 3). For instance, patient and caregiver app users reported interests in “daily tips,” “new ideas,” and “solutions or recommendations” for symptom management [26,27,36]. Furthermore, an app that could serve as a resource for multiple management strategies was preferable [26,28,31]. Quantitatively, 4% of patients receiving acute treatment in a partial hospitalization program for MH conditions, including mood and psychotic disorders, reported that the primary purpose of their DMHT use was therapy skills practice [18]. HCPs similarly appreciated that DMHTs could facilitate patients practicing skills outside of formal treatment sessions [5]. In particular, clinicians from a youth behavioral health clinic noted that DMHTs might be especially beneficial for young users because they could be conveniently and discreetly incorporated into their daily lives [30].

Users valued easy data sharing with clinicians, particularly for mood- or symptom-tracking features, which could improve communication and the accuracy of symptom reporting during clinical visits [5,25-27,34,36]. For instance, 53% of a general college student population believed that the potential to share information with their clinician was “one of the top benefits” of using DMHTs [34]. In addition, many HCPs reported active use or interest in the use of DMHTs in clinical practice to facilitate asynchronous communication and increase patient engagement with treatments outside of formal appointments; however, some preferred traditional care strategies for their personalization and flexibility (Table 3) [5,26,30].

Theme 5: Clinical Foundation

A total of 8 (31%) studies reported findings related to the clinical foundation of DMHTs, that is, their utility and appropriateness for patients. Clinical validity was the most reported subtheme, with evidence identified from 6 (75%) studies; 2 (25%) studies reported on the user feedback subtheme and 2 (25%) on the impressions of use subtheme, which captured users' perceptions of app content as accurate and relevant.

Across subthemes, users valued evidence of DMHT benefit or efficacy from various sources. A total of 71.8% of surveyed veterans said that they would use a DMHT if they "saw proof that it worked" for their MH conditions [32]. Similarly, among the 811 general population participants surveyed, 69.5% ranked direct research evidence as "important" or "very important" for DMHT, and 66.8% ranked indirect research evidence the same [35]. Qualitative data identified recommendations from HCPs or academic institutions, as well as evidence of DMHT benefit from publications or research studies, as specific sources for clinically valid evidence of benefits (Table 3) [27,34,35].

In addition to academic and professional support, the user feedback subtheme captured user interest in whether DMHTs were beneficial for peers or recommended by other trusted individuals. Patients with depression reported that other users' experiences influenced their app use, with one user wanting to know "...if other people had success using it" [27]. Quantitatively, user ratings and user reviews were ranked as "important" or "very important" factors in DMHT use by 59.4% and 58.7% of the general population participants, respectively [35].

Quality Assessment

The risk of bias was overall moderate. Of the 14 studies including quantitative components, only 1 (7%) used relevant validated outcome measurement instruments [33]; all others used custom questionnaires. Of the 18 studies with qualitative components, 4 (22%) were at risk of selection bias due to participants being exclusively recruited using web-based postings and research registries [33-35,37], and only 1 (6%) considered the relationship between researcher and participant when interpreting the results [36]. Full quality assessments for qualitative and quantitative study components can be found in Tables S11 and S12 in Multimedia Appendix 1, respectively.

Discussion

Principal Findings

This SLR aimed to identify and synthesize qualitative and quantitative evidence on how DMHTs are valued by users, payers, and employers in the United States. Evidence from users with or without diagnosed relevant disorders, caregivers, and HCPs was captured across a wide range of demographics. No study reported evaluating an app from a payer or employer perspective. Furthermore, all but one included study focused on mobile apps.

No relevant appraisals of DMHTs were identified from the FDA website searches; however, 8 relevant FDA approval labels or notifications for MH apps or guidance documents for industry and FDA staff were identified. The content of these materials overlapped with some valuation factors identified in this SLR, including evidence of clinical efficacy and safety, app maintenance, and privacy and security.

Engagement style, although not covered by the FDA materials, was the most reported theme by the studies included in this SLR and was found to overlap heavily with other themes. Engagement may be a key consideration for app developers, as app user retention can be low: 1 study showed that >90% of users had abandoned free MH apps within 30 days of installation [41]. Engagement is also a key clinical concern in terms of DMHT efficacy; one meta-analysis of 25 studies showed that increased use of DMHT modules was significantly associated with positive outcomes regardless of the target MH condition [42]. The findings of this SLR may therefore be informative to both DMHT designers and HCPs who integrate DMHTs into clinical care by providing insight on DMHT valuation and thus how use and benefit can be improved. For instance, users valued DMHTs that were easy to use and aligned with their needs and priorities, particularly through features that supported their therapeutic goals. In addition, content presented through multiple delivery modes, such as both text and visuals, promoted engagement as well as accessibility.

However, engagement and feature preference varied across populations. For instance, DMHT valuation was affected by technical literacy, which may relate to user demographics; in this SLR, veterans repeatedly emphasized technical literacy as a barrier to DMHT use [25]. Similarly, offline functionality may be more important for some users. Although 85% of the total United States population owns smartphones, only 59% of Medicare beneficiaries have access to a smartphone with a wireless plan. Moreover, beneficiaries who are older, less educated, disabled, or Black or Hispanic have even lower digital access [43,44]. These findings emphasize the importance of customizability and suggest that app development and selection in the clinical setting should consider the demographics of the target population, particularly in relation to ease of use and offline functionality.

Background and accessibility findings also identified up-front and hidden costs as barriers to DMHT use, with the willingness to pay varying among individuals. This has important implications for app development, considering that many MH apps currently on the market are direct-to-consumer sales and require out-of-pocket payment. App developers often take this approach as it does not require the accumulation of formal evidence of clinical benefit for FDA approval [45], but it may present a financial barrier to use for consumers.

Privacy and security, reported by 13 (50%) out of 26 studies, was a prevalent theme, with users primarily concerned with data and PHI security within apps. This finding reflects wider research; a 2019 review of 116 depression-related apps retrieved from iTunes and Google Play stores in 2017 found that only 4% of the identified apps had acceptable transparency in privacy and security, with many completely lacking a privacy policy [46]. Similarly, 39% of MH apps recommended by college counseling centers had no privacy policy, and of those with a policy, 88% collected user data, and 49% shared that data with third parties [4]. Most evidence identified in this SLR under this theme, as well as findings previously published in the wider literature, focuses on these remote privacy risks. However, local privacy concerns are also important to users. In particular, inconspicuous naming and the ability to hide sensitive modules within MH apps were rated as highly important by both patients and HCPs to maintain user privacy. Users emphasized a desire to avoid the stigma associated with mental illness, which was also reflected by the findings in the engagement style theme: more young adults with psychosis were more interested in in-app messaging with other patients in psychosis recovery (67.1%) than a provider and family member together (47.3%) or their personal support network (59.8%) [22]. Similarly, youths were interested in apps that could be used discreetly in school or other public settings to avoid potential MH stigma. This is a key, novel finding of this SLR, considering that many app or DMHT components on the market are named after their target disorder.

The use of DMHTs to achieve therapeutic goals was discussed from patient, caregiver, and HCP perspectives, all of which valued DMHTs that had evidence of efficacy, presented clinically actionable information, and facilitated patient-clinician relationships. Of the 5 studies that explored how HCPs value DMHTs in clinical practice, 2 (40%) were restricted to the oncology or ASD settings and were not readily generalizable to wider MH settings [20,28]. In other studies, providers reported interest in using DMHTs to facilitate asynchronous communication with patients and their caregivers, promote patient skill practice, and improve care for children through the use of games and visuals [26,30]. However, while HCPs overall believed that DMHTs improved care, some believed that their clinical training allowed for care personalization beyond what DMHTs could provide. Feature customizability and receipt of input from HCPs and users during app development and testing may be a way to mitigate these concerns, as well as concerns about safety and efficacy, as many available apps do not appropriately address user health concerns [47].

Findings additionally suggested that training and resources on DMHTs would be beneficial to ensure that HCPs were equipped to integrate DMHTs into their practices [5]. Collaboration between DMHT specialists and HCPs, along with a shift from randomized controlled trials to effectiveness-implementation hybrid trials, may be a way to streamline the integration of DMHTs into clinical care and provide more training and resources for HCPs [30,48].

Strengths

This review followed a prespecified protocol and used systematic methods in line with the York Centre for Reviews and Dissemination guidelines [49] to conduct an exhaustive search of the literature, identifying evidence relevant to the review objectives from multiple databases and supplementary sources. The framework synthesis approach allowed for the inclusion and analysis of both qualitative and quantitative data, providing a detailed picture of not only what DMHT features users value but *why* they value them, especially in areas where valuation varies across patient demographics. In addition, the APA framework is a robust model created with patient and HCP input that incorporates key valuation themes broadly shared by other frameworks and widely acknowledged in the literature [11-13].

Limitations

Methodological limitations should be considered when interpreting the findings of this SLR. Only publications in English and in United States populations were included. As perceptions of value are influenced by factors including cultures, laws, and health care settings, the findings of this SLR should not be generalized to other countries. For instance, trust in HCPs and rates of longstanding relationships between patients and primary care providers are lower in the United States than in many European nations [50,51], which could impact the type of support users want from DMHTs (ie, engagement style) or interest in DMHT integration with therapeutic goals.

In addition to the prespecified eligibility criteria, deprioritization strategies were implemented due to the large volume of the identified evidence, and this may have resulted in missing relevant articles. In particular, the deprioritization of secondary research and opinion pieces likely led to the exclusion of relevant discussion around payer perspectives and reimbursement, for which no evidence was included in this SLR. Furthermore, although unlikely, there may have been reporting biases in the included studies due to missing results, which this SLR was not able to assess.

This SLR identified no evidence for 3 subthemes included in the APA framework: business model (background and accessibility), which covers DMHT funding sources and potential sources of conflict, medical claims (background and accessibility), which examines whether DMHTs claim to be medical and the trustworthiness of their creators, and data ownership, access, and export (therapeutic goal), which includes sharing data with eHealth records or wellness devices (eg, Apple HealthKit [Apple Inc], Fitbit [Google LLC]). The valuation of these subthemes should be evaluated in future research.

Conclusions

In summary, app usability, cost, accessibility and other technical considerations, and alignment with therapeutic goals were the most reported valuation factors identified by this SLR. Many studies also reported user preference for apps that incorporated privacy and security features that provided protection from stigma. However, individual DMHTs and their features are valued differently across individuals based on demographics and personal preferences. MH apps should be developed and selected with these specific user needs in mind. Feature customizability and input from users and HCPs during development may improve app usability and clinical benefit.

Acknowledgments

The authors thank Max Lee, Costello Medical, US, for medical writing and editorial assistance based on the authors' input and direction.

Conflicts of Interest

DCM is a consultant for Otsuka Pharmaceutical Development & Commercialization (OPDC) Inc for this project and has received consulting funds from Pear Therapeutics, Sanofi, Avidity, Sarepta, Novartis, and BioMarin. ML, HG, and HCW are employees of OPDC. JC, SB, RSK, and EW are employees of Costello Medical. This research was funded by OPDC.

Multimedia Appendix 1
Electronic database and supplementary search terms, systematic literature review eligibility criteria, publications excluded or deprioritized at full-text review, quality assessments of included studies, and the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the identified publications.
[DOCX File , 189 KB - mental_v11i1e57401_app1.docx]

Multimedia Appendix 2
PRISMA checklist.
[PDF File (Adobe PDF File), 119 KB - mental_v11i1e57401_app2.pdf]

References

1. Content of premarket submissions for device software functions: guidance for industry and Food and Drug Administration staff. U.S. Food and Drug Administration. 2023 Jun 14. URL: https://www.fda.gov/media/153781/download [accessed 2024-07-19]
2. Patel NA, Butte AJ. Characteristics and challenges of the clinical pipeline of digital therapeutics. NPJ Digit Med 2020 Dec 11;3(1):159 [FREE Full text] [doi: 10.1038/s41746-020-00370-8] [Medline: 33311567]
3. Policy for device software functions and mobile medical applications: guidance for industry and Food and Drug Administration staff. U.S. Food and Drug Administration. URL: https://www.fda.gov/media/80958/download [accessed 2023-01-06]
4. Melcher J, Torous J. Smartphone apps for college mental health: a concern for privacy and quality of current offerings. Psychiatr Serv 2020 Nov 01;71(11):1114-1119. [doi: 10.1176/appi.ps.202000098] [Medline: 32664822]
5. Hoffman L, Benedetto E, Huang H, Grossman E, Kaluma D, Mann Z, et al. Augmenting mental health in primary care: a 1-year study of deploying smartphone apps in a multi-site primary care/behavioral health integration program. Front Psychiatry 2019;10:94 [FREE Full text] [doi: 10.3389/fpsyt.2019.00094] [Medline: 30873053]
6. Tong A, Flemming K, McInnes E, Oliver S, Craig J. Enhancing transparency in reporting the synthesis of qualitative research: ENTREQ. BMC Med Res Methodol 2012 Nov 27;12(1):181 [FREE Full text] [doi: 10.1186/1471-2288-12-181] [Medline: 23185978]
7. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009 Jul 21;339(jul21 1):b2535 [FREE Full text] [doi: 10.1136/bmj.b2535] [Medline: 19622551]
8. CASP critical appraisal checklists. Critical Appraisal Skills Programme. URL: https://casp-uk.net/casp-tools-checklists/ [accessed 2022-01-06]
9. Carroll C, Booth A, Cooper K. A worked example of "best fit" framework synthesis: a systematic review of views concerning the taking of some potential chemopreventive agents. BMC Med Res Methodol 2011 Mar 16;11(1):29 [FREE Full text] [doi: 10.1186/1471-2288-11-29] [Medline: 21410933]
10. Carroll C, Booth A, Leaviss J, Rick J. "Best fit" framework synthesis: refining the method. BMC Med Res Methodol 2013 Mar 13;13(1):37 [FREE Full text] [doi: 10.1186/1471-2288-13-37] [Medline: 23497061]
11. Kolasa K, Kozinski G. How to value digital health interventions? a systematic literature review. Int J Environ Res Public Health 2020 Mar 23;17(6):2119 [FREE Full text] [doi: 10.3390/ijerph17062119] [Medline: 32209988]
12. Lagan S, Aquino P, Emerson MR, Fortuna K, Walker R, Torous J. Actionable health app evaluation: translating expert frameworks into objective metrics. NPJ Digit Med 2020;3:100 [FREE Full text] [doi: 10.1038/s41746-020-00312-4] [Medline: 32821855]
13. Lagan S, Emerson MR, King D, Matwin S, Chan SR, Proctor S, et al. Mental health app evaluation: updating the American Psychiatric Association's framework through a stakeholder-engaged workshop. Psychiatr Serv 2021 Sep 01;72(9):1095-1098. [doi: 10.1176/appi.ps.202000663] [Medline: 33882716]
14. Hong QN, Pluye P, Bujold M, Wassef M. Convergent and sequential synthesis designs: implications for conducting and reporting systematic reviews of qualitative and quantitative evidence. Syst Rev 2017 Mar 23;6(1):61 [FREE Full text] [doi: 10.1186/s13643-017-0454-2] [Medline: 28335799]
15. Beel J, Gipp B, Langer S, Genzmehr M. Docear: an academic literature suite for searching, organizing and creating academic literature. In: Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries. 2011 Presented at: JCDL '11; June 13-17, 2011; Ottawa, ON p. 565-566 URL: https://dl.acm.org/doi/10.1145/1998076.1998188 [doi: 10.1145/1998076.1998188]

16. Forma F, Chiu K, Shafrin J, Boskovic DH, Veeranki SP. Are caregivers ready for digital? caregiver preferences for health technology tools to monitor medication adherence among patients with serious mental illness. Digit Health 2022;8:20552076221084472 [FREE Full text] [doi: 10.1177/20552076221084472] [Medline: 35295765]
17. Afra P, Bruggers CS, Sweney M, Fagatele L, Alavi F, Greenwald M, et al. Mobile software as a medical device (SaMD) for the treatment of epilepsy: development of digital therapeutics comprising behavioral and music-based interventions for neurological disorders. Front Hum Neurosci 2018;12:171 [FREE Full text] [doi: 10.3389/fnhum.2018.00171] [Medline: 29780310]
18. Beard C, Silverman AL, Forgeard M, Wilmer MT, Torous J, Björgvinsson T. Smartphone, social media, and mental health app use in an acute transdiagnostic psychiatric sample. JMIR Mhealth Uhealth 2019 Jun 07;7(6):e13364 [FREE Full text] [doi: 10.2196/13364] [Medline: 31199338]
19. Borghouts J, Neary M, Palomares K, de Leon C, Schueller SM, Schneider M, et al. Understanding the potential of mental health apps to address mental health needs of the deaf and hard of hearing community: mixed methods study. JMIR Hum Factors 2022 Apr 11;9(2):e35641 [FREE Full text] [doi: 10.2196/35641] [Medline: 35404259]
20. Boster JB, McCarthy JW. Designing augmentative and alternative communication applications: the results of focus groups with speech-language pathologists and parents of children with autism spectrum disorder. Disabil Rehabil Assist Technol 2018 May 10;13(4):353-365. [doi: 10.1080/17483107.2017.1324526] [Medline: 28488902]
21. Buck B, Chander A, Monroe-DeVita M, Cheng SC, Stiles B, Ben-Zeev D. Mobile health for caregivers of young adults with early psychosis: a survey study examining user preferences. Psychiatr Serv 2021 Aug 01;72(8):955-959 [FREE Full text] [doi: 10.1176/appi.ps.202000193] [Medline: 34235943]
22. Buck B, Chander A, Tauscher J, Nguyen T, Monroe-DeVita M, Ben-Zeev D. mHealth for young adults with early psychosis: user preferences and their relationship to attitudes about treatment-seeking. J Technol Behav Sci 2021;6(4):667-676. [doi: 10.1007/s41347-021-00223-5] [Medline: 34604506]
23. Carpenter-Song E, Noel VA, Acquilano SC, Drake RE. Real-world technology use among people with mental illnesses: qualitative study. JMIR Ment Health 2018 Nov 23;5(4):e10652 [FREE Full text] [doi: 10.2196/10652] [Medline: 30470681]
24. Casarez RL, Barlow E, Iyengar SM, Soares JC, Meyer TD. Understanding the role of m-health to improve well-being in spouses of patients with bipolar disorder. J Affect Disord 2019 May 01;250:391-396. [doi: 10.1016/j.jad.2019.03.035] [Medline: 30877862]
25. Connolly SL, Miller CJ, Koenig CJ, Zamora KA, Wright PB, Stanley RL, et al. Veterans' attitudes toward smartphone app use for mental health care: qualitative study of rurality and age differences. JMIR Mhealth Uhealth 2018 Aug 22;6(8):e10748 [FREE Full text] [doi: 10.2196/10748] [Medline: 30135050]
26. Cummings JR, Gaydos LM, Mensa-Kwao A, Song M, Blake SC. Perspectives on caregiver-focused mHealth technologies to improve mental health treatment for low-income youth with ADHD. J Technol Behav Sci 2019 Mar 9;4(1):6-16 [FREE Full text] [doi: 10.1007/s41347-018-0066-2] [Medline: 31106258]
27. Dinkel D, Harsh Caspari J, Fok L, Notice M, Johnson DJ, Watanabe-Galloway S, et al. A qualitative exploration of the feasibility of incorporating depression apps into integrated primary care clinics. Transl Behav Med 2021 Sep 15;11(9):1708-1716. [doi: 10.1093/tbm/ibab075] [Medline: 34231855]
28. Huberty J, Bhuiyan N, Neher T, Joeman L, Mesa R, Larkey L. Leveraging a consumer-based product to develop a cancer-specific mobile meditation app: prototype development study. JMIR Form Res 2022 Jan 14;6(1):e32458 [FREE Full text] [doi: 10.2196/32458] [Medline: 35029528]
29. Kern A, Hong V, Song J, Lipson SK, Eisenberg D. Mental health apps in a college setting: openness, usage, and attitudes. Mhealth 2018 Jun;4:20 [FREE Full text] [doi: 10.21037/mhealth.2018.06.01] [Medline: 30050916]
30. Knapp AA, Cohen K, Nicholas J, Mohr DC, Carlo AD, Skerl JJ, et al. Integration of digital tools into community mental health care settings that serve young people: focus group study. JMIR Ment Health 2021 Aug 19;8(8):e27379 [FREE Full text] [doi: 10.2196/27379] [Medline: 34420928]
31. Kornfield R, Meyerhoff J, Studd H, Bhattacharjee A, Williams JJ, Reddy MC, et al. Meeting users where they are: user-centered design of an automated text messaging tool to support the mental health of young adults. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022 Presented at: CHI '22; April 29-May 5 2022; New Orleans, LA p. 1-6 URL: https://dl.acm.org/doi/abs/10.1145/3491102.3502046
32. Lipschitz J, Miller CJ, Hogan TP, Burdick KE, Lippin-Foster R, Simon SR, et al. Adoption of mobile apps for depression and anxiety: cross-sectional survey study on patient interest and barriers to engagement. JMIR Ment Health 2019 Jan 25;6(1):e11334 [FREE Full text] [doi: 10.2196/11334] [Medline: 30681968]
33. Mata-Greve F, Johnson M, Pullmann MD, Friedman EC, Griffith Fillipo I, Comtois KA, et al. Mental health and the perceived usability of digital mental health tools among essential workers and people unemployed due to COVID-19: cross-sectional survey study. JMIR Ment Health 2021 Aug 05;8(8):e28360 [FREE Full text] [doi: 10.2196/28360] [Medline: 34081592]
34. Melcher J, Camacho E, Lagan S, Torous J. College student engagement with mental health apps: analysis of barriers to sustained use. J Am Coll Health 2022 Oct 13;70(6):1819-1825. [doi: 10.1080/07448481.2020.1825225] [Medline: 33048626]

35. Schueller SM, Neary M, O'Loughlin K, Adkins EC. Discovery of and interest in health apps among those with mental health needs: survey and focus group study. J Med Internet Res 2018 Jun 11;20(6):e10141 [FREE Full text] [doi: 10.2196/10141] [Medline: 29891468]
36. Schueller SM, Neary M, Lai J, Epstein DA. Understanding people's use of and perspectives on mood-tracking apps: interview study. JMIR Ment Health 2021 Aug 11;8(8):e29368 [FREE Full text] [doi: 10.2196/29368] [Medline: 34383678]
37. Stiles-Shields C, Montague E, Lattie EG, Kwasny MJ, Mohr DC. What might get in the way: barriers to the use of apps for depression. Digit Health 2017 Jun 08;3:2055207617713827 [FREE Full text] [doi: 10.1177/2055207617713827] [Medline: 29942605]
38. Storm M, Venegas M, Gocinski A, Myers A, Brooks J, Fortuna KL. Stakeholders' perspectives on partnering to inform the software development lifecycle of smartphone applications for people with serious mental illness: enhancing the software development lifecycle through stakeholder engagement. In: Proceedings of the 2021 IEEE Global Humanitarian Technology Conference. 2021 Presented at: GHTC '21; October 19-23, 2021; Seattle, WA p. 195-199 URL: https://ieeexplore.ieee.org/document/9612444 [doi: 10.1109/ghtc53159.2021.9612444]
39. Torous J, Wisniewski H, Liu G, Keshavan M. Mental health mobile phone app usage, concerns, and benefits among psychiatric outpatients: comparative survey study. JMIR Ment Health 2018 Nov 16;5(4):e11715 [FREE Full text] [doi: 10.2196/11715] [Medline: 30446484]
40. Zhou L, Parmanto B. User preferences for privacy protection methods in mobile health apps: a mixed-methods study. Int J Telerehabil 2020 Dec 08;12(2):13-26 [FREE Full text] [doi: 10.5195/ijt.2020.6319] [Medline: 33520091]
41. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. J Med Internet Res 2019 Sep 25;21(9):e14567 [FREE Full text] [doi: 10.2196/14567] [Medline: 31573916]
42. Gan DZ, McGillivray L, Han J, Christensen H, Torok M. Effect of engagement with digital interventions on mental health outcomes: a systematic review and meta-analysis. Front Digit Health 2021;3:764079 [FREE Full text] [doi: 10.3389/fdgth.2021.764079] [Medline: 34806079]
43. Mobile fact sheet. Pew Research Center. 2021. URL: https://www.pewresearch.org/internet/fact-sheet/mobile/ [accessed 2024-04-29]
44. Roberts ET, Mehrotra A. Assessment of disparities in digital access among Medicare beneficiaries and implications for telemedicine. JAMA Intern Med 2020 Oct 01;180(10):1386-1389 [FREE Full text] [doi: 10.1001/jamainternmed.2020.2666] [Medline: 32744601]
45. Powell AC, Torous JB, Firth J, Kaufman KR. Generating value with mental health apps. BJPsych Open 2020 Feb 05;6(2):e16 [FREE Full text] [doi: 10.1192/bjo.2019.98] [Medline: 32019619]
46. O'Loughlin K, Neary M, Adkins EC, Schueller SM. Reviewing the data security and privacy policies of mobile apps for depression. Internet Interv 2019 Mar;15:110-115 [FREE Full text] [doi: 10.1016/j.invent.2018.12.001] [Medline: 30792962]
47. Akbar S, Coiera E, Magrabi F. Safety concerns with consumer-facing mobile health applications and their consequences: a scoping review. J Am Med Inform Assoc 2020 Feb 01;27(2):330-340 [FREE Full text] [doi: 10.1093/jamia/ocz175] [Medline: 31599936]
48. Curran GM, Bauer M, Mittman B, Pyne JM, Stetler C. Effectiveness-implementation hybrid designs: combining elements of clinical effectiveness and implementation research to enhance public health impact. Med Care 2012 Mar;50(3):217-226 [FREE Full text] [doi: 10.1097/MLR.0b013e3182408812] [Medline: 22310560]
49. Systematic reviews: CRD's guidance for undertaking reviews in health care. Centre for Reviews and Dissemination. 2008. URL: https://www.york.ac.uk/media/crd/Systematic_Reviews.pdf [accessed 2024-04-29]
50. Gumas ED, Lewis C, Horstman C, Gunja MZ. Finger on the pulse: the state of primary care in the U.S. and nine other countries. The Commonwealth Fund. URL: https://www.commonwealthfund.org/publications/issue-briefs/2024/mar/finger-on-pulse-primary-care-us-nine-countries [accessed 2024-04-29]
51. Huang EC, Pu C, Chou YJ, Huang N. Public trust in physicians-health care commodification as a possible deteriorating factor: cross-sectional analysis of 23 countries. Inquiry 2018 Mar 05;55:46958018759174 [FREE Full text] [doi: 10.1177/0046958018759174] [Medline: 29502479]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
APA: American Psychiatric Association
ASD: autism spectrum disorder
BD: bipolar disorder
DMHT: digital mental health technology
FDA: Food and Drug Administration
HCP: health care provider
MDD: major depressive disorder
MH: mental health

PHI: personal health information
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SLR: systematic literature review
SPIDER: Sample, Phenomenon of Interest, Design, Evaluation, Research type

Edited by J Torous; submitted 15.02.24; peer-reviewed by A Mathieu-Fritz, K Stawarz; comments to author 05.05.24; revised version received 20.06.24; accepted 21.06.24; published 30.08.24.

Please cite as:
Catania J, Beaver S, Kamath RS, Worthington E, Lu M, Gandhi H, Waters HC, Malone DC
Evaluation of Digital Mental Health Technologies in the United States: Systematic Literature Review and Framework Synthesis
JMIR Ment Health 2024;11:e57401
URL: https://mental.jmir.org/2024/1/e57401
doi:10.2196/57401
PMID:

Review

Preventive Interventions for Internet Addiction in Young Children: Systematic Review

Yansen Theopilus[1,2], MEng; Abdullah Al Mahmud[1], PhD; Hilary Davis[3], PhD; Johanna Renny Octavia[2], PhD

[1]Centre for Design Innovation, Swinburne University of Technology, Melbourne, Australia

[2]Centre for Ergonomics, Parahyangan Catholic University, Bandung, Indonesia

[3]Centre for Social Impact, Swinburne University of Technology, Melbourne, Australia

Corresponding Author:
Abdullah Al Mahmud, PhD
Centre for Design Innovation
Swinburne University of Technology
John St Hawthorn
Melbourne, 3122
Australia
Phone: 61 39214383
Email: aalmahmud@swin.edu.au

Abstract

Background: In this digital age, children typically start using the internet in early childhood. Studies highlighted that young children are vulnerable to internet addiction due to personal limitations and social influence (eg, family and school). Internet addiction can have long-term harmful effects on children's health and well-being. The high risk of internet addiction for vulnerable populations like young children has raised questions about how best to prevent the problem.

Objective: This review study aimed to investigate the existing interventions and explore future directions to prevent or reduce internet addiction risks in children younger than 12 years.

Methods: The systematic review was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. We searched for relevant literature from 4 research databases (Scopus, Web of Science, PubMed, and PsycINFO). We included 14 primary studies discussing the interventions to prevent or reduce internet addiction risks in young children and their efficacy outcomes.

Results: The preventive interventions identified were categorized into four approaches as follows: (1) children's education, (2) parenting strategy, (3) strategic physical activity, and (4) counseling. Ten interventions showed promising efficacy in preventing or reducing internet addiction risks with small-to-medium effect sizes. Interventions that enhance children's competencies in having appropriate online behaviors and literacy were more likely to show better efficacy than interventions that force children to reduce screen time. Interventions that shift children's focus from online activities to real-world activities also showed promising efficacy in reducing engagement with the internet, thereby preventing addictive behaviors. We also identified the limitations of each approach (eg, temporariness, accessibility, and implementation) as valuable considerations in developing future interventions.

Conclusions: The findings suggest the need to develop more sustainable and accessible interventions to encourage healthy online behaviors through education, appropriate parenting strategies, and substitutive activities to prevent children's overdependence on the internet. Developing digital tools and social support systems can be beneficial to improve the capability, efficiency, and accessibility of the interventions. Future interventions also need to consider their appropriateness within familial context or culture and provide adequate implementation training. Last, policy makers and experts can also contribute by making design guidelines to prevent digital product developers from making products that can encourage overuse in children.

(JMIR Ment Health 2024;11:e56896) doi:10.2196/56896

KEYWORDS

children; digital device; internet addiction; intervention; prevention; problematic internet use; technology; young children; problematic use; preventive; interventions; systematic review; internet; addiction; prevent; reduce; risk; risks; database; databases; child; PICOS; thematic analysis; Population, Intervention, Comparison, Outcome, and Study type

Introduction

Internet Addiction in Young Children

The internet provides powerful functions and benefits in supporting human lives and work. Globally, people spend more than 6.5 hours daily on online activities, meaning we spend almost a third of our lives using the internet [1]. In this digital age, most children begin using the internet since early childhood (younger than 5 years) [2-5]. The existing guidelines suggest that parents do not give their children access to internet devices before they are 2 years old [6,7]. The increasing access to the internet for young children raises concern about the negative impacts of internet overuse that could lead to addictive behavior. This emerging phenomenon is often referred to as internet addiction [8,9]. Behavioral disorders related to the internet and gaming activities have been recognized as a diagnosable mental health condition that needs further studies in the *International Statistical Classification of Diseases* and the *Diagnostic and Statistical Manual of Mental Disorders* [10,11].

Internet addiction is identified as a behavioral disorder caused by the excessive and uncontrolled use of the internet and digital devices, which can lead to physical, mental, and social problems [12,13]. Internet addiction could bring many negative consequences for children, including mental health problems such as anxiety, emotional instability, and depression [14,15]; physical health problems such as headaches, eye problems, and musculoskeletal pains [16,17]; declining performance [18,19]; sleep disorder [20]; antisocial behavior [21,22]; speech delay [23]; and hindering child growth and development [24]. Prior studies highlighted that children are considered vulnerable to internet addiction [8,9,25]. In 2022, the estimated global prevalence of internet addiction in children was 13.82% [26]. Multiple studies highlighted some reasons that underlie the high risks of internet addiction in young children, such as limited self-control [18], incomplete brain development [20], parental limitations [27,28], and influence from children's environment [29,30].

According to the Interactional Theory of Childhood Problematic Media Use, some distal, proximal, and maintaining factors jointly contribute to determining the risks of internet addiction in children younger than 12 years [31]. For context, the media referred to in the model is digital media that can be accessed and distributed through the internet [31]. The distal factors include the family socioeconomic conditions, family dysfunctions (eg, behavioral, academic, and social dysfunctions), and digital environments (eg, types of devices used, online activities, and content accessed). The proximal factors include the access, behavior, and attitude toward the internet and media use from the children, their family, and their peers. The maintaining factors include parent-child relationships, peers' influence on the internet and media use, and self-efficacy and self-regulation in children. Internet addiction in young children becomes more complex than in adults since the people around them (eg, parents, siblings, or peers) may significantly influence their online behavior [31,32]. In addition, prior studies were concerned about product features that could encourage children to have more screen time, which can exacerbate the problem [33,34].

Related Work and the Objective of This Study

This review study was conducted to fill gaps in the discussion of the current state and future directions of preventive interventions for internet addiction in young children. We identified the gaps from the prior studies. Vondráčková and Gabrhelík [25] and Lee et al [35] conducted review studies about the prevention of internet addiction. They discussed the topic from several aspects, such as conceptual model, target groups, specific skills, characteristics, and environmental (or social). However, they did not discuss the preventive interventions and their outcomes. They highlighted that research into and development of preventive interventions for internet addiction are still scarce, especially for vulnerable populations like children. The finding from Vondráčková and Gabrhelík [25] suggested the need for more intervention studies of children by involving their environment (eg, parents, teachers, and peers). Young children typically have unique characteristics related to their internet use and risky online behavior, thus requiring appropriate intervention developed for them [31,36]. The previous findings indicate the need to conduct more investigation on appropriate preventive interventions for internet addiction in children.

Prior studies reviewed the existing internet addiction treatments for various target groups. Xu et al [37] and Kuss and Lopez-Fernandez [9] discussed psychological treatments and therapies for internet addiction. Ayub et al [38] discussed treatment modalities for addressing internet addiction in children and adolescents. Those studies similarly investigated internet addiction treatments' methods, domains, and effectiveness. The use of psychotherapies like cognitive behavioral therapy and electro-acupuncture were reported as promising treatments for reducing the symptoms of internet addiction [9,37,38]. However, there is a lack of studies discussing the preventive interventions for internet addiction in young children. Preventive interventions are essentially needed for children to prevent them from experiencing the negative consequences of internet addiction [8,25].

The high risks and prevalence of internet addiction in young children raised the urgency in exploring how best to prevent this problem. There is a need to understand how the existing preventive interventions have been developed, implemented, and assessed to prevent or reduce the risks of internet addiction in young children. Young children are categorized as persons 12 years and younger based on the theory of child cognitive development [39,40]. Children older than 12 years are generally classified as adolescents with different characteristics and internet behavior than younger children [31,40-42]. Therefore, this review study aimed to investigate the existing interventions and explore future directions to prevent or reduce internet addiction risks in children younger than 12 years. This review study contributed to filling the gaps in understanding the current approaches, efficacy outcomes, and strengths and limitations of preventive interventions to address internet addiction in young children. In addition, this study provided recommendations on future intervention study opportunities to overcome the limitations of the existing interventions.

Methods

Overview

This review was conducted following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines to identify, report, and synthesize the evidence systematically [43]. We performed the systematic literature review in six key stages as follows: (1) determining the research questions, (2) defining the search strategy and conducting the literature search, (3) selecting the relevant studies, (4) assessing the risk of bias, (5) extracting the data, and (6) analyzing and reporting the data.

Research Questions

We formulated two research questions to achieve the objective of this study as follows: (1) What intervention approaches have been developed, and what are their efficacy outcomes to prevent or reduce internet addiction risks in children younger than 12 years? and (2) What are the strengths and limitations of the existing interventions to prevent internet addiction in children younger than 12 years?

Search Terms and Strategy

This study reviewed primary studies discussing interventions to prevent internet addiction in young children and their efficacy outcomes. In addition, some terms are commonly used to refer to the phenomenon that could lead to internet addiction, such as problematic internet use, compulsive internet use, and excessive internet use [9,12,44-46]. In this study, we also included those related terms in the literature search to obtain comprehensive evidence on the current state of preventive interventions.

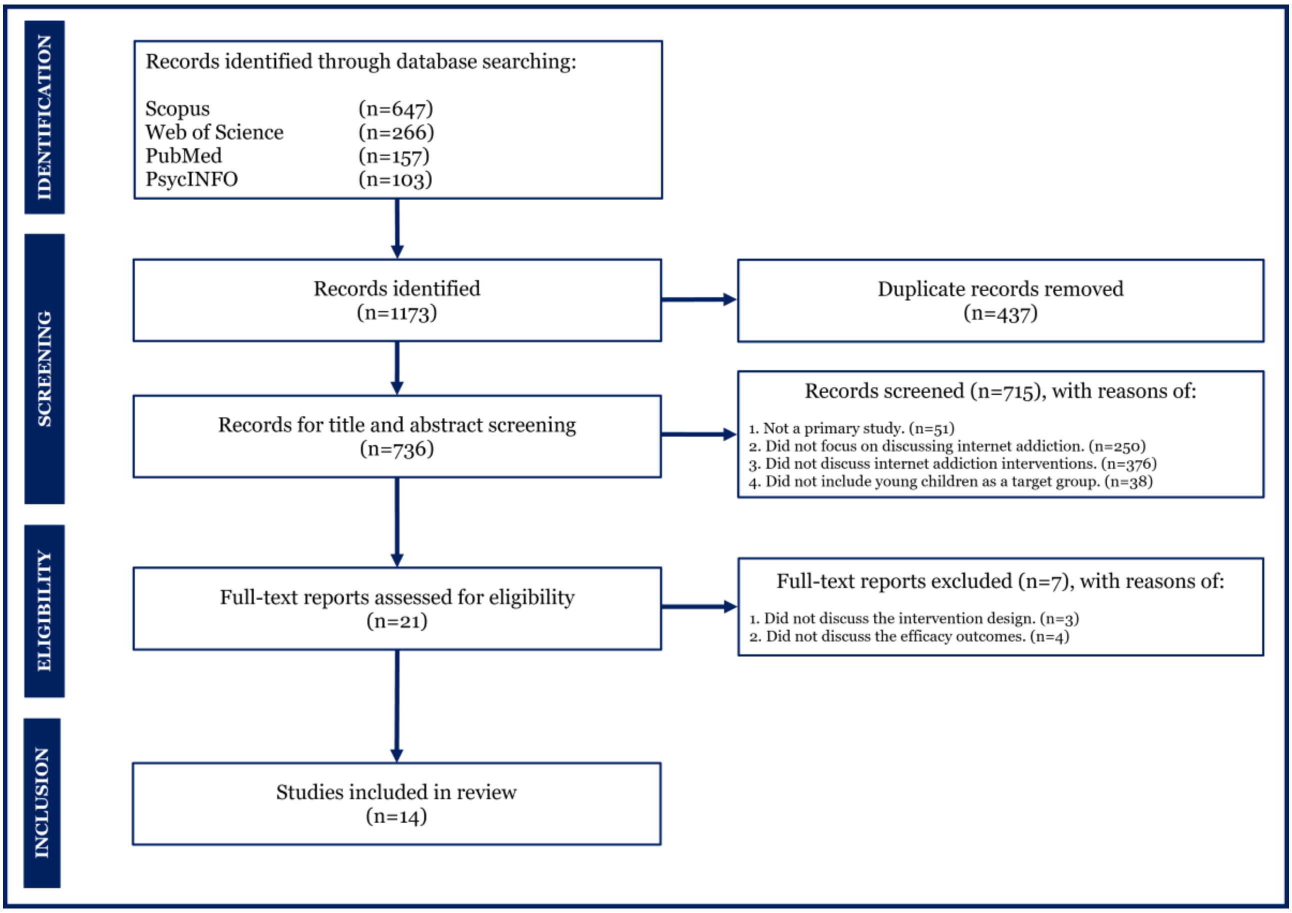
We searched for relevant literature from 4 credible research databases (Scopus, Web of Science, PubMed, and PsycINFO). The literature search was conducted within the title and abstract from the databases with the following search string: internet AND (addiction OR problematic OR compulsive OR excessive) AND (prevent* OR intervention) AND (child*). We searched for peer-reviewed journals or conference articles written in English. The search was conducted on January 5, 2024. The term “internet addiction” was initially introduced in around 1996, and the conceptualization remained relevant to date [47,48]. Therefore, we searched for relevant literature using the publication timeframe between 1996 and January 2024.

Inclusion Criteria for Study Selection

The literature search used inclusion and exclusion criteria to select relevant studies. The inclusion criteria were as follows: (1) the intervention discussed in the study was intended to prevent internet addiction in children younger than 12 years; (2) the study discussed the intervention design and its efficacy outcomes—various types of efficacy assessments were allowed in our evidence search, such as randomized controlled trials [RCTs], quasi-experimental designs (QEDs), or other quantitative study designs; (3) the study was available in a full-text article; (4) the study was peer-reviewed; and (5) the study was written in English.

In this systematic review, we focused on discussing the evidence from primary studies. Therefore, we excluded some types of articles: (1) editorial, (2) review, (3) study protocol, and (4) commentary. Two authors (YT and AAM) performed the study selection, and all the authors checked the results. The screening and selection processes in this study are shown in the PRISMA flow diagram in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.



Risk of Bias Assessment

To assess the risk of bias, we evaluated each study using the Mixed Methods Appraisal Tool (MMAT; version 2018; McGill University) [49]. This tool has been proven valid and reliable for assessing the methodological quality of empirical studies with various study designs [50]. MMAT is suitable in this study since we included primary studies with multiple study designs.

The MMAT consists of 2 screening indicators for all study types and 5 unique quality indicators for each type of study (eg, quantitative RCT, quantitative nonrandomized, quantitative descriptive, qualitative, and mixed methods) [49]. Therefore, each study design has 7 indicators to assess. However, for a mixed methods study, there are 17 indicators to assess (2 screening indicators + 5 mixed methods indicators + 5 quantitative indicators + 5 qualitative indicators). The risk-of-bias assessment was conducted on the studies that had passed the screening and selection process. The assessment was initially performed by 2 authors (YT and AAM), and the final decisions were made based on the authors' consensus.

Data Extraction

The data extraction aimed to summarize the included studies systematically. The PICOS (Population, Intervention, Comparison, Outcome, and Study type) framework was used to systematically report the key evidence of each study [51]. The critical information of each included study was collected, such as authors, year, country, study design (including measurement scale and timeframe), participants, intervention design, control condition, and key outcomes (Multimedia Appendix 1 [52-66]).

Data Analysis

The included studies in this review were thematically analyzed based on their intervention characteristics, approaches, and efficacy outcomes [67]. The mechanisms, strengths, and limitations of each intervention approach were also investigated to identify gaps and directions for future preventive interventions to address the problem. Recommendations for future studies were provided based on the findings of this study. The data extraction and analysis were performed by 2 authors (YT and AAM).

Results

Overview of the Studies

The initial literature search found a total of 1173 articles. After screening and selecting relevant articles (Figure 1), we included 14 studies that met all the inclusion criteria. The summary of included studies based on the PICOS framework is shown in Multimedia Appendix 1.

Although we searched publications using a timeframe between 1996 and 2024, all included studies that suited our criteria were conducted after 2013. The sample size ranged from 10 to 3141. Most studies (13/14, 93%) focused on children aged between 9 and 12 years (or older) as the target group. Only 1 study was intended for children younger than 7 years. The final samples came from Europe (7/14, 50%), Asia (6/14, 42.9%), and the

Middle East (1/14, 7.1%). Based on the country, most of the included studies were conducted in Turkey (n=5) and Hong Kong (n=3). The rest of the studies came from various countries, including South Korea (n=1), Germany (n=1), Lebanon (n=1), Norway (n=1), Taiwan (n=1), and Thailand (n=1).

This review explored the existing preventive interventions that have been assessed to understand the efficacy outcomes of the intervention discussed. Four testing designs were used in the studies, including RCT (n=5), QED (n=5), single-arm trial (n=3), and cross-sectional study (n=1). There was variability in the implementation duration, measurement timeframe, and scale used to measure the efficacy outcomes. The implementation of the interventions discussed in this study varied between 1 week and 6 months. The measurement timeframe was also varied between 1 week and 12 months. Some common internet addiction scales were used, such as Young's Internet Addiction Scale (n=3) [8] and Korean Internet Addiction Proneness Scale (n=2) [68].

The interventions discussed in the included studies were aimed at preventing or reducing the risks of internet addiction (n=9), internet gaming addiction (n=6), and problematic internet use (n=1). Two studies claimed that their interventions were intended to prevent internet addiction and internet gaming addiction [57,58], and 1 study developed an intervention to prevent both internet addiction and smartphone addiction [59]. The included studies involved children's environment (eg, families and schools) in their intervention design, such as teachers (n=7), parents (n=4), peers (n=2), and school nurses (n=1).

The included studies mentioned the theoretical underpinnings that underlie their intervention designs. The interventions used various well-established theories related to parenting (eg, parental mediation [52], positive parenting [55], and parenting styles [58]), psychosocial (eg, self-regulation [60], social cognitive [61], self-determination [62], ecological systems [62], family systems [59], and operant conditioning [58]), and learning (eg, participatory learning [60], media literacy [57], and gamified learning [53]).

Risk of Bias Assessment

The risk of bias assessment was conducted on the 14 studies (Multimedia Appendix 2). All included studies passed the 2 screening indicators for clear research questions and data collection to address the questions. Five RCT studies were assessed based on 5 quality indicators in terms of randomization, baseline comparison, outcome data, outcome assessor, and intervention adherence. Four studies provided adequate explanations of all required quality indicators. One study did not clearly explain one indicator about the comparison of the baseline conditions between intervention and control groups [52].

Nine studies were assessed using 5 quality indicators for quantitative nonrandomized studies (including QED, single-arm trial, and cross-sectional study designs). This includes participants' representativeness, measurement appropriateness, outcome data, confounders' accountability, and intervention or exposure administration. Seven of 9 studies met all the quality indicators. Two studies did not provide adequate rationale about the representativeness of their samples to the target population [53,54]. One study did not clearly describe the potential confounders in their studies [54].

According to the risk of bias assessment results, 11 studies met all quality indicators, 2 did not meet 1 indicator, and 1 did not meet 2 indicators. We decided to include all the studies by taking into account their methodological limitations in extracting and analyzing the data.

Preventive Intervention Approaches and Their Outcomes

Overview

The preventive interventions included in this study were categorized based on the similarity of the working mechanism to achieve the objective. Four different approaches were identified from the existing interventions. These were children's education, parenting strategy, strategic physical activity, and counseling.

Children's Education (n=5)

The interventions in this category aim to prevent or reduce internet addiction risks in young children by enhancing their knowledge or skills. The educational materials provided through this approach can be divided into two main goals as follows: (1) improving children's digital literacy and encouraging healthy internet use (n=4) and (2) developing children's competencies in combating addictive behavior (n=1). The interventions in this category were developed as school materials or curricula, thus involving teachers in delivering the materials to their students in classes.

Three educational interventions developed limited meetings (4-8 sessions, 30-90 minutes each) for children to enhance their knowledge and skills about digital literacy and healthy internet use: School-Based Media Literacy [57], School- and Family-based Intervention [60], and Healthy Internet Use [61]. These programs were delivered as offline seminars, training, or classes to educate children about internet use, risky online behavior, and how to prevent and anticipate internet addiction. Although the goals were similar, they used different theoretical underpinnings to develop the intervention, such as media literacy, social cognition, self-regulation, and participatory learning. Based on the efficacy assessment, these interventions showed promising efficacy in lowering the risks of internet addiction or internet gaming addiction [57,60,61].

One intervention (Wise IT-use) used a hybrid format (combining 3-month online training and an offline workshop) to deliver similar materials [53]. This intervention was developed based on gamification learning and flow theory. It provided multimedia learning and flexible online training to enhance children's experience in learning the materials. This program showed promising efficacy in lowering the risks of internet gaming addiction (χ^2_4=42.89, P<.001; d=.5) [63].

The last intervention in this category (B.E.S.T. Teen) aimed to equip children with youth development competencies to combat addictive behavior, including cognitive, emotional, social, and behavioral competencies [63]. Similar to 3 other interventions in this category, it consisted of limited offline meetings (10 sessions, 30 minutes each) to deliver the materials in school settings. This intervention significantly lowered the odds of addictive behavior (*B*=–0.61, SE 0.19; odds ratio 0.55; *P*<.002).

Parenting Strategy (n=4)

This approach provides parental rules, skills, or guidelines to support parents in regulating children's internet use. Two studies in this category developed learning materials to improve parenting knowledge and skills to prevent risky internet use in children [55,62]. The Positive Parenting Program (Triple P) provided a set of seminars (3 sessions, 2 hours each) to educate parental guidelines on how to cope with online behavior problems and health care services in young children [55]. Developed based on the positive parenting theory, this program showed promising efficacy in reducing children's screen time and improving minor psychiatric disorders and family perception [55]. The Game Over Intervention was developed based on self-determination and ecological systems theories to provide parents with parental monitoring, parental care, and psychoeducation skills [62]. This intervention could reduce children's screen time and addiction risks. However, the control group also showed a similar reduction, so there was not enough evidence that the intervention was better than the control group [62].

Two other studies in this category developed and discussed parenting strategies to reduce children's internet use. The e-Discipline program used screen time as a discipline tool for parents to manage children's behavior and attitude [58]. This program was based on parenting styles and operant conditioning theories. Through this intervention, parents reward and punish their children by adding or reducing their children's screen time. Ironically, this intervention made their children more likely to exceed recommended screen time (2 hours a day) than before [58]. Therefore, this strategy was not promising in preventing risky internet use in children. The other intervention (Guardian Guidelines to Prevent Problematic Gaming) was developed based on the parental mediation theory [52]. This is a set of guidelines for parents or caregivers to manage children's device use at home (eg, children should not use digital devices in their bedrooms, children should not use digital devices more than 5 days a week, and no screen time during meals). This intervention was not promising because it showed no significant difference between the intervention and control group [52]. In addition, many parents involved in this study could not understand how to implement the guidelines properly and were not consistent in implementing the guidelines.

Strategic Physical Activity (n=4)

The strategic physical activity approach encourages children to do more physical activities in order to become less attached to online activities. This approach could also provide many positive developments for children, such as self-regulation, executive functions, and social engagement [56,59,64].

Two interventions used sports activities to reduce the risks of internet addiction in young children. The first intervention (Intensive Sports Activity) was a program based on optimism theory where children were involved in multiple types of intensive sports activities for 12 consecutive weeks [54]. This study reported promising efficacy in reducing addiction risks ($t(185)=20.091$, $P<.001$), improving optimism ($t(185)=–13.205$, $P<.001$), and improving communication skills ($t(185)=–14.903$, $P<.001$) [54]. One similar program (Strategic Physical Activity) was a 12-week strategic basketball activity [56]. The intervention was developed based on the principle that increasing children's executive functions would reduce their addictive behavior. However, although the intervention successfully improved motor competencies ($P=.04$; $r=–0.38$), there was not sufficient evidence that the intervention reduced the risks of addiction.

The other 2 programs in this category focused on preventing addiction by encouraging children's physical activities through play and art activities with peers. Peer Relationship Enhancement and Traditional Children's Game interventions were developed to encourage children to interact with their peers through playing and making art [59,64]. The first intervention used social systems and family systems theories to build the intervention, whereas the second used the psychosocial development theory as a framework. Promising efficacy was reported by the Peer Relationship Enhancement ($P<.05$; $d=.4$) and Traditional Children's Game ($P<.05$; $d=.77$) interventions [59,64].

Counseling (n=1)

Counseling is professional assistance that gives advice or recommendations for coping with particular personal problems [69]. The intervention in this category prevents internet addiction through professional counseling sessions to help children reflect on their internet use, recognize their internet use problems, and find appropriate solutions. One study (Solution-Focused Intervention) was developed based on positive psychology theory to guide children in identifying problems, setting goals, and finding appropriate solutions relating to their internet use [65]. This short-term intervention (3-month implementation) consisted of 6 group interviews every 2 weeks. The intervention showed promising efficacy in lowering internet addiction risks ($P<.01$; $d=.5$) [65].

Strengths and Limitations of the Existing Approaches

The most frequent intervention approach to prevent internet addiction in young children is children's education (n=5). This approach showed great potential since all included studies in this category reported promising efficacy in preventing or reducing internet addiction risks. The main strength of this approach is that it provides children with understanding rather than forces them to engage in particular online behaviors (eg, reducing screen time and avoiding specific apps). In addition, this approach is flexible and can be attached to children's daily activities, such as at school or a child community center. The challenges may appear in providing appropriate materials for the children, training the provider to deliver the materials in exciting ways, and increasing awareness to educate healthy internet use in children [8]. Besides, the family's or parent's roles in educating and modeling healthy internet use could also significantly influence how children can implement the materials [27,28,70]. Educational interventions typically need initial awareness and active commitment from children to be involved. In addition, the educational interventions included in this study are temporary programs (limited seminar, training, or workshop sessions).

Parents had a significant influence in providing internet access and controlling children's online behavior [31,71,72]. However, some parenting strategy interventions included in this study were not efficacious [52,58,62]. Parents can contribute significantly to preventing internet addiction through education, role modeling, or positive relationships with children [8,27,28]. Some limitations of the parenting approach may support the outcomes of the existing interventions: (1) parents had limited capacity, capability, and consistency in implementing the strategy [52,62]; (2) children with better knowledge about technology might outsmart their parents so that they could violate parental rules easily [27,62]; (3) forcing children's online behavior by implementing restrictions and limitations without giving proper understanding might not be effective and favorable [52,58,73]; and (4) harnessing screen time as a tool for rewarding or punishing children's behavior might not work to prevent internet addiction [58,74].

The strategic physical activity approach showed promising efficacy outcomes in preventing addictive behavior in children. The strengths of this approach are its ability to enhance peer relationships and smoothly shift children's attention to physical activities to reduce children's engagement with online activities. In addition, the interventions in this approach could bring additional positive values for children, such as self-regulation, executive functions, and social engagement [56,59,64]. However, all interventions in this approach provided limited physical activity sessions for children; thus, we could not determine the sustainable effects of the interventions. Matching the physical activities with children's interests would also be crucial because children may have various activity preferences (eg, some children may not like sports or arts).

The counseling approach utilized the capability of health practitioners to provide children with proper advice or recommendations to address internet addiction [8]. The intervention included in this review used a solution-focused approach in the counseling sessions to achieve the objective [65]. The use of counseling or psychotherapy approaches was common and promising in reducing the symptoms of internet addiction [37,38]. Several common approaches exist, such as cognitive behavioral therapy and family therapy [9,37]. However, there is a lack of discussion about the application and outcomes of those in preventing internet addiction in young children. The counseling approach needs the commitment of children and their parents or guardians to spend time, money, and energy attending the sessions. In addition, this intervention needs to be delivered by professional health practitioners, who may not always be available or accessible, especially in low-resource or rural areas [75]. Some parents or children may also have a negative stigma towards counseling or therapy because they may be considered different or "abnormal" [76].

Discussion

Principal Findings

This review study has provided an exploration of the current approaches, efficacy outcomes, and strengths and limitations of the existing interventions to prevent or reduce the risks of internet addiction in young children. Ten (71%) out of 14 preventive interventions for young children reported promising efficacy in preventing or reducing the risks of internet addiction. Those interventions showed small to medium effect sizes of their interventions [77]. However, 3 studies with promising efficacy did not provide effect size information in their articles (we have tried to follow up this information with the corresponding author via email).

According to the outcomes, interventions that enhance children's knowledge and skills in having appropriate digital literacy and healthy online behavior were more likely to show promising efficacy than interventions that force children to reduce screen time. Interventions with this objective showed promising efficacy, regardless of the approaches used (eg, children's education, parenting strategy, and counseling) [53,55,57,60,61,63,65]. Another study (Game Over Intervention) with a similar aim also reported a significant risk reduction [62]. However, there was insufficient evidence because a similar reduction was also found in the control groups. In contrast, interventions that forced children to restrict their screen time without proper education and communication were not efficacious [52,58]. It showed that instilling an awareness of healthy online behavior in children had a better effect than enforcing restrictions. Previous studies similarly found that shaping children's behavior would be more effective than forcing them to do certain behaviors [27,78]. Forcing the children too much may also provide a negative experience [74].

Interventions that shift children's focus from online activities to real-world activities also showed promise in reducing children's engagement with the internet, thereby preventing addictive behavior [54,59,64]. Those interventions leveraged children's social relationships with peers through various activities (eg, sports, plays, and arts) to prevent overengagement with online activities. Prior studies also highlighted the importance of improving peer relationships and encouraging more real-world activities in combating internet addiction in children [79-82]. In addition, it may be beneficial if the intervention can suit the physical activities with children's or families' preferences to increase its acceptability.

Prior studies highlighted the vital role of parents in preventing internet addiction in young children [27,28]. However, 3 parenting strategy interventions included in this study did not show promising efficacy [52,58,62]. We identified that the interventions might be ineffective for two main reasons: (1) inappropriate strategies and (2) the parents' failure to implement the intervention as intended. Regardless of the outcomes, involving parents is essential in developing interventions to encourage healthy internet use in children [27,83]. To overcome the first limitation, we suggest further studies to collaborate with related experts and health practitioners to develop appropriate parental guidelines. Technology may also be used to provide tailoring or personalized strategies for parents based on their preferences or conditions [84,85]. To overcome the second limitation, we suggest further studies to provide adequate training and understanding for parents in implementing the strategies. Appropriate strategies would not be useful if parents cannot apply them well. Therefore, we should also consider the motivation, usability, and learnability factors when parents apply the strategies to their children.

According to the Interactional Theory of Childhood Problematic Media Use model, internet addiction in young children could be significantly influenced by factors related to the family and peers (eg, relationships, behavior, attitude, and media influence) [31]. The role of children's environment may improve or exacerbate the risks of internet addiction in young children. Therefore, involving people who can influence children's online behavior may be beneficial for the success of preventive interventions. The existing interventions in the included studies also involved children's families or schools in delivering the intervention. For instance, parents were trained to manage children's internet use, teachers delivered education materials and physical activity programs, peers collaborated to do physical activities together, and health practitioners provided professional counseling. However, no single intervention involved more than 1 stakeholder. Therefore, we recommend that the intervention design and implementation involve stakeholders that can significantly influence children's behavior. They should be used to reinforce positive online behavior in children and prevent negative influence [86]. For example, parents or teachers can be role models or educators, peers or siblings can be social facilitators, and health practitioners can create educational materials to combat addictive behavior. Their combined contributions will create positive environments for children to prevent addictive behavior.

The findings of this study showed that each intervention approach has some limitations in design and implementation that need to be further improved. Although some interventions reported promising efficacy outcomes, most of them were temporary programs with limited sessions and accessibility (eg, seminar or training, professional counseling, and strategic physical activity). Sustainable interventions may be needed to improve long-term effects in young children [87,88]. In addition, interventions that need much money or expert involvement (eg, counseling) may not always be accessible in low-resource regions [89]. Therefore, we suggest developing interventions with better accessibility to reach various families with different backgrounds.

Combining multiple approaches may improve efficacy in overcoming the limitations of each approach [90]. For example, we may develop an integrated intervention that facilitates families in educating healthy internet use, determining appropriate internet use regulations, and suggesting attractive physical activities to prevent their children from over-engagement with online activities. The use of digital technology may be beneficial in achieving these goals [91]. Digital technology can increase the capability, efficiency, and accessibility of the intervention in encouraging children to have healthy online behavior [92,93].

Some digital tools, such as parental control or digital well-being software, have been developed to support managing children's device use [88,92]. Parental control software could be beneficial in improving children's online safety and parental mediation [74,94]. However, in this study, we did not find studies that investigated the design and efficacy outcomes of digital interventions or digital tools to prevent internet addiction in young children. Some studies developed digital tools to manage children's internet use, but they were not developed specifically to prevent internet addiction, and their efficacy outcomes have not been tested [74,88,95]. Therefore, further studies are needed to investigate, develop, and evaluate appropriate digital tools to prevent internet addiction in young children.

The interventions discussed in this study mainly focused on educating or regulating young children as problem owners. However, there were concerns about digital product features or content that could encourage children to have more online activities [22,33,34]. Considering that excessive and uncontrolled online activities can cause internet addiction, we suggest future studies to investigate how product developers for children can contribute to preventing addictive behavior in their users. We encourage product developers, related experts, or policy makers to consider safe child-computer interaction in supporting internet addiction prevention in children. This can be manifested in various forms, such as making child-friendly design guidelines, interaction strategies, or policies.

In this study, we searched for evidence of the existing interventions for preventing internet addiction in children younger than 12 years. However, the interventions identified in this study mainly focused on children aged 9-12 years. There is a lack of intervention studies intended for children younger than 8 years. Since today's children start using the internet from early childhood (1-5 years old) [3-5], we suggest investigating more intervention studies focusing on children younger than 8 years. It is crucial to ensure children do not have addictive behavior in early childhood since optimal cognitive development typically starts from that period [96,97].

Limitations of the Study

This study may have some possible limitations. This systematic review focused on investigating relevant evidence about preventive interventions to address internet addiction in children. To date, the conceptualization between internet addiction and other related terminologies (eg, digital addiction and smartphone addiction) is still under debate due to some similarities in symptoms, mechanisms, and harmful effects [13,98,99]. Consequently, interventions for preventing related problems like digital or smartphone addiction may also have the prospect of preventing internet addiction. However, we did not include other related terms in this study to avoid biases. Future studies are needed to define and standardize this conceptualization issue before considering them as similar constructs.

In this study, we did not limit the regions where the studies were conducted to avoid selection biases. However, the studies included in this review mainly came from European and Asian countries. In this review, we did not get samples from some regions (eg, North America and Australia) due to our inclusion criteria to achieve the objectives of this particular study, which might be a limitation of our study. For instance, we initially found 15 studies with relevant topics (preventing internet addiction in young children) from North America and Australia. However, we excluded them because they did not report their intervention design or efficacy outcomes. Accordingly, this study might have limited generalizations that must be considered when applying the findings. This limitation also indicated the need for future studies to develop, implement, and evaluate new or existing preventive interventions in different regions to extend our knowledge on preventing the problem effectively in multiple contexts.

Although the existence of addiction to the internet and online gaming has been recognized as a diagnosable mental condition (eg, in *Diagnostic and Statistical Manual of Mental Disorders* and *International Statistical Classification of Diseases*), the diagnosis may have a cultural limitation. For example, in the United States, internet addiction was considered a comorbid condition, not a primary diagnosis [100], and approximately 86% of internet addiction cases were comorbid with other conditions [101]. Other studies similarly reported the possible comorbidity of internet addiction with other diagnoses such as attention deficit and hyperactivity disorder and depression [102,103]. We recognize that considering the comorbid conditions of the participants is essential to have a more accurate diagnosis of internet addiction. However, only 1 study [56] in our sample screened the participants' comorbid conditions, which might be a limitation of our study. We suggest further internet addiction studies to pay attention to this comorbidity issue to have better results and validity.

The variability of the measurement scale and timeframe in the included studies raised challenges in comparing the efficacy outcomes of the interventions. Therefore, we could not perform a comparative analysis of the interventions in this study. Nevertheless, we have provided a deep exploration and discussion of the potential preventive intervention mechanisms and approaches as valuable insights to address this problem better in the future.

Conclusions

There is a growing concern about internet addiction in young children due to the increasing number of childhood internet users and their vulnerability to this problem. This review study has investigated and discussed the current approaches, efficacy outcomes, and strengths and limitations of the existing interventions to prevent or reduce the risks of internet addiction in young children. This study identified 14 preventive interventions categorized into 4 groups based on their approaches to achieving the objective. This includes children's education, parenting strategy, strategic physical activity, and counseling. Ten interventions showed promising efficacy outcomes in preventing or reducing internet addiction risks in young children with small-to-medium effect sizes.

Overall, preventive interventions that enhance children's competencies in having appropriate online behavior and literacy were more likely to have better efficacy than interventions that force children to reduce screen time. Interventions that shift children's focus from online activities to real-world activities also showed promise in reducing children's engagement with the internet, thereby preventing addictive behavior. In this study, we have also identified the limitations of each intervention approach as valuable considerations in developing future interventions to address the problem. The current limitations include several domains, such as the temporariness of the program, accessibility, parental capability, and implementation.

The findings of this study suggest the need to develop more sustainable and accessible interventions in educating healthy internet use, determining appropriate internet use regulations for children, and suggesting attractive activities to prevent children from overengagement with online activities. Involving children's stakeholders (eg, parents, teachers, and peers) can be beneficial in reinforcing positive online behavior in children and preventing negative influence. The use of technology-mediated interventions is recommended to improve the capability, efficiency, and accessibility of the intervention. Further studies are needed to investigate, develop, and evaluate appropriate digital tools to prevent internet addiction in young children. In developing parental control interventions, we must consider the appropriateness of the strategies with familial contexts or cultures and provide adequate training or understanding for parents to apply the strategies as intended. Future interventions may also emphasize the role of product developers, related experts, or policy makers by developing child-friendly product design guidelines to prevent developers from making products that can encourage overuse. Last, future studies may be needed to develop preventive interventions for children younger than 8 years. This was lacking in the current literature but urgently needed, given that today's children start interacting with technology at a very young age.

Acknowledgments

We would like to thank the Indonesia Endowment Fund for Education and the Indonesian Center of Higher Education Funding for funding the research project associated with this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of included studies.
[DOCX File , 55 KB - mental_v11i1e56896_app1.docx]

Multimedia Appendix 2
Risk-of-bias assessment summary.
[DOCX File , 39 KB - mental_v11i1e56896_app2.docx]

Multimedia Appendix 3
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 33 KB - mental_v11i1e56896_app3.docx]

References

1. Kemp S. Digital 2023: global overview report. DataReportal. 2023. URL: https://datareportal.com/reports/digital-2023-global-overview-report [accessed 2024-01-03]
2. Milanovic J. Every fourth child in preschool age owns a digital device. UNICEF. 2018. URL: https://www.unicef.org/serbia/en/press-releases/every-fourth-child-preschool-age-owns-digital-device [accessed 2024-01-03]
3. Sahlberg P, Graham A. Children own around 3 digital devices on average, and few can spend a day without them. The Conversation. 2021. URL: https://theconversation.com/children-own-around-3-digital-devices-on-average-and-few-can-spend-a-day-without-them-159546 [accessed 2024-01-03]
4. Chang HY, Park EJ, Yoo HJ, Lee JW, Shin Y. Electronic media exposure and use among toddlers. Psychiatry Investig 2018;15(6):568-573 [FREE Full text] [doi: 10.30773/pi.2017.11.30.2] [Medline: 29788698]
5. Gummer A. Childwise preschool report 2021. Kidscreen. URL: https://kidscreen.com/tag/childwise-preschool-report-2021/ [accessed 2023-11-20]
6. Joshi A, Hinkley T. Too much time on screens? screen time effects and guidelines for children and young people. Australian Institute of Family Studies. 2021. URL: https://aifs.gov.au/resources/short-articles/too-much-time-screens [accessed 2023-11-20]
7. Digital guidelines: promoting healthy technology use for children. American Psychological Association. 2019. URL: https://www.apa.org/topics/social-media-internet/technology-use-children [accessed 2023-11-20]
8. Young KS. In: Abreu CND, editor. Internet Addiction in Children and Adolescents Risk Factors, Assessment, and Treatment. New York: Springer Publishing Company; 2017.
9. Kuss DJ, Lopez-Fernandez O. Internet addiction and problematic internet use: a systematic review of clinical research. World J Psychiatry 2016;6(1):143-176 [FREE Full text] [doi: 10.5498/wjp.v6.i1.143] [Medline: 27014605]
10. International Statistical Classification of Diseases. Geneva: World Health Organization; 2017.
11. DSM-V: Diagnostic and Statistical Manual of Mental Disorders. Virginia: American Psychiatric Association; 2013.
12. Shaw M, Black DW. Internet addiction: definition, assessment, epidemiology and clinical management. CNS Drugs 2008;22(5):353-365. [doi: 10.2165/00023210-200822050-00001] [Medline: 18399706]
13. Basel A, McAlaney J, Skinner T, Pleva M, Ali R. Defining digital addiction: Key features from the literature. Psihologija 2020;53(3):237-253. [doi: 10.2298/psi191029017a]
14. Mo PKH, Chen JH, Lau JTF, Wu AMS. Internet-related addictions: from measurements to interventions. Int J Environ Res Public Health 2020;17(7):2539 [FREE Full text] [doi: 10.3390/ijerph17072539] [Medline: 32276328]
15. Sohn SY, Rees P, Wildridge B, Kalk NJ, Carter B. Prevalence of problematic smartphone usage and associated mental health outcomes amongst children and young people: a systematic review, meta-analysis and GRADE of the evidence. BMC Psychiatry 2019;19(1):356 [FREE Full text] [doi: 10.1186/s12888-019-2350-x] [Medline: 31779637]
16. Rashid SMM, Mawah J, Banik E, Akter Y, Deen JI, Jahan A, et al. Prevalence and impact of the use of electronic gadgets on the health of children in secondary schools in Bangladesh: a cross-sectional study. Health Sci Rep 2021;4(4):e388 [FREE Full text] [doi: 10.1002/hsr2.388] [Medline: 34622022]
17. Smahel D, Wright MF, Cernikova M. The impact of digital media on health: children's perspectives. Int J Public Health 2015;60(2):131-137. [doi: 10.1007/s00038-015-0649-z] [Medline: 25601331]
18. Schulz van Endert T. Addictive use of digital devices in young children: associations with delay discounting, self-control and academic performance. PLoS One 2021;16(6):e0253058 [FREE Full text] [doi: 10.1371/journal.pone.0253058] [Medline: 34157026]
19. Foerster M, Henneke A, Chetty-Mhlanga S, Röösli M. Impact of adolescents' screen time and nocturnal mobile phone-related awakenings on sleep and general health symptoms: a prospective cohort study. Int J Environ Res Public Health 2019;16(3):518 [FREE Full text] [doi: 10.3390/ijerph16030518] [Medline: 30759792]
20. Zhou M, Zhu W, Sun X, Huang L. Internet addiction and child physical and mental health: evidence from panel dataset in China. J Affect Disord 2022;309:52-62. [doi: 10.1016/j.jad.2022.04.115] [Medline: 35469911]
21. Chagas Brandão L, Sanchez ZM, de O Galvão PP, da Silva Melo MH. Mental health and behavioral problems associated with video game playing among Brazilian adolescents. J Addict Dis 2022;40(2):197-207. [doi: 10.1080/10550887.2021.1971941] [Medline: 34515623]
22. Ma HK. Internet addiction and antisocial internet behavior of adolescents. Sci World J 2011;11:2187-2196 [FREE Full text] [doi: 10.1100/2011/308631] [Medline: 22125466]

23. Hermawati D, Rahmadi FA, Sumekar TA, Winarni TI. Early electronic screen exposure and autistic-like symptoms. Intractable Rare Dis Res 2018;7(1):69-71 [FREE Full text] [doi: 10.5582/irdr.2018.01007] [Medline: 29552452]
24. El Asam A, Samara M, Terry P. Problematic internet use and mental health among British children and adolescents. Addict Behav 2019;90:428-436. [doi: 10.1016/j.addbeh.2018.09.007] [Medline: 30579146]
25. Vondráčková P, Gabrhelík R. Prevention of internet addiction: a systematic review. J Behav Addict 2016;5(4):568-579 [FREE Full text] [doi: 10.1556/2006.5.2016.085] [Medline: 27998173]
26. Meng SQ, Cheng JL, Li YY, Yang XQ, Zheng JW, Chang XW, et al. Global prevalence of digital addiction in general population: a systematic review and meta-analysis. Clin Psychol Rev 2022;92(102128):1-12. [doi: 10.1016/j.cpr.2022.102128] [Medline: 35150965]
27. Auxier B, Anderson M, Perrin A, Turner E. Parenting children in the age of screens. Pew Research Center. 2021. URL: https://www.pewresearch.org/internet/2020/07/28/parenting-children-in-the-age-of-screens/ [accessed 2023-04-26]
28. Valcke M, Bonte S, De Wever B, Rots I. Internet parenting styles and the impact on internet use of primary school children. Comput Educ 2010;55(2):454-464. [doi: 10.1016/j.compedu.2010.02.009]
29. Kim HJ, Min JY, Min KB, Lee TJ, Yoo S. Relationship among family environment, self-control, friendship quality, and adolescents' smartphone addiction in South Korea: findings from nationwide data. PLoS One 2018;13(2):e0190896 [FREE Full text] [doi: 10.1371/journal.pone.0190896] [Medline: 29401496]
30. Gao Q, Zheng H, Sun R, Lu S. Parent-adolescent relationships, peer relationships, and adolescent mobile phone addiction: the mediating role of psychological needs satisfaction. Addict Behav 2022;129:107260. [doi: 10.1016/j.addbeh.2022.107260] [Medline: 35151093]
31. Domoff SE, Borgen AL, Radesky JS. Interactional theory of childhood problematic media use. Hum Behav Emerg Technol 2020;2(4):343-353 [FREE Full text] [doi: 10.1002/hbe2.217] [Medline: 36381426]
32. Soh PCH, Chew KW, Koay KY, Ang PH. Parents vs peers' influence on teenagers' internet addiction and risky online activities. Telemat Inform 2018;35(1):225-236. [doi: 10.1016/j.tele.2017.11.003]
33. Petrescu M, Krishen AS. The dilemma of social media algorithms and analytics. J Market Anal 2020;8:187-188. [doi: 10.1057/s41270-020-00094-4]
34. Yang Q, Gong X. The engagement–addiction dilemma: an empirical evaluation of mobile user interface and mobile game affordance. Internet Res 2021;31(5):1745-1768. [doi: 10.1108/intr-11-2020-0622]
35. Lee SY, Kim MS, Lee HK. Prevention strategies and interventions for internet use disorders due to addictive behaviors based on an integrative conceptual model. Curr Addict Rep 2019;6(3):303-312. [doi: 10.1007/s40429-019-00265-z]
36. Ferrara P, Corsello G, Ianniello F, Sbordone A, Ehrich J, Giardino I, et al. Internet addiction: starting the debate on health and well-being of children overexposed to digital media. J Pediatr 2017;191:280-282. [doi: 10.1016/j.jpeds.2017.09.054] [Medline: 29637892]
37. Xu LX, Wu LL, Geng XM, Wang ZL, Guo XY, Song KR, et al. A review of psychological interventions for internet addiction. Psychiatry Res 2021;302:114016. [doi: 10.1016/j.psychres.2021.114016] [Medline: 34087672]
38. Ayub S, Jain L, Parnia S, Bachu A, Farhan R, Kumar H, et al. Treatment modalities for internet addiction in children and adolescents: a systematic review of randomized controlled trials (RCTs). J Clin Med 2023;12(9):3345 [FREE Full text] [doi: 10.3390/jcm12093345] [Medline: 37176784]
39. Child development center of disease control and prevention. Center of Disease Control and Prevention. 2021. URL: https://www.cdc.gov/ncbddd/childdevelopment/positiveparenting/adolescence.html [accessed 2023-11-20]
40. American Academy of Pediatrics. 2022. URL: https://www.healthychildren.org/English/ages-stages/Pages/default.aspx [accessed 2023-11-20]
41. Bonnie RJ, Backes EP. The Promise of Adolescence: Realizing Opportunity for All Youth. Washington DC: National Academies Press; 2019.
42. Cabello-Hutt T, Cabello P, Claro M. Online opportunities and risks for children and adolescents: the role of digital skills, age, gender and parental mediation in Brazil. New Media Soc 2018;20(7):2411-2431. [doi: 10.1177/1461444817724168]
43. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
44. Yellowlees PM, Marks S. Problematic internet use or internet addiction? Comput Hum Behav 2007;23(3):1447-1453. [doi: 10.1016/j.chb.2005.05.004]
45. Weinstein A, Lejoyeux M. Internet addiction or excessive internet use. Am J Drug Alcohol Abuse 2010;36(5):277-283. [doi: 10.3109/00952990.2010.491880] [Medline: 20545603]
46. McIntyre E, Wiener KKK, Saliba AJ. Compulsive internet use and relations between social connectedness, and introversion. Comput Hum Behav 2015;48:569-574. [doi: 10.1016/j.chb.2015.02.021]
47. Young KS. Psychology of computer use: XL. addictive use of the internet: a case that breaks the stereotype. Psychol Rep 1996;79(3 Pt 1):899-902. [doi: 10.2466/pr0.1996.79.3.899] [Medline: 8969098]
48. YOUNG KS. Internet addiction: the emergence of a new clinical disorder. Cyberpsychol Behav Soc Netw 1998;1(3):237-244. [doi: 10.1089/cpb.1998.1.237]
49. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The mixed methods appraisal tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018;34(4):285-291. [doi: 10.3233/EFI-180221]

50. Pace R, Pluye P, Bartlett G, Macaulay AC, Salsberg J, Jagosh J, et al. Testing the reliability and efficiency of the pilot mixed methods appraisal tool (MMAT) for systematic mixed studies review. Int J Nurs Stud 2012;49(1):47-53. [doi: 10.1016/j.ijnurstu.2011.07.002] [Medline: 21835406]
51. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med 2009;6(7):e1000100 [FREE Full text] [doi: 10.1371/journal.pmed.1000100] [Medline: 19621070]
52. Krossbakken E, Torsheim T, Mentzoni RA, King DL, Bjorvatn B, Lorvik IM, et al. The effectiveness of a parental guide for prevention of problematic video gaming in children: a public health randomized controlled intervention study. J Behav Addict 2018;7(1):52-61 [FREE Full text] [doi: 10.1556/2006.6.2017.087] [Medline: 29313731]
53. Chau CL, Tsui YYY, Cheng C. Gamification for internet gaming disorder prevention: evaluation of a wise IT-Use (WIT) program for Hong Kong primary students. Front Psychol 2019;10:2468 [FREE Full text] [doi: 10.3389/fpsyg.2019.02468] [Medline: 31736842]
54. Sari S, Karagun E. The effect of sports on children's internet addiction, optimism and communication skills. Int J Appl Exerc Physiol 2020;9(11):157-166 [FREE Full text]
55. Özyurt G, Dinsever C, Çalişkan Z, Evgin D. Effects of triple p on digital technological device use in preschool children. J Child Fam Stud 2017;27(1):280-289. [doi: 10.1007/s10826-017-0882-6]
56. Tseng Y, Chao HH, Hung CL. Effect of a strategic physical activity program on cognitive flexibility among children with internet addiction: a pilot study. Children (Basel) 2022;9(6):798 [FREE Full text] [doi: 10.3390/children9060798] [Medline: 35740735]
57. Walther B, Hanewinkel R, Morgenstern M. Effects of a brief school-based media literacy intervention on digital media use in adolescents: cluster randomized controlled trial. Cyberpsychol Behav Soc Netw 2014;17(9):616-623. [doi: 10.1089/cyber.2014.0173] [Medline: 25126888]
58. Hawi NS, Rupert MS. Impact of e-Discipline on children's screen time. Cyberpsychol Behav Soc Netw 2015;18(6):337-342 [FREE Full text] [doi: 10.1089/cyber.2014.0608] [Medline: 26075921]
59. Jo J, Bang KS. The effect of peer relationship enhancement programs on the prevention of smartphone addiction among late school-age children in South Korea. J Pediatr Nurs 2022;63:e127-e135. [doi: 10.1016/j.pedn.2021.09.025] [Medline: 34625302]
60. Apisitwasana N, Perngparn U, Cottler LB. Effectiveness of school- and family-based interventions to prevent gaming addiction among grades 4-5 students in Bangkok, Thailand. Psychol Res Behav Manag 2018;11:103-115 [FREE Full text] [doi: 10.2147/PRBM.S145868] [Medline: 29695939]
61. Uysal G, Balci S. Evaluation of a school-based program for internet addiction of adolescents in Turkey. J Addict Nurs 2018;29(1):43-49. [doi: 10.1097/JAN.0000000000000211] [Medline: 29505460]
62. Li AYL, Chau CL, Cheng C. Development and validation of a parent-based program for preventing gaming disorder: the game over intervention. Int J Environ Res Public Health 2019;16(11):1984 [FREE Full text] [doi: 10.3390/ijerph16111984] [Medline: 31167457]
63. Shek DTL, Yu L, Leung H, Wu FKY, Law MYM. Development, implementation, and evaluation of a multi-addiction prevention program for primary school students in Hong Kong: the B.E.S.T. teen program. Asian J Gambl Issues Public Health 2016;6(1):5 [FREE Full text] [doi: 10.1186/s40405-016-0014-z] [Medline: 27630812]
64. Kacar D, Ayaz-Alkaya S. The effect of traditional children's games on internet addiction, social skills and stress level. Arch Psychiatr Nurs 2022;40:50-55. [doi: 10.1016/j.apnu.2022.04.007] [Medline: 36064245]
65. Akgül-Gündoğdu N, Selçuk-Tosun A. Effect of solution-focused approach on problematic internet use, health behaviors in schoolchildren. J Pediatr Nurs 2023;68:e43-e49. [doi: 10.1016/j.pedn.2022.11.002] [Medline: 36446694]
66. Janssen I, Leblanc AG. Systematic review of the health benefits of physical activity and fitness in school-aged children and youth. Int J Behav Nutr Phys Act 2010;7:40 [FREE Full text] [doi: 10.1186/1479-5868-7-40] [Medline: 20459784]
67. Vaismoradi M, Turunen H, Bondas T. Content analysis and thematic analysis: implications for conducting a qualitative descriptive study. Nurs Health Sci 2013;15(3):398-405. [doi: 10.1111/nhs.12048] [Medline: 23480423]
68. Kim D, Lee Y, Lee J, Nam JK, Chung Y. Development of Korean smartphone addiction proneness scale for youth. PLoS One 2014;9(5):e97920 [FREE Full text] [doi: 10.1371/journal.pone.0097920] [Medline: 24848006]
69. APA dictionary of psychology. American Psychological Association. 2024. URL: https://dictionary.apa.org/counseling [accessed 2023-01-03]
70. Sun Y, Wilkinson JS. Parenting style, personality traits, and interpersonal relationships: a model of prediction of Internet addiction. Int J Commun 2020;14:2163-2185 [FREE Full text]
71. Leijten P, Gardner F, Melendez-Torres GJ, Knerr W, Overbeek G. Parenting behaviors that shape child compliance: a multilevel meta-analysis. PLoS One 2018;13(10):e0204929 [FREE Full text] [doi: 10.1371/journal.pone.0204929] [Medline: 30289928]
72. Brito R, Dias P. “Which apps are good for my children?”: how the parents of young children select apps. Int J Child Comput Interact 2020;26:100188. [doi: 10.1016/j.ijcci.2020.100188]
73. Lanette S, Chua PK, Hayes G, Mazmanian M. How much is 'Too Much'? the role of a smartphone addiction narrative in individuals' experience of use. : USA: Association for Computing Machinery; 2018 Presented at: Proceedings of the 21st

ACM Conference on Computer-Supported Cooperative Work and Social Computing; 01 November 2018; New York p. 1-22. [doi: 10.1145/3274370]
74. Wang G, Zhao J, van Kleek M, Shadbolt N. Protection or punishment? relating the design space of parental control apps and perceptions about them to support parenting for online safety. 2021 Presented at: Proceedings of the 23rd ACM Conference on Computer-Supported Cooperative Work and Social Computing; 18 October 2021; New York p. 1-26. [doi: 10.1145/3476084]
75. Rassool GH. Counselling and addiction. In: Islamic Counselling: An Introduction to Theory and Practice. United Kingdom: Routledge; 2016:237-254.
76. Deres AT, Bürkner PC, Klauke B, Buhlmann U. The role of stigma during the course of inpatient psychotherapeutic treatment in a German sample. Clin Psychol Psychother 2020;27(2):239-248. [doi: 10.1002/cpp.2423] [Medline: 31910308]
77. Maher JM, Markey JC, Ebert-May D. The other half of the story: effect size analysis in quantitative research. CBE Life Sci Educ 2013;12(3):345-351 [FREE Full text] [doi: 10.1187/cbe.13-04-0082] [Medline: 24006382]
78. Canadian Pediatric Society. Effective discipline for children. Paediatr Child Health 2004;9(1):37-50 [FREE Full text] [doi: 10.1093/pch/9.1.37] [Medline: 19654979]
79. Zhou P, Zhang C, Liu J, Wang Z. The relationship between resilience and internet addiction: a multiple mediation model through peer relationship and depression. Cyberpsychol Behav Soc Netw 2017;20(10):634-639. [doi: 10.1089/cyber.2017.0319] [Medline: 29039703]
80. Tsai SM, Wang YY, Weng CM. A study on digital games internet addiction, peer relationships and learning attitude of senior grade of children in elementary school of chiayi county. JEL 2020;9(3):13. [doi: 10.5539/jel.v9n3p13]
81. Nayak SG, Pai MS. Smartphone de-addiction among children and adolescents: an hour of need. J Child Adolesc Psychiatr Nurs 2022;35(2):196-198. [doi: 10.1111/jcap.12369] [Medline: 35133675]
82. Kopczewski M, Napieralska A. Cyber space – risks to children and young people - research results. SJMULF 2018;189(4):46-59. [doi: 10.5604/01.3001.0013.0720]
83. Keya FD, Rahman MM, Nur MT, Pasa MK. Parenting and child's (five years to eighteen years) digital game addiction: a qualitative study in North-Western part of Bangladesh. Comput Hum Behav Rep 2020;2:100031. [doi: 10.1016/j.chbr.2020.100031]
84. Hennessy EA, Johnson BT, Acabchuk RL, McCloskey K, Stewart-James J. Self-regulation mechanisms in health behavior change: a systematic meta-review of meta-analyses, 2006-2017. Health Psychol Rev 2020;14(1):6-42 [FREE Full text] [doi: 10.1080/17437199.2019.1679654] [Medline: 31662031]
85. Venuleo C, Rollo S, Marinaci T, Calogiuri S. Towards a cultural understanding of addictive behaviours. the image of the social environment among problem gamblers, drinkers, internet users and smokers. Addict Res Theory 2016;24(4):274-287. [doi: 10.3109/16066359.2015.1126257]
86. Yang SY. Effects of self-efficacy and self-control on internet addiction in middle school students: a social cognitive theory-driven focus on the mediating influence of social support. Child Health Nurs Res 2020;26(3):357-365 [FREE Full text] [doi: 10.4094/chnr.2020.26.3.357] [Medline: 35004479]
87. Lo CKM, Chan KL, Yu L, Chui WWH, Ip P. Long-term effects of psychosocial interventions on internet-related disorders: a meta-analysis. Comput Hum Behav 2023;138:107465. [doi: 10.1016/j.chb.2022.107465]
88. Alrusaini O, Beyari H. The sustainable effect of artificial intelligence and parental control on children's behavior while using smart devices' apps: the case of Saudi Arabia. Sustainability 2022;14(15):9388. [doi: 10.3390/su14159388]
89. Love SM, Sanders MR, Metzler CW, Prinz RJ, Kast EZ. Enhancing accessibility and engagement in evidence-based parenting programs to reduce maltreatment: conversations with vulnerable parents. J Public Child Welf 2013;7(1):20-38 [FREE Full text] [doi: 10.1080/15548732.2012.701837] [Medline: 23710156]
90. Cemiloglu D, Almourad MB, McAlaney J, Ali R. Combatting digital addiction: current approaches and future directions. Technol Soc 2022;68(101832):1-12. [doi: 10.1016/j.techsoc.2021.101832]
91. Theopilus Y, Al Mahmud A, Davis H, Octavia JR. Digital interventions for combating internet addiction in young children: qualitative study of parent and therapist perspectives. JMIR Pediatr Parent 2024;7:e55364 [FREE Full text] [doi: 10.2196/55364] [Medline: 38669672]
92. Almourad MB, Alrobai A, Skinner T, Hussain M, Ali R. Digital wellbeing tools through users lens. Technol Soc 2021;67:101778. [doi: 10.1016/j.techsoc.2021.101778]
93. Liverpool S, Mota CP, Sales CMD, Čuš A, Carletto S, Hancheva C, et al. Engaging children and young people in digital mental health interventions: systematic review of modes of delivery, facilitators, and barriers. J Med Internet Res 2020;22(6):e16317 [FREE Full text] [doi: 10.2196/16317] [Medline: 32442160]
94. Stoilova M, Bulger M, Livingstone S. Do parental control tools fulfil family expectations for child protection? a rapid evidence review of the contexts and outcomes of use. J Child Media 2023;18(1):29-49. [doi: 10.1080/17482798.2023.2265512]
95. Sadeghi S, Pouretemad HR, Shalani B. Internet-based versus face-to-face intervention training for parents of young children with excessive screen-time and autism spectrum disorder-like symptoms: a comparative study. Int J Dev Disabil 2022;68(5):744-755 [FREE Full text] [doi: 10.1080/20473869.2021.1895699] [Medline: 36210895]

96. Piaget J. Cognitive development in children: piaget development and learning. J Res Sci Teach 1964;2(3):176-186. [doi: 10.1002/tea.3660020306]
97. Dhuey E, Figlio D, Karbownik K, Roth J. School starting age and cognitive development. J Policy Anal Manage 2019;38(3):538-578. [doi: 10.1002/pam.22135]
98. Jin Jeong Y, Suh B, Gweon G. Is smartphone addiction different from internet addiction? comparison of addiction-risk factors among adolescents. Behav Inf Technol 2020;39(5):578-593. [doi: 10.1080/0144929x.2019.1604805]
99. Christakis DA. The challenges of defining and studying "Digital Addiction" in children. JAMA 2019;321(23):2277-2278. [doi: 10.1001/jama.2019.4690] [Medline: 31095260]
100. Block JJ. Pathological computer use in the USA. In: Proceedings of the International Symposium on the Counseling and Treatment of Youth Internet Addiction. Seoul, Korea: National Youth Commission; 2007.
101. Block JJ. Issues for DSM-V: internet addiction. Am J Psychiatry 2008;165(3):306-307. [doi: 10.1176/appi.ajp.2007.07101556] [Medline: 18316427]
102. Yen JY, Ko CH, Yen CF, Wu HY, Yang MJ. The comorbid psychiatric symptoms of internet addiction: attention deficit and hyperactivity disorder (ADHD), depression, social phobia, and hostility. J Adolesc Health 2007;41(1):93-98. [doi: 10.1016/j.jadohealth.2007.02.002] [Medline: 17577539]
103. González-Bueso V, Santamaría JJ, Fernández D, Merino L, Montero E, Ribas J. Association between internet gaming disorder or pathological video-game use and comorbid psychopathology: a comprehensive review. Int J Environ Res Public Health 2018;15(4):668 [FREE Full text] [doi: 10.3390/ijerph15040668] [Medline: 29614059]

Abbreviations

MMAT: Mixed Methods Appraisal Tool
PICOS: Population, Intervention, Comparison, Outcome, and Study type
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QED: quasi-experimental design
RCT: randomized controlled trial

Edited by J Torous; submitted 06.02.24; peer-reviewed by M Ellur , S Karlapati, Y Chen; comments to author 23.06.24; revised version received 25.06.24; accepted 12.07.24; published 30.08.24.

Please cite as:
Theopilus Y, Al Mahmud A, Davis H, Octavia JR
Preventive Interventions for Internet Addiction in Young Children: Systematic Review
JMIR Ment Health 2024;11:e56896
URL: https://mental.jmir.org/2024/1/e56896
doi:10.2196/56896
PMID:

Review

Smartphone-Delivered Attentional Bias Modification Training for Mental Health: Systematic Review and Meta-Analysis

Bilikis Banire[1,2], PhD; Matt Orr[3], PhD; Hailey Burns[4], MSc; Youna McGowan[1], BSc; Rita Orji[2], Prof Dr; Sandra Meier[1], Prof Dr

[1]Department of Psychiatry, Faculty of Medicine, Dalhousie University, Halifax, NS, Canada

[2]Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada

[3]Department of Psychology, Faculty of Pure and Applied Sciences, Acadia University, Wolfville, NS, Canada

[4]Faculty of Medicine, Dalhousie University, Saint John, NB, Canada

Corresponding Author:
Bilikis Banire, PhD
Department of Psychiatry, Faculty of Medicine
Dalhousie University
Sir Charles Tupper Building, 13th Fl.
5850 College Street
Halifax, NS, B34 3H7
Canada
Phone: 1 17828825242
Email: banire.bilikis.o@gmail.com

Abstract

Background: Smartphone-delivered attentional bias modification training (ABMT) intervention has gained popularity as a remote solution for alleviating symptoms of mental health problems. However, the existing literature presents mixed results indicating both significant and insignificant effects of smartphone-delivered interventions.

Objective: This systematic review and meta-analysis aims to assess the impact of smartphone-delivered ABMT on attentional bias and symptoms of mental health problems. Specifically, we examined different design approaches and methods of administration, focusing on common mental health issues, such as anxiety and depression, and design elements, including gamification and stimulus types.

Methods: Our search spanned from 2014 to 2023 and encompassed 4 major databases: MEDLINE, PsycINFO, PubMed, and Scopus. Study selection, data extraction, and critical appraisal were performed independently by 3 authors using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. When necessary, we pooled the standardized mean difference with a 95% CI. In addition, we conducted sensitivity, subgroup, and meta-regression analyses to explore moderator variables of active and placebo ABMT interventions on reducing symptoms of mental health problems and attentional bias.

Results: Our review included 12 papers, involving a total of 24,503 participants, and we were able to conduct a meta-analysis on 20 different study samples from 11 papers. Active ABMT exhibited an effect size (Hedges *g*) of –0.18 (*P*=.03) in reducing symptoms of mental health problems, while the overall effect remained significant. Similarly, placebo ABMT showed an effect size of –0.38 (*P*=.008) in reducing symptoms of mental health problems. In addition, active ABMT (Hedges *g* –0.17; *P*=.004) had significant effects on reducing attentional bias, while placebo ABMT did not significantly alter attentional bias (Hedges *g* –0.04; *P*=.66).

Conclusions: Our understanding of smartphone-delivered ABMT's potential highlights the value of both active and placebo interventions in mental health care. The insights from the moderator analysis also showed that tailoring smartphone-delivered ABMT interventions to specific threat stimuli and considering exposure duration are crucial for optimizing their efficacy. This research underscores the need for personalized approaches in ABMT to effectively reduce attentional bias and symptoms of mental health problems.

Trial Registration: PROSPERO CRD42023460749; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=460749

(JMIR Ment Health 2024;11:e56326) doi:10.2196/56326

KEYWORDS
attentional bias; mental health problem; anxiety; depression; systematic review; meta-analysis; smartphone; mobile phone

Introduction

Background

Smartphone-delivered attentional bias modification training (ABMT) has emerged as a promising intervention for alleviating symptoms of mental health conditions amid a notable increase in their prevalence [1]. As mental health problems, such as anxiety, depression, and substance use disorders, continue to rise globally, traditional treatment options face challenges of accessibility and scalability [1,2]. In response to this growing concern, researchers are exploring innovative approaches such as ABMT, leveraging the ubiquity of smartphones to provide convenient and flexible support for individuals experiencing psychological distress. This systematic review and meta-analysis aimed to evaluate the efficacy of smartphone-delivered ABMT in addressing attentional biases and symptoms of mental health problems, with a particular focus on exploring the impact of different design approaches and methods of administration.

Statistical records have supported a substantial rise, with the number of people with mental health problems increasing from 80.8 million to 125.3 million between 1990 and 2019 [3]. This upward trend has prompted a growing inclination among individuals to seek in-person treatment options for addressing their mental health problems. However, face-to-face therapy also presents societal challenges, including heightened demands on health care systems, a pressing need for additional mental health professionals, and the potential for disparities in access to care [4-6]. Furthermore, various forms of stigma emanate from diverse sources, including families and friends [7]. People's increasing recognition of the significance of addressing mental health problems explains the urgent requirement for comprehensive and accessible mental health services to effectively tackle the broader societal implications of these conditions, alongside the need to protect individuals' privacy when seeking mental health assistance.

In addressing the rise in mental health problems, researchers have come up with evidence-based treatments such as pharmacotherapy and psychological interventions that involve medications and behavior modification, respectively [8]. For example, cognitive behavioral therapy (CBT) focuses on modifying behaviors and maladaptive thoughts through language and communication to address dysfunctional cognitions, fostering behavioral change and proving highly effective and versatile across various mental health conditions. Despite the effectiveness of CBT in addressing mental health problems among young individuals, approximately 40% do not exhibit a positive response to this intervention [9,10]. One of the key possible factors of ineffective CBT is the limitation of the youth's language and communication skills [8]. Therefore, there is continued interest in developing novel interventions.

As options for mental health treatment continue to develop, traditional modalities, such as cognitive restructuring and behavioral activation [11], along with newer approaches, such as third-wave acceptance and mindfulness [12] and ABMT, have been widely used. Prioritizing attentional bias is crucial because it is an automatic process [13]. As such, ABMT can not only be effective in itself but also enhance the effectiveness of other therapeutic interventions and provide a targeted, evidence-based strategy for improving mental health outcomes [14-16].

ABMT stands out as a promising alternative, targeting cognitive processes using visual cues, such as directing attention away from threat-related or addiction-related stimuli. This stimuli-design approach allows ABMT to be more accessible and effective for individuals with limited language and communication skills, overcoming challenges posed by linguistic barriers in the CBT interventions [14,17,18]. Unlike traditional therapies such as CBT, which often involve interpreting complex sentences and verbal interactions, ABMT uses visual and cognitive tasks. For instance, patients respond to visual stimuli rather than needing to interpret text or verbal instructions. This approach reduces the cognitive load and makes it easier for patients to engage effectively in therapy sessions, regardless of their language proficiency or communication abilities. Research indicates that modifying attentional biases through ABMT can have long-lasting effects on emotional regulation and anxiety reduction [19]. ABMT can also be a fully automated, computer-based intervention designed to modify attentional preferences, making it highly scalable and easily accessible for clinical use [16]. In addition, ABMT does not require language communication, which can be particularly advantageous in treating patients who have language barriers or communication impairments.

Recent years have borne witness to a growing interest in ABMT as an empirically supported treatment strategy for an array of mental health problems, including anxiety, posttraumatic stress disorder (PTSD), depression, and substance use. ABMT revolves around the fundamental tenet of training attention away from threat-related stimuli for anxiety, depression, and PTSD and from addictive-cue stimuli for substance use, thereby fostering an internal competition between stimuli that evoke threats or cravings, respectively, and those that are neutral. This internal contest induces a recalibration of attentional mechanisms, leading to a diminished bias toward threat stimuli. The common application of ABMT, grounded in phenomenological characteristics, involves 4 primary experimental tasks: Posner task, Stroop task, dot-probe task, and visual search task [20].

In the context of psychological research, ABMT involves 2 key paradigms: active and placebo ABMT. Active ABMT strategically redirects attention by consistently guiding individuals to focus on neutral stimuli, thereby modifying attentional biases and reducing symptoms associated with anxiety and other mental health issues. In contrast, placebo ABMT serves as a control condition, maintaining the same task structure as active ABMT but placing the cue on both neutral and negative stimuli. This distinction allows researchers to assess the specific therapeutic effects of actively pacing cues to neutral stimuli in ABMT interventions while controlling for nonspecific factors, such as task engagement or participant expectations. These 2 paradigms are pivotal in evaluating the effectiveness of computer-based ABMT and understanding its potential clinical applications [21,22].

The medium through which ABMT is administered has experienced a transformative evolution, aligning itself with the digital tapestry of contemporary health care. While traditionally executed through computer-based platforms, ABMT has recently embarked on a trajectory toward smartphone-mediated delivery [23,24]. This paradigm shift holds great promise, poised to address several pivotal challenges associated with the in-person mode of delivery. The use of smartphones as a conduit for ABMT promises to revolutionize the accessibility and privacy of mental health interventions. The ubiquity of smartphones transcends geographical constraints, rendering mental health support accessible to individuals across diverse locations. In addition, smartphone-based delivery holds the potential to attenuate the omnipresent specter of stigma, an entrenched barrier that has historically dissuaded individuals from engaging with traditional, in-person therapeutic interventions. The discrete and private nature of smartphone-delivered ABMT may sidestep potential stigma, potentially fostering a more expansive adoption of mental health interventions [25]. Furthermore, the incorporation of gamification within smartphone apps enhances user engagement, potentially bolstering treatment adherence and overall efficacy [26]. The gamified interface capitalizes on users' inherent motivation to participate, cultivates sustained engagement, and optimizes treatment outcomes.

Given the advantages of enhanced accessibility, reduced stigma, increased engagement through gamification, extensive customization, real-time feedback, and seamless integration into daily routines, this review focused on ABMT delivered through smartphones and not ABMT delivered through computers. Recent advancements in psychiatry have seen significant contributions from smartphone-delivered interventions for mental health problems. The previous meta-analyses used a narrow lens for the evaluation of smartphone-delivered ABMT, focusing on specific conditions [27-29]. They found that such interventions were effective in addressing mental health problems, improving quality of life, and reducing symptoms of depression and anxiety. The previous studies focused on smartphone-delivered ABMT on reducing the symptoms of specific mental health problems; however, understanding how ABMT operates across a spectrum of mental health symptoms can help evaluate its effectiveness more comprehensively in reducing symptoms of mental health problems. This review not only assessed the impact of ABMT on a wide range of symptoms of mental health problems but also delved into the mechanisms by which ABMT reduces attentional bias. This investigation included both the active and placebo forms of ABMT. By adopting this comprehensive approach, our review provides insights into how smartphone-delivered ABMT affects not only various mental health symptoms but also attentional bias, offering a more holistic perspective on its effects.

Objectives

This systematic review and meta-analysis aimed to analyze the efficacy of smartphone-delivered ABMT in reducing mental health symptoms and attentional biases and to understand how different ABMT design strategies influence these outcomes. The mechanisms by which ABMT operates are directly related to these objectives, as ABMT works by retraining the brain to reduce automatic attention to negative stimuli, thus alleviating symptoms of anxiety, depression, and other mental health issues. By modifying attentional patterns through repetitive training tasks, ABMT aims to improve emotional regulation. Different design strategies, such as the type of stimuli, delivery method, and training frequency, may impact the effectiveness of this retraining, making the exploration of these mechanisms essential for optimizing ABMT interventions and achieving better mental health outcomes.

Methods

Overview

Our systematic literature review adhered to the Cochrane recommendations [30] and followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines for the planning, execution, and reporting of this study (Multimedia Appendix 1) [31]. In addition, our review protocol was registered with PROSPERO (CRD42023460749).

Ethical Considerations

In this review, there was no need for informed consent or ethics approval because the data were extracted from previously published studies.

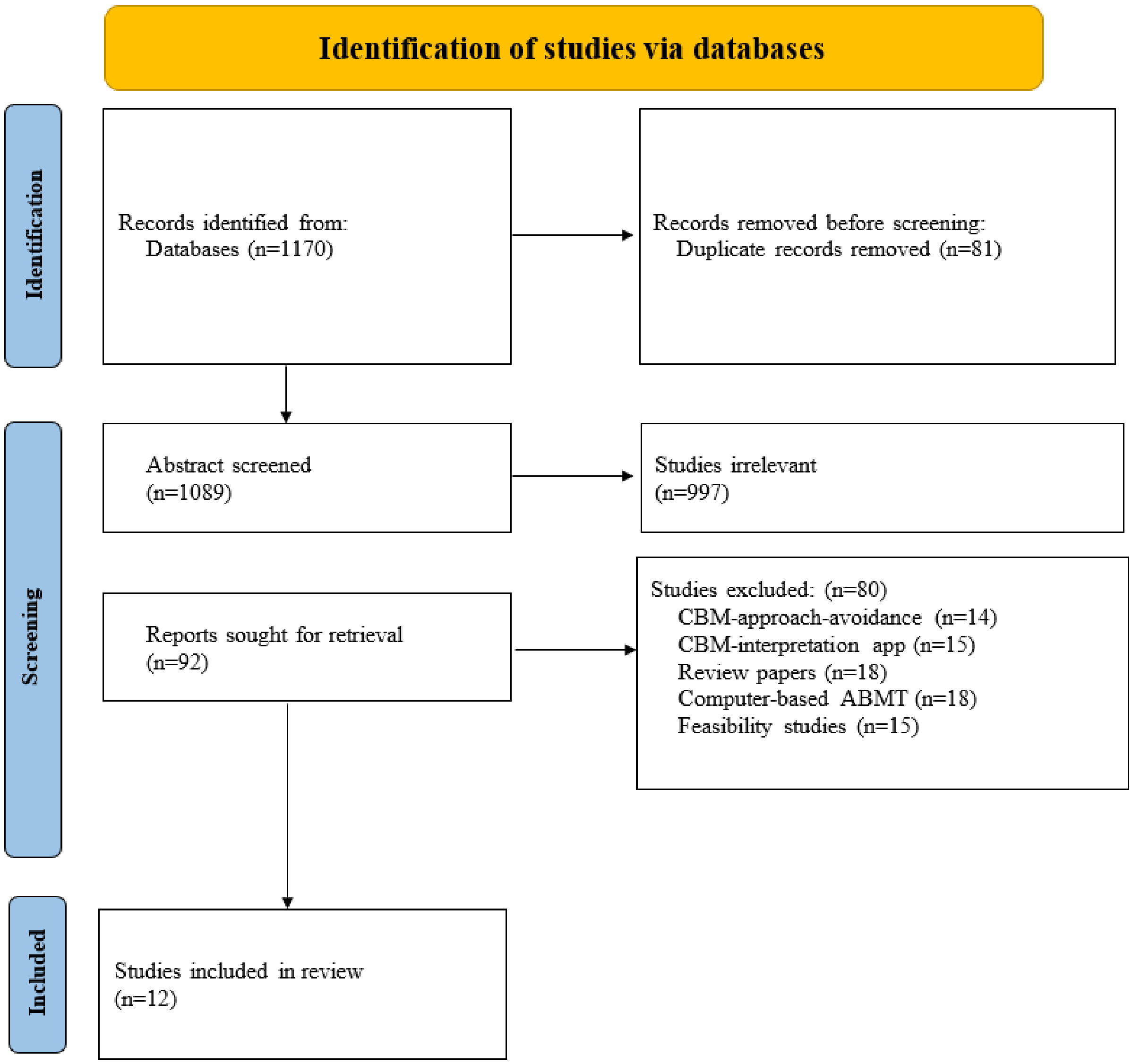
Search Strategy and Study Selection

In February 2023, MEDLINE, PsycINFO, PubMed, and Scopus were searched systematically for eligible studies published between 2014 and 2023 using keywords related to attentional bias and mobile apps: "attention bias" OR "cognitive bias" AND "smartphone" OR "smartphone application" OR "smartphone app" OR "mobile phones" OR "mobile application" OR "mobile app" OR "personal digital assistant." These keywords and databases were chosen based on a related prior study [26]. This decision was made to extend their research on the efficacy and design characteristics of ABMT. The rationale behind starting the search in 2023 was based on 2 similar reviews on smartphone-delivered cognitive bias modification interventions, which indicated that the first smartphone-delivered ABMT was developed in 2014 [32,33]. In addition, while other internet-based ABMT interventions can be accessed on mobile phones, our focus is exclusively on interventions developed specifically for mobile phones. Recognized articles were exported to the web-based systematic review software Rayyan (Rayyan Systems, Inc) [34], and duplicates were removed. The remaining articles were reviewed for inclusion by 3 independent authors using Rayyan.

Inclusion and Exclusion Criteria

To be considered for inclusion in this review, studies had to satisfy the following criteria: they had to be written in English and assess ABMT for mental health symptoms, such as anxiety, depression, stress, PTSD, or substance use (eg, smoking and alcohol consumption). ABMT interventions were explicitly defined and limited to those delivered exclusively through mobile devices. We included studies that met the following conditions: (1) ABMT was administered using a mobile device, such as a mobile phone, smartphone, or PDA, and (2) the delivery method took the form of a dedicated app or game. Articles that met any of the following criteria were excluded: (1) reviews, (2) interpretation bias modification delivered in any format, (3) avoidance bias modification delivered in any format, (4) web-based ABMT, and (5) computer-based ABMT. Web-based and computer-based ABMT interventions are similar in that both aim to modify attentional biases by training individuals to shift their focus away from negative stimuli. They offer interactive tasks and can be accessed remotely. However, web-based ABMT is accessed through an internet browser, making it more flexible and accessible across various devices, whereas computer-based ABMT usually requires specific software installed on a computer, potentially limiting accessibility. The objective of this review was to investigate ABMT that can be accessed only through smartphones, excluding the option of using other types of devices, to understand its unique effectiveness and accessibility. Therefore, ABMT interventions delivered through other platforms were excluded from this review. Figure 1 provides a visual representation of the inclusion and exclusion processes.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analyses) flowchart. ABMT: attentional bias modification training; CBM: cognitive bias modification.



Data Extraction and Risk-of-Bias Assessment

We used the Rayyan management software to initially screen studies based on their titles and abstracts. Subsequently, 3 authors (BB, MO, and HB) independently performed data extraction based on the predetermined eligibility criteria. The disagreements that arose during this process were effectively resolved through collaborative discussion among the authors.

The data extracted from each included study encompassed several key elements: author information, publication date, sample size, the delineation of sample groups (active and placebo), the description of the type of treatment, specifics regarding experimental tasks (dot-probe task, Stroop task, or visual search task), the type of threat stimuli (faces, pictures, or words), characteristics of threat stimuli, details about the number of stimuli and the type of stimulus array presented, the number of trials and sessions, stimulus presentation duration, and the outcome measurements used. Of the 5 authors, 2 (BB and YM) independently used the Cochrane risk-of-bias assessment tool to evaluate the risk of bias in the selected studies for the meta-analyses [35]. We also discussed the discrepancies to reach a consensus.

Data Analysis

We used the R Studio analysis packages (The R Foundation) [36] to conduct the meta-analyses. To perform these analyses, we used sample sizes for each group (active and placebo), along with means and SDs of mental health symptoms and attentional biases observed before and after the intervention (preintervention and postintervention assessments). These data were instrumental in calculating meta-estimates for both attentional bias levels and the reduction in mental health symptoms, encompassing anxiety, depression, stress, PTSD, or substance use. These meta-estimates were derived through random-effects meta-analyses.

The rationale for choosing random-effects meta-analyses is the anticipated substantial heterogeneity and aim to obtain a comprehensive overview of the true effect size while accounting for the variability among studies. Random-effects meta-analyses, fixed-effects meta-analyses, and Bayesian meta-analyses are common types of meta-analyses, each with distinct characteristics [37]. The choice of a random-effects model is often preferred when conducting a meta-analysis due to several key reasons. First, it is a flexible approach that can accommodate significant heterogeneity, which is common in meta-analyses involving diverse study populations and research questions. By allowing for varying effect sizes between studies, the random-effects model acknowledges the inherent variability in study results and provides more conservative estimates with wider CIs. This conservative approach is valuable, as it acknowledges the uncertainty associated with the underlying effect sizes and is less influenced by potential outliers.

The primary meta-analysis aimed to compute a comprehensive Hedges *g* effect size, accompanied by 95% CIs. This effect size was computed for both active and placebo ABMT interventions across all the studies included in our analysis. To interpret the effect sizes, we applied the Hedges *g* values of 0.20, 0.50, and 0.80, which correspond to small, moderate, and large effect sizes, respectively [38]. Heterogeneity was quantified using the I^2 statistic, and I^2>50% was considered evidence of substantial heterogeneity. We also used the inverse variance approach, a restricted maximum-likelihood estimator for τ^2, and the Q-profile method to establish CIs for τ^2 and τ, ensuring a robust analytical framework. Publication bias was examined using funnel plots, and the presence of asymmetry was assessed using the Egger regression test [39]. If the Egger test yields a significant result (indicating asymmetry), it suggests potential publication bias in the meta-analysis.

In addition, we performed separate sensitivity analyses using meta-regression with random-effects models in cases where there were enough study samples (at least 3), drawing reference from a previous study [27]. Sensitivity analysis is crucial for assessing the robustness of meta-analysis results and understanding the impact of potential sources of bias or variability, particularly in the presence of significant heterogeneity [40]. Moderators refer to specific factors or variables that can influence the relationship between the use of the smartphone-delivered ABMT intervention and its impact on mental health symptoms or attentional bias. These moderators can help researchers better understand the conditions under which the smartphone-delivered ABMT is effective and provide insights into the nuances of its outcomes. These meta-regression analyses allowed us to explore the influence of 5 moderators identified from the ABMT design characteristics reviewed, namely type of threat stimuli (face, images, or words); stimulus array type (left right or top down); design style (gamified or not gamified); display duration (200 ms or 500 ms) where applicable, along with the risk of bias (low or some concerns); and treatment groups (mental health and attentional bias) as additional considerations. These factors were identified as potential moderators.

Results

Study Selection

The initial search produced 1170 results, and we reduced this number to 1089 (93.08%) by eliminating duplicates. We excluded 997 (91.55%) articles after reviewing their titles and abstracts. Moving on to the full-text level, we reviewed the remaining 92 (8.4%) articles. Out of these 92 articles, we retained only 12 (13%) articles that were relevant to this paper. During the extraction process, we excluded 80 (87%) articles out of the initially considered 92. The article screening process is detailed in Figure 1. In Table 1, we have summarized the characteristics of the included studies. The sample sizes varied, ranging from 18 to 22,993 participants across different studies. The studies evaluated active and placebo interventions for 5 different symptoms identified in 24,503 participants. These 5 symptoms include anxiety, depression, stress, PTSD, and substance use. The treatments involved tasks used include dot-probe, Stroop, and visual search tasks, with threat or addictive-cue stimuli, including faces, words, and images. The duration of stimulus presentation was typically 500 milliseconds, and the number of stimuli varied between 2 and 16 per array type. The number of trials conducted in these studies ranged from 60 to 800. Overall, these studies encompassed a diverse range of sample sizes and characteristics, reflecting their focus on different mental health problems and intervention strategies.

Table 1. Study characteristics (N=12).

Study	Sample size, N	Active, n (%)	Placebo, n (%)	Treatment	Tasks	Threat or addictive-cue stimuli	Duration (ms)	Stimulus, n and array type	Trials, N
Dennis and O'Toole [41], 2014	76	Short: 18 (24); long: 19 (25)	Short: 20 (26); long: 19 (25)	Anxiety	Dot probe	Face	500	2 and top down	640
Enock et al [42], 2014	326	158 (48.5)	141 (43.3); WL[a]: 27 (8.3)	Anxiety or depression	Dot probe	Face	500	2 and top down	160
Dennis-Tiwary et al [43], 2016	42	19 (45)	23 (55)	Anxiety	Dot probe	Face	500	2 and top down	480
Yang et al [44], 2017	40	20 (50)	20 (50)	Anxiety	Dot probe	Face	500	2 and top down	800
Dennis-Tiwary et al [24], 2017	29	15 (52)	14 (48)	Anxiety or stress	Dot probe	Face	500	2 and top down	160
Teng et al [45], 2019	82	30 (37)	30 (37); WL: 22 (27)	Anxiety	Dot probe	Word	500	Left and right	82
Flaudias et al [46], 2020	41	18 (44)	Memory group: 15 (37); no AB[b]: 8 (20)	Alcohol use	Stroop	Image	500	3 and grid	240
Niles et al [47], 2020	546	Personalized: 177 (32.4); nonpersonalized: 179 (32.8)	190 (34.8)	Anxiety or PTSD[c]	Dot probe	Word	500	2 and top down	70
Charvet et al [23], 2021	35	High anxiety: 17 (49); low anxiety: 13 (37)	No placebo group	Anxiety or depression	Dot probe	Face	500	2 and top down	120
Chelliah and Robinson [48], 2022	22,993	Dot 4448: 4448 (19.34); Dot 2588: 2588 (11.26)	Dot 4301: 4301 (18.7); Dot 4818: 4818 (20.95); no training: 6778 (29.47)	Anxiety	Visual search	Face	500	16 and grid	100
Flaudias et al [49], 2022	47	20 (43)	27 (57)	Alcohol use	Stroop	Image	500	4 and grid	60
Robinson et al [50], 2022	246	124 (50.4)	122 (49.6)	Anxiety or substance use disorder	Dot probe and Stroop	Image and word	500	2 and left right	440

[a]WL: waiting list.

[b]AB: attentional bias.

[c]PTSD: posttraumatic stress disorder.

Risk of Bias in the Studies

The risk-of-bias assessment for the included studies was conducted across 5 specific domains (Figure 2 [23,24,41-47,49,50] and Figure 3): *randomization, deviation from intended intervention, missing outcome data, measurement of the outcome, and selection of the reported result.* Among the 11 studies analyzed, the majority demonstrated a low risk of bias across all these domains, indicating a generally robust methodology in these investigations. Specifically, 9 (82%) out of 11 studies were categorized as having a low overall risk of bias, implying a high level of confidence in their research findings. However, 1 (9%) out of 11 studies raised some concerns, primarily in the domains of randomization, deviation from intended intervention, and missing outcome data. In addition, 1 (9%) out of 11 studies showed a high risk of bias, particularly in the domain of measurement of the outcome. Our assessment of the risk of bias in individual studies is presented in Multimedia Appendix 2 [23,24,41-47,49,50]. These findings underscore the overall quality and reliability of the studies, with the majority (9/11, 82%) exhibiting a low risk of bias in their design and execution, while a small proportion raised some concerns (1/11, 9%) or demonstrated a high risk of bias (1/11, 9%) in specific domains.

Figure 2. Risk-of-bias domains.

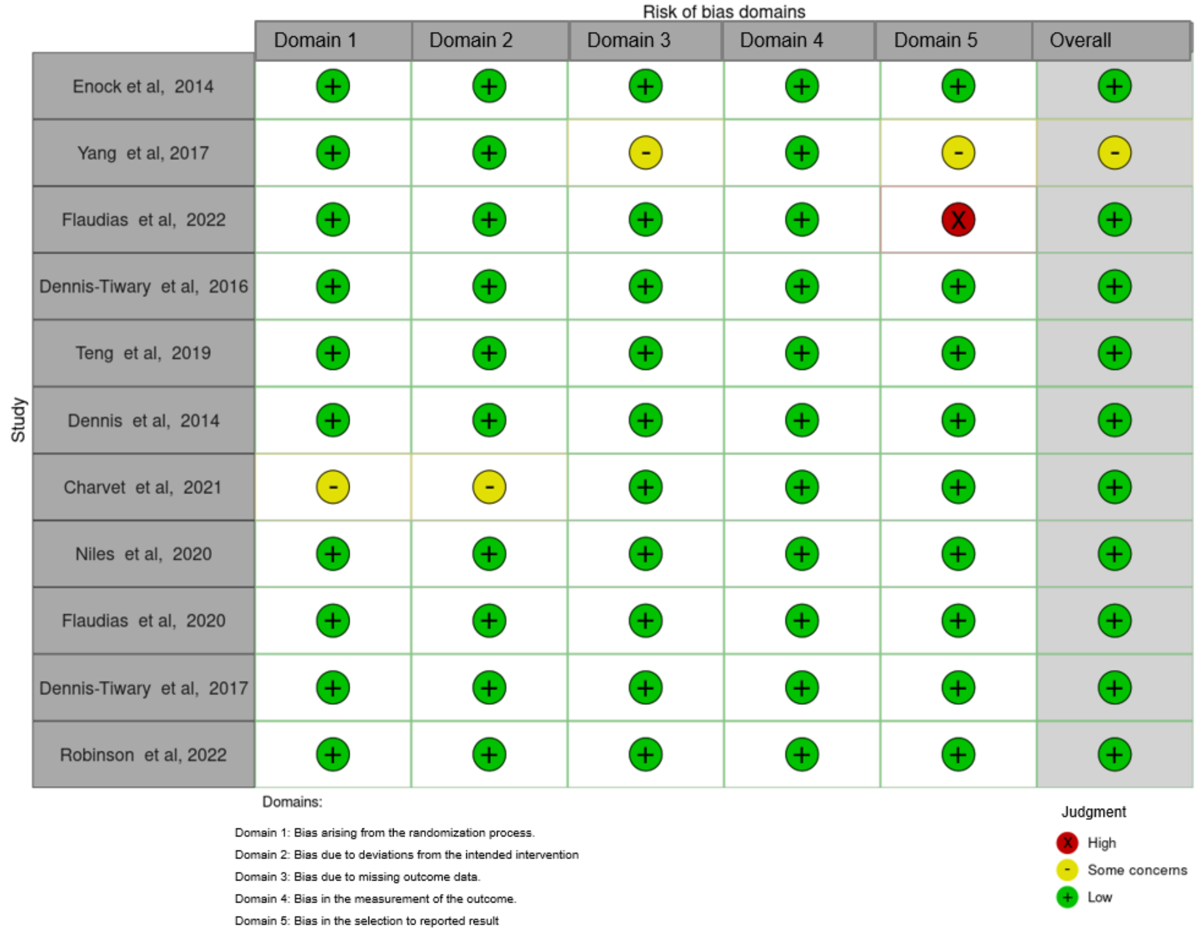
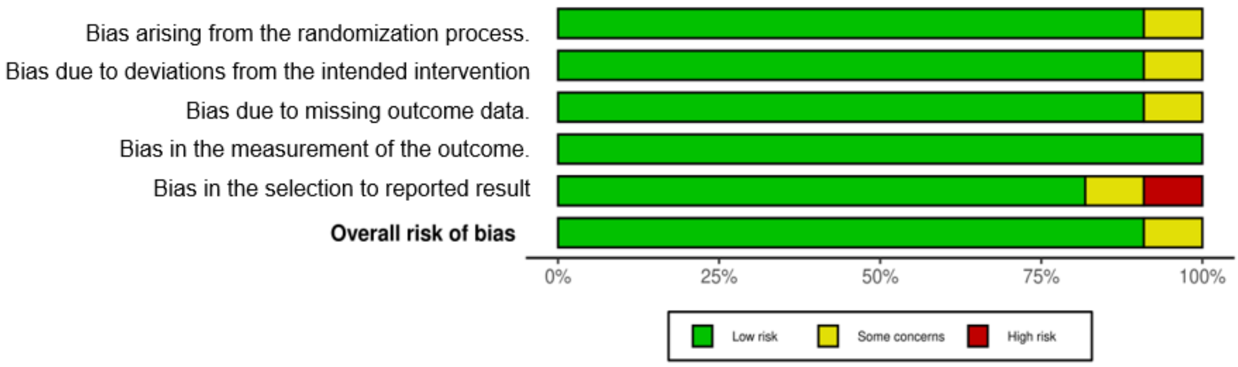


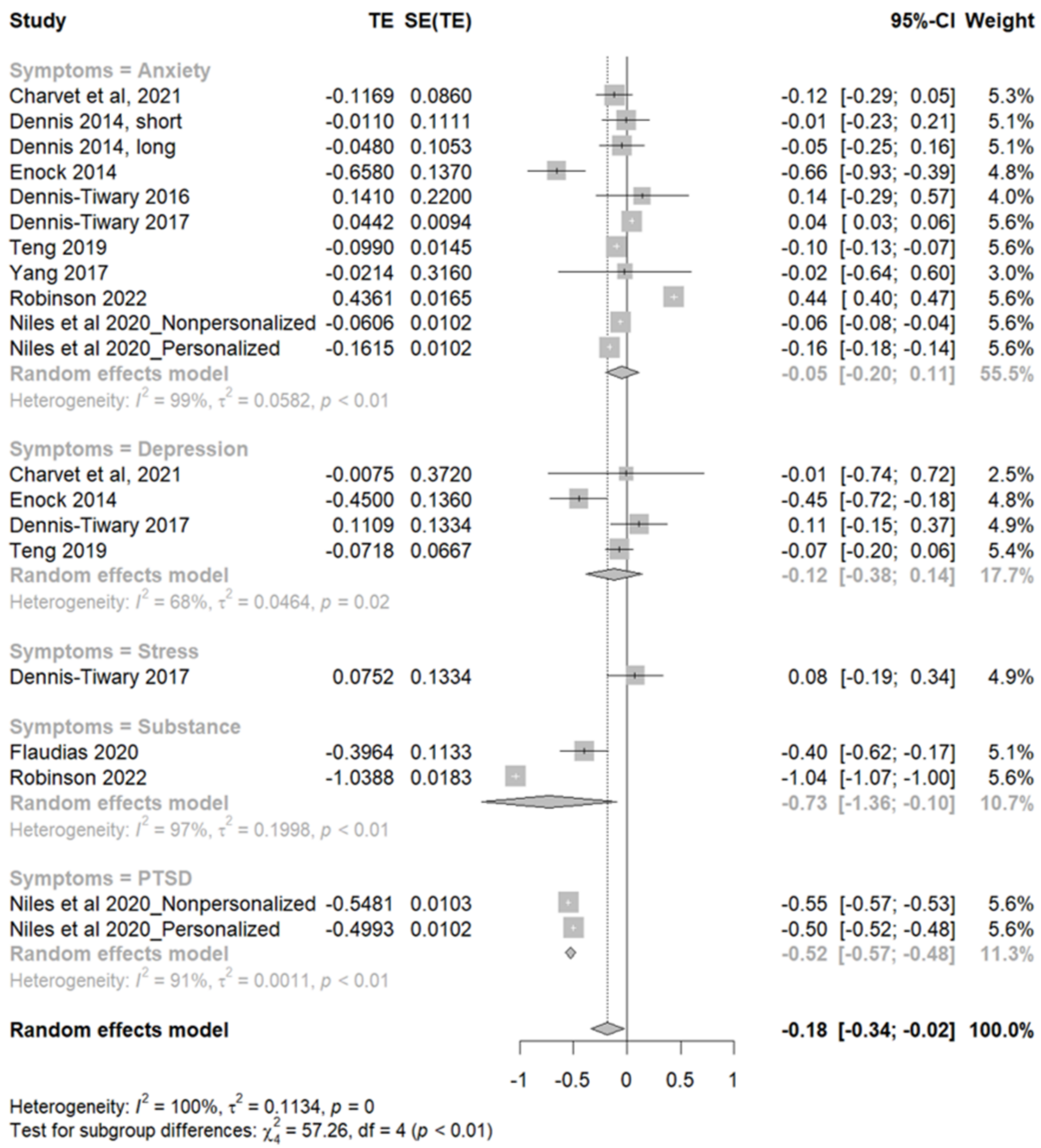
Figure 3. Overall risk of bias.



Effectiveness of Smartphone-Delivered Active ABMT for Mental Health Symptoms

In this comprehensive analysis involving 20 study samples (Figure 4 [23,24,41-45,47,50]), the pooled effect size for the study samples reflected a significant effect of active ABMT on mental health symptoms (Hedges g=–0.18, 95% CI –0.340 to –0.02; z score=–2.27; P=.03). Specifically, the negative value of the effect size suggests that the symptoms decreased after the intervention. In addition, the statistical tests conducted confirm that this reduction is unlikely to have occurred by chance, suggesting that active ABMT can be effective in alleviating mental health symptoms. Significant heterogeneity was observed among the study samples (Q=6526.76; P<.001; I^2=99.7%). The subgroup meta-analysis revealed diverse effects of interventions across 5 distinct symptom categories, namely anxiety, depression, stress, substance use, and PTSD, with the varying impacts of interventions and heterogeneity. Sensitivity analysis of the compiled *P* values presented in Multimedia Appendix 3 [23,24,41-47,49,50] showed that specific studies, including the studies by Dennis-Tiwary et al [43] (*P*=.02) and Robinson et al [50] (*P*=.005), emerged as influential contributors to the overall significance. Despite their exclusion, the meta-analysis maintained statistical significance, reaffirming the primary findings' solidity. The test of the asymmetry funnel plots presented in Multimedia Appendix 4 showed no evidence of publication bias (t_{18}=0.31, *P*=.76).

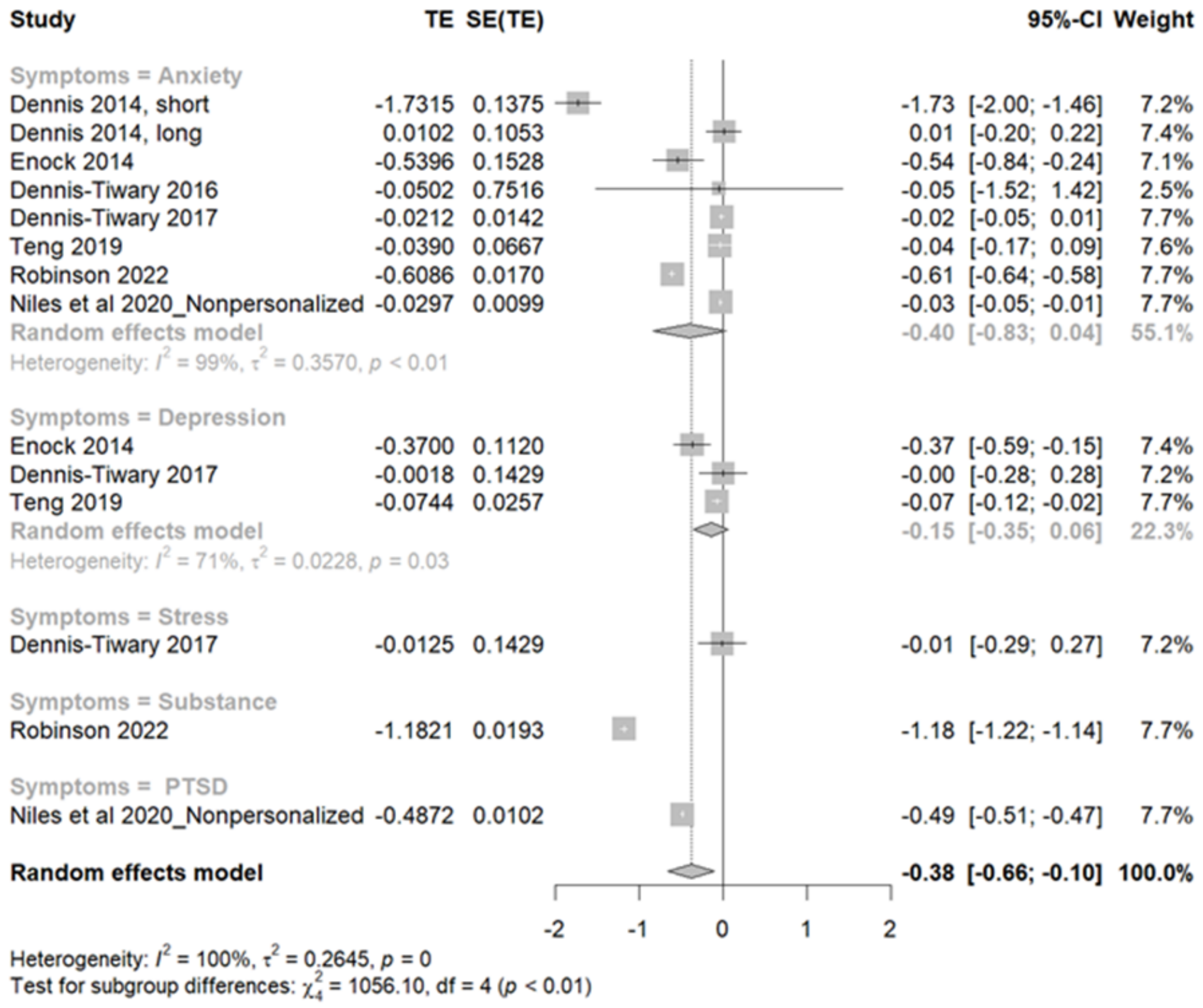
Figure 4. Forest plots for active attentional bias modification training for mental health symptoms. PTSD: posttraumatic stress disorder; TE: treatment effect.



Effectiveness of Smartphone-Delivered Placebo ABMT for Mental Health Symptoms

In this analysis involving 14 study samples (Figure 5 [24,41-43,45,47,50]), the outcomes revealed a significant effect size (Hedges *g*=–0.381, 95% CI –0.8307 to 0.0403; *z* score=–2.66; *P*=.008), and significant heterogeneity was observed among the samples (*Q*=559.83; *P*=.002; I^2=99.7%). In essence, the negative value of the effect size suggests that the symptoms decreased after the intervention. In addition, the statistical tests conducted confirm that this reduction is unlikely to have occurred by chance, suggesting that placebo ABMT can be effective in reducing mental health symptoms. In the subgroup analysis, diverse effects of interventions across 5 distinct symptom categories, namely anxiety, depression, stress, substance use, and PTSD, were observed. Sensitivity analysis of the compiled *P* values presented in Multimedia Appendix 3 revealed that among the 14 study samples considered, the exclusion of 2 specific study samples, short active AMBT (*P*=.009) and long active AMBT (*P*=.006) from the study by Dennis and O'Toole [41], was found to exert substantial influence, significantly impacting the overall statistical significance. Notably, even with the removal of these influential study samples, the overall analysis sustained its statistical significance. The results of the test of the asymmetry funnel plots are displayed in Multimedia Appendix 4. The results of the linear regression test conducted to assess funnel plot asymmetry yielded a nonsignificant outcome (t_{12}=–0.35, *P*=.73), indicating that there was no publication bias.

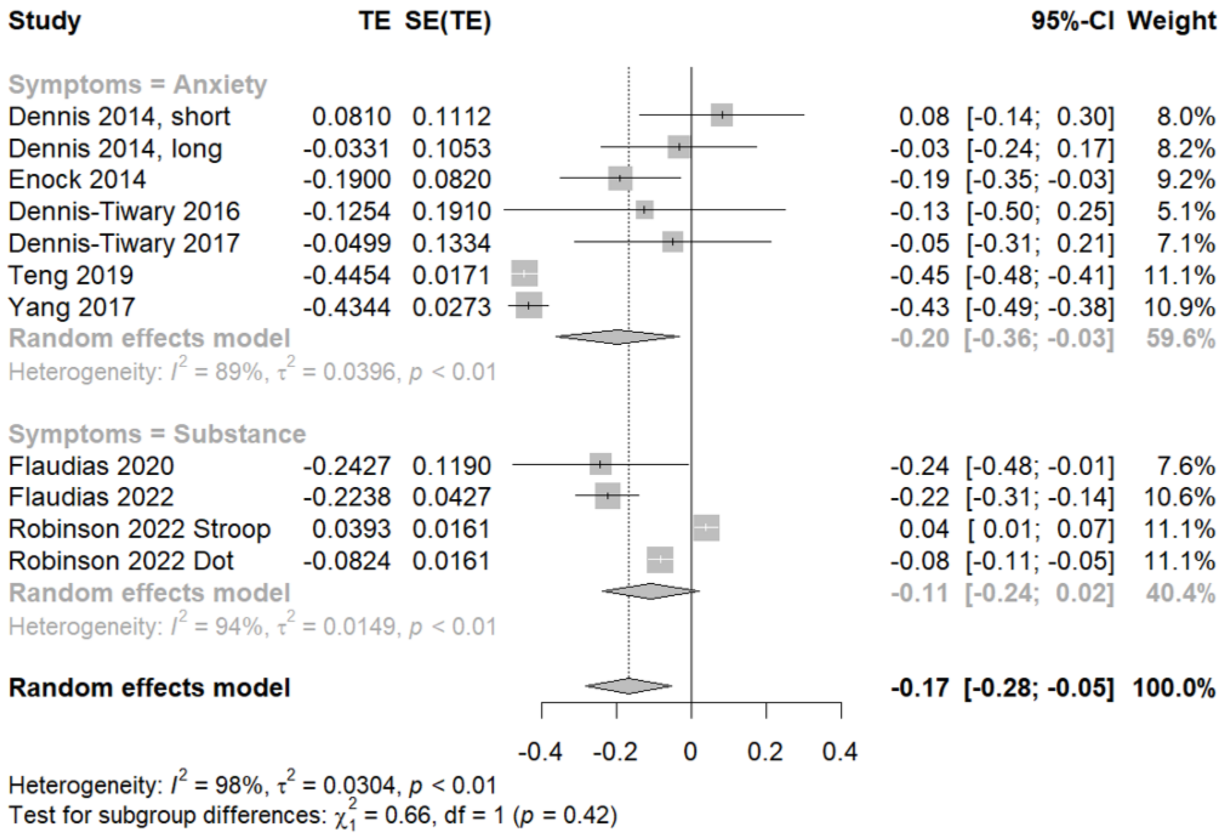
Figure 5. Forest plots for placebo attentional bias modification training for mental health problems. PTSD: posttraumatic stress disorder; TE: treatment effect.



Effectiveness of Smartphone-Delivered Active ABMT for Attentional Bias

In the analysis involving 11 study samples evaluating attentional bias, we observed a significant outcome using a random-effects model (Figure 6 [24,41-45]). The pooled effect size analysis for the study samples showed a significant effect of active ABMT on attentional biases (Hedges *g*=–0.17, 95% CI –0.28 to 0.05; *z* score=–2.87; *P*=.004). In summary, the negative value of the effect size suggests that the attentional bias decreased after the intervention, and the statistically significant effect did not occur by chance, suggesting that active ABMT is effective in reducing attentional bias.

Figure 6. Forest plots for active attentional bias modification training for attentional bias.

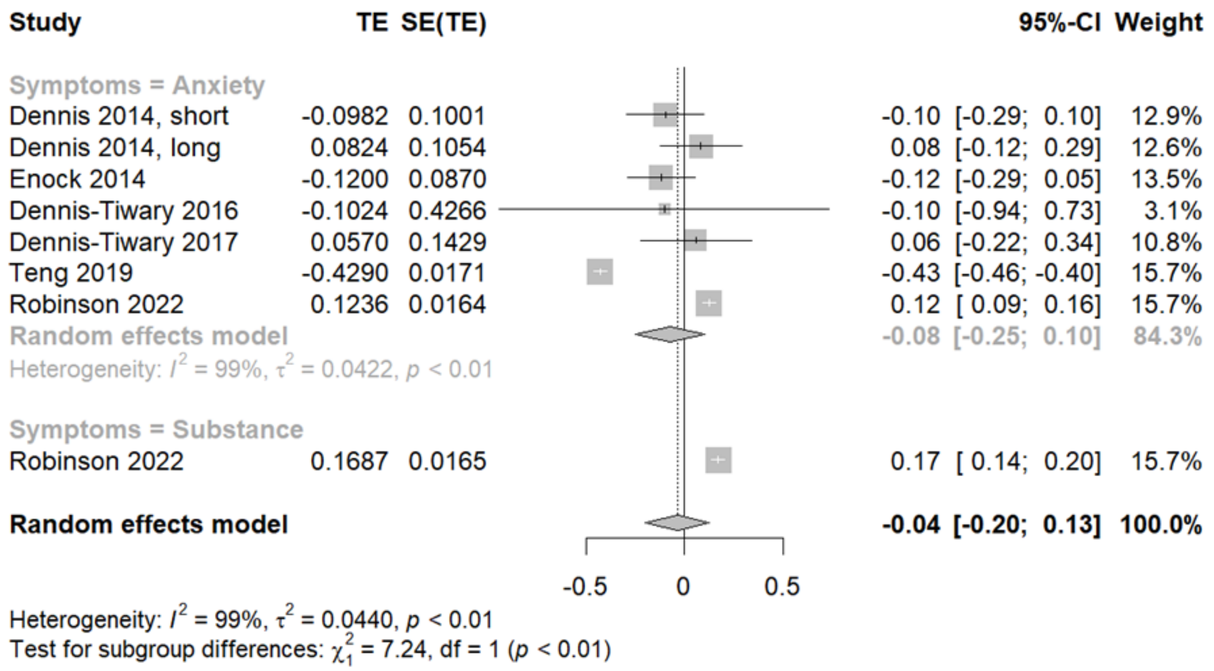


Significant heterogeneity was observed among all the samples (Q=559.83; P=.004; I^2=98.2%). The analysis of the subgroups within the study samples showed different effects and significant heterogeneity among the 2 different category symptoms: anxiety and substance use. Sensitivity analysis of the compiled P values presented in Multimedia Appendix 3 revealed that even after the removal of specific study samples, the overall analysis maintained statistical significance, reaffirming the efficacy of ABMT in reducing attentional bias. The test of the asymmetry funnel plot is displayed in Multimedia Appendix 4. The Egger regression test found no evidence of publication (t_9=–0.09, P=.93).

Effectiveness of Smartphone-Delivered Placebo ABMT for Attentional Bias

In this analysis comprising 8 study samples, as shown in Figure 7 [24,41-43,45,50], the pooled effect size did not reflect a significant effect of placebo ABMT on attentional biases (Hedges g=–0.04, 95% CI 0.200-0.13; z score=–.44; P=.66). The negative value of the effect size suggests that the attentional bias decreased after the intervention; however, the effect size was not statistically significant, suggesting that placebo ABMT is not effective in reducing attentional bias. Significant heterogeneity was observed among the study samples (Q=779.84; P<.001; I^2=99.1%). The subgroups within the random-effects model unveiled varying effects and heterogeneity linked to the different symptom categories of anxiety and substance use. The sensitivity analysis based on the compiled P values in Multimedia Appendix 3 indicated that even when certain studies were excluded, the overall analysis did not attain statistical significance. The test of asymmetry funnel plot is displayed in Multimedia Appendix 4; however, the analysis was not conducted because the study samples were too small to be included in the meta-regression, but the funnel plot showed asymmetry.

Figure 7. Forest plots for placebo attentional bias modification training for attentional bias.



Moderator Analyses

The moderator analysis focused exclusively on anxiety and depression, given the limited samples in other subgroups. Only the anxiety subgroup is discussed, as the depression subgroup did not exhibit significant effects on all the moderating parameters, as shown in Multimedia Appendix 5 based on the raw data provided in Multimedia Appendix 6 [23,24,41-47,49,50]. Results of the moderator analysis using meta-regression showed that the choice of "stimuli" played a significant role in shaping treatment outcomes. Specifically, when "images" were used as stimuli for individuals with anxiety symptoms, a significant effect on anxiety reduction was observed, suggesting that this stimulus type may be particularly effective in this subgroup. In contrast, the use of "face" stimuli for patients with primary anxiety symptoms and "word" stimuli for those with anxiety and PTSD did not yield significant effects on reducing anxiety symptoms, and the use of "word" stimuli for those with anxiety related to PTSD did not yield significant effects on anxiety outcomes.

Moreover, the "display duration" of 200 milliseconds emerged as a significant moderator (β=.436; *P*=.002), indicating that shorter exposure durations may lead to more substantial reductions in symptoms, while a longer duration of 500 milliseconds showed a negative effect on outcomes (β=–.537; *P*<.001). The other factors, such as "stimulus array type," "design style," and "risk of bias," did not exhibit significant moderating effects on anxiety outcomes within ABMT interventions. These findings illustrate the importance of tailoring ABMT interventions based on the specific type of threat stimuli and the characteristics of the target population to optimize their efficacy, thus highlighting the nature of ABMT's impact on anxiety reduction.

Discussion

Principal Findings

Overview

This paper presents a systematic review and meta-analysis of 12 individual studies including 20 independent samples. The overarching goals were to (1) compute the overall effect sizes for active ABMT and placebo ABMT on the reduction of attentional bias and symptoms of mental health problems and (2) separately evaluate which ABMT design characteristics moderate the effect sizes in reducing mental health problems and attentional bias. Importantly, this analysis included only randomized controlled trial designs and study samples that included pre-post modification comparisons of mental health symptoms and attentional biases. Our study diverges from earlier reviews focused on gamification elements and commercialized apps by prioritizing the core phenomenological characteristics of ABMT: Posner task, Stroop task, dot-probe task, and visual search task. Embedding these established tasks in smartphone interventions ensures scientific rigor and enhances ABMT effectiveness. Furthermore, our research identifies current smartphone approaches and explores novel adaptations of these tasks for optimal integration into mobile platforms. This exploration provides practical insights for designers and developers, guiding interface design, interaction mechanics, and task presentation to boost engagement and adherence among mobile users. These insights are pivotal for advancing mobile-based ABMT, ensuring interventions are scientifically grounded and efficacious in enhancing mental health outcomes. In addition, our study represents the first meta-analysis focused on smartphone-based ABMT. Unlike previous reviews, we included both placebo and active designs to rigorously assess the efficacy of smartphone-based ABMT interventions. Using a randomized controlled trial design with a diverse participant sample, we enhanced methodological robustness and applicability. These contributions set a high standard for future research in smartphone-based ABMT, aiming to improve the accessibility and effectiveness of mental health interventions broadly.

Overall Effect

Overall, this review found that using ABMT through smartphones had a small but significant impact (Hedges *g*=0.17), whereas Hakamata et al [51] reported a larger effect size (Cohen *d*=0.51). The studies by Heeren et al [52] and Hang et al [53] reported small effect sizes for computer-based ABMT in reducing anxiety, with Hedges *g*=0.41 for social anxiety and Hedges *g*=0.26 for anxiety disorders, respectively. It is important to note that the study by Hakamata et al [51] encompassed various delivery methods without a specific focus, while the studies by Heeren et al [52] and Hang et al [53] concentrated on the computer-based medium. In contrast, this review's investigation specifically centered on ABMT via smartphones. These results highlight that both smartphone-based and computer-based ABMT interventions can be effective in reducing social anxiety symptoms, despite variations in effect size. The effect size variation across media may be due to the difference in screen sizes between computers and smartphones. Exploring the differences in platforms for the delivery of ABMT will be insightful intervention for mental health problems . Despite variations in effect size between studies, the findings by Hakamata et al [51] reinforce the notion that design characteristics play a crucial role in determining the effectiveness of ABMT interventions, aligning with the observations of this study. This suggests that while the magnitude of the effect may differ, the underlying factors contributing to the efficacy of ABMT remain consistent across different research contexts. In essence, the emphasis on design characteristics highlights the need for the careful consideration of intervention parameters to optimize outcomes in ABMT research and practice. Furthermore, the meta-analysis of combined ABMT and CBT also showed small but significant effects on clinician-rated anxiety symptoms and attentional bias toward threat [54]. Smartphone-delivered ABMT showed similar effects to other mental health apps for anxiety and depression [27,55,56]. In contrast, the placebo ABMT showed a small and not significant effect size for reducing attentional bias toward negative stimuli (Hedges *g*=–0.04; *P*=.66), but there was a significant small to moderate effect size for reducing mental health for anxiety and depression (Hedges *g*=–0.38; *P*=.008).

Our findings on placebo ABMT revealed a significant and moderate effect size in reducing mental health problems but not attentional bias. Other studies have shown that both active and placebo ABMT interventions significantly reduce mental health problems [21,45,57], which supports our findings. These findings regarding symptom reduction across active and placebo ABMT interventions suggest a potential placebo effect, emphasizing the need for further investigation into the mechanisms driving these outcomes and the broader implications for understanding placebo treatment effects in therapeutic interventions.

The significant effects of both active ABMT and placebo ABMT on symptoms of mental health problems can be explained through several mechanisms supported by previous studies. Active ABMT works by specifically targeting and modifying attentional biases toward threat-related stimuli, which are often implicated in anxiety and other mental health conditions. This modification reduces cognitive load and emotional reactivity, leading to symptom reduction [58,59]. In addition, active ABMT enhances emotional regulation by training individuals to redirect their attention away from negative stimuli. In contrast, placebo ABMT may improve symptoms through expectation effects, where participants' belief in receiving effective treatment triggers neurobiological responses contributing to symptom improvement [60]. Moreover, the structured nature of placebo ABMT improves therapeutic engagement and support, enhancing feelings of control and self-efficacy [42]. Finally, the cognitive engagement involved in placebo ABMT tasks can improve cognitive functioning, indirectly benefiting mental health. These mechanisms collectively explain the significant effects observed in both treatment conditions.

The significant heterogeneity observed in the effectiveness of smartphone-delivered active ABMT for mental health symptoms across studies could be attributed to several factors. First, there are variations in study populations and sample sizes, ranging from small groups of 29 participants [43] to large samples of 22,993 participants [48], which could affect the generalizability and robustness of the results. Second, the type of mental health conditions addressed varies, with studies targeting anxiety, depression, PTSD, alcohol use, and substance use disorders, potentially leading to different outcomes based on the specific symptoms being treated. Third, the intervention protocols differ significantly; for instance, while most studies use the dot-probe task with face stimuli, others use the Stroop task with images or words or a visual search task. In addition, the duration of exposure to stimuli and the number of trials vary widely, from 60 trials [49] to 800 trials [44], which might influence the effectiveness of the intervention. These methodological differences, including those in the type of stimuli, the structure of the tasks, and the specific parameters of the intervention, contribute to the observed heterogeneity in the outcomes of these studies.

However, sensitivity analysis emerged as a critical tool in navigating this heterogeneity, allowing for a thorough examination of the robustness of the findings [27]. Despite the significant observed heterogeneity among the study samples, sensitivity analysis revealed that the meta-analysis maintained its statistical significance even after excluding influential studies. This indicates that the overall findings remained robust and reliable, despite the presence of variability across the included studies. By systematically assessing the impact of individual studies or study characteristics on the overall results, sensitivity analysis provided valuable insights into the stability of the meta-analysis outcomes.

Comparison of Smartphone-Delivered ABMT With Computer-Delivered ABMT

The current findings on smartphone-delivered ABMT and its effects on attentional bias and mental health symptoms, particularly for anxiety and depression, are compared with mixed significant effects reported in other meta-analyses on ABMT delivered through other platforms, that is, computer-based (internet-based) ABMT interventions.

Considering computer-based ABMT platforms, the study by Heeren et al [52] primarily focused on their application for social anxiety disorder. The meta-analysis reported small but significant effect sizes for ABMT in reducing attentional bias (Hedges *g*=0.30) and social anxiety disorder symptoms (Hedges *g*=0.41 for training toward neutral stimuli vs control condition) after multiple sessions. The study also noted that the control conditions often performed similarly to the ABMT, suggesting a potential placebo effect or the nonspecific benefits of participating in a study. Similarly, the study by Hang et al [53] focused on computer-based ABMT but extended the discussion to children and adolescents with anxiety disorders. This meta-analysis found that ABMT had small but significant effects on clinician-rated anxiety symptoms (Hedges *g*=0.26) and attentional bias toward threat (Hedges *g*=0.21) but not on self-reported or parent-reported anxiety measures (Hedges *g*=–0.08). The control conditions used in these studies, such as attention control training, did not show significant effects, which contrasts with the significant effects seen in this study for placebo ABMT on reducing mental health symptoms for anxiety and depression.

This review's findings suggest a potential advancement in the delivery method of ABMT. The use of smartphones could increase accessibility and adherence to ABMT protocols, potentially enhancing their effectiveness. These results build upon the findings from the studies by Heeren et al [52] and Hang et al [53] by suggesting that the delivery method of ABMT (eg, via smartphone) and the nature of the control condition can significantly influence the outcomes of ABMT interventions. They also highlight the importance of considering placebo effects in the design and interpretation of ABMT studies, as nonspecific factors can sometimes produce significant improvements in mental health symptoms. This underscores the need for well-designed studies to carefully assess the specific contributions of ABMT techniques to changes in attentional bias and mental health outcomes.

Moderating Factors of ABMT Intervention

The moderator analysis conducted in this review comprehensively evaluated the effectiveness of ABMT design characteristics across various mental health problems, including anxiety, depression, PTSD, and substance use disorders. The analysis shows that certain characteristics, such as the threat stimulus type, spatial arrangement of stimuli, and duration of stimuli display, have an influence on ABMT's effectiveness. This finding is consistent with a previous meta-analysis that showed that the use of ABMT interventions may be owing to their varying design characteristics and optimal protocols (eg, task types, target stimuli, stimulus directions, and display settings) [61]. However, due to the limited number of study samples (<3) available in the other mental health problem categories, the analysis focused only on anxiety and depression. Delving into the intricacies of ABMT, we first explored the role of its design characteristics, including the type of threat stimuli, the spatial arrangement of stimuli, and the duration of stimuli display, in influencing ABMT's effectiveness.

Given the findings from our moderator analyses highlighting the significant influence of various design characteristics on ABMT's effectiveness, there is a compelling rationale for exploring personalized approaches in future smartphone-based ABMT interventions. While our review did not directly analyze personalized ABMT due to the limited studies with personalized features, the identified moderating factors offer valuable insights into potential avenues for customization. For example, the choice of stimuli, spatial arrangement, and the duration of stimulus display emerged as critical factors influencing treatment outcomes. Building on these insights, future research could investigate how tailoring ABMT protocols to individual preferences and needs, based on these design characteristics, could enhance treatment efficacy. By developing personalized ABMT interventions that align with patients' specific attentional biases and cognitive profiles, we may optimize treatment outcomes and improve overall engagement and adherence. Therefore, we propose personalized approaches as a promising direction for further exploration in the field of smartphone-delivered ABMT interventions.

Personalization of ABMT Interventions on Smartphone Platforms

Personalization in the context of ABMT refers to the customization of intervention components to align with individual preferences, needs, and characteristics. This customization can encompass various aspects of the intervention, including the stimulus selection, presentation format, difficulty levels, and session duration. By tailoring the intervention to everyone, personalization aims to enhance engagement, relevance, and effectiveness, ultimately optimizing treatment outcomes.

The personalization of ABMT interventions on smartphone platforms and that of ABMT interventions on PCs can differ significantly due to the unique characteristics and capabilities of each device. However, the personalization of ABMT design can be achieved on both smartphones and PCs by adding the option of selecting the type of stimuli and duration of stimuli display. However, personalization on smartphone platforms tends to be more dynamic, context aware, and automated, leveraging the device's sensors and data processing capabilities to tailor ABMT interventions to the user's current context and needs.

Smartphone-based ABMT offers unique advantages for personalization compared to traditionally delivered (computer-based) ABMT. With smartphones, users have greater control and flexibility in customizing intervention parameters according to their preferences. For example, individuals with anxiety may respond differently to various types of stimuli, as highlighted by this study. Smartphone apps can allow users to select their preferred stimulus types, such as images, faces, or words, based on their personal preferences and comfort levels. This level of customization enables users to engage with stimuli that resonate the most with them, potentially enhancing their attentional training experience and improving treatment outcomes.

Similarly, smartphones can be leveraged to track personalized data, such as location and physical activity patterns, using their sensors. These data may be used to inform real-time feedback on user performance. In contrast, personalization on PCs may rely more on user input and manual customization, offering greater control but potentially limiting the adaptability and responsiveness of the intervention. The use of mobile sensors in personalizing mobile health interventions is highlighted in the meta-analysis by Tong et al [62]. The study found that interventions using system-captured data, which can be obtained from mobile sensors, were associated with higher effectiveness compared to those using user-reported data. Specifically, the standardized difference in means for interventions using system-captured data was 1.48 (95% CI 0.76-2.19), indicating a more substantial impact on lifestyle behavior outcomes when mobile sensors are used for personalization. Despite the advantages of the smartphone-delivery platform, it also comes with potential challenges, such as limited screen size, potential distractions, and variability in device capabilities across different users. It is essential to carefully consider these factors when designing smartphone-based ABMT interventions to ensure optimal effectiveness and user engagement.

The Impact of ABMT Characteristics on Anxiety and Depression Symptoms

Our research findings shed light on the impact of ABMT design characteristics on intervention outcomes for anxiety symptoms, providing valuable insights into the diverse factors that shape the efficacy of ABMT interventions. One pivotal aspect of these design characteristics is the influence of threat stimuli, a factor that significantly affects the effectiveness of ABMT. Our results align with previous studies, such as those conducted by Xia et al [61], emphasizing how the nature of these stimuli notably shapes the efficacy of ABMT interventions. Another critical factor is the spatial arrangement of stimuli, whether presented in a top-down or left-right fashion. Interestingly, the top-down arrangement is found to significantly reduce anxiety symptoms, echoing the findings of a previous meta-analysis that demonstrated the impact of spatial stimulus display on the outcome of ABMT treatment [61]. It is noteworthy that the design style, whether gamified or not, does not significantly impact ABMT outcomes. This finding can be linked to a recent review on gamified ABMT, where 2 (50%) out of 4 studies did not reduce mental health problems, while the other 2 (50%) studies did. These mixed results highlight the need for the further exploration of gamified ABMT, as identified in the review, given the limited number of studies conducted in this field [61].

Moreover, the duration for which stimuli are displayed significantly influences ABMT's effectiveness, with a 200-millisecond display duration emerging as a significant moderator compared to a 500-millisecond display duration. This indicates that shorter exposure durations may lead to more substantial reductions in anxiety symptoms. This aligns with prior research, such as the work by Charles et al [63], emphasizing the importance of optimizing stimulus presentation duration in ABMT protocols. In addition, a co-design study by Zhang et al [64], involving both health care professionals and patients, aimed to enhance conventional ABMT. Their recommendation to initiate training with a lengthier stimulus presentation interval and then gradually reduce the interval has proven instrumental in enhancing engagement and reducing assessment time. Subsequently, Zhang et al [65] adopted this approach, reinforcing the effectiveness of a 200-millisecond duration by presenting participants with a 500-millisecond fixation cross, followed by images for 200 milliseconds.

From the ABMT protocol perspective, the risk of bias does not appear to impact ABMT outcomes, as evidenced by nonsignificant trends in both low and some concern categories. This corresponds with an existing meta-analysis study that has shown the role of bias risk and intervention types in determining the outcomes of ABMT interventions [27]. Finally, the findings from this study on the effect of intervention types (active and placebo) on ABMT outcomes reveal that the intervention types in the anxiety treatment group do not significantly impact ABMT outcomes. In summary, these significant moderators offer nuanced insights into optimizing the design and implementation of ABMT interventions for anxiety, establishing direct connections to existing literature and enhancing the understanding of the multifaceted influences on treatment effectiveness.

When considering threat stimuli, there is a nonsignificant negative effect for both face and words, suggesting a subtle reduction in depression symptoms. The stimulus array type shows no significant impact for either left-right or top-down arrangement. The design style, whether gamified or not, does not significantly influence ABMT outcomes for depression. Notably, within the depression treatment group, both active and placebo interventions exhibit negative but nonsignificant effects, indicating comparable impacts on ABMT outcomes. These findings provide a detailed understanding of the role of these moderators in shaping the effectiveness of ABMT interventions for depression.

Limitations

This meta-analysis provides valuable insights into the effectiveness of active and placebo ABMT interventions for reducing mental health problems, particularly anxiety and depression. However, it is crucial to acknowledge several limitations that should be considered when interpreting these findings. First, the study's reliance on small sample sizes within the selected studies limits the generalizability of the results. Future research with larger and more diverse samples could provide a more comprehensive understanding of the effects of ABMT on mental health problems, as current research has primarily focused on high-income countries. Second, the high heterogeneity observed among the included study samples poses a challenge to drawing definitive conclusions. This heterogeneity calls for further investigation into which specific elements of ABMT are the most impactful in reducing mental health problems and whether certain subgroups of individuals may benefit more than others. Despite the significant heterogeneity observed among the study samples, sensitivity analysis revealed that the meta-analysis maintained its statistical significance. The sensitivity analysis conducted in this study should be interpreted with caution. While it helps assess the robustness of the findings, it relies on assumptions that may not always hold. The role of moderators in influencing the effectiveness of

ABMT interventions deserves further attention. This meta-analysis highlights certain moderators, such as the type of threat stimuli and intervention duration, but the complex interplay of these factors requires more in-depth investigation to determine their precise impact on treatment outcomes. The availability of internet facilities could be considered an obstacle when contemplating smartphone-delivered ABMT, as it might limit access for certain patients. Last, most of the studies focused on symptoms of mental health conditions and not a formal diagnosis of mental health conditions; further research is needed to validate the use of smartphone-delivered ABMT in this patient collective.

Future Directions

The findings of this meta-analysis point toward several promising avenues for the future of ABMT research. First, there is a need for the further exploration of design styles in ABMT interventions, with a particular focus on creating engaging and gamified programs that enhance user engagement and motivation. Second, the personalization of ABMT based on individual characteristics and preferences holds significant potential, enabling tailored interventions that match specific symptom profiles and cognitive processes. The standardization of ABMT protocols, including stimulus types, array formats, trial numbers, and intervention durations, is crucial to address the heterogeneity among study samples. Long-term follow-up studies are essential to assess the durability of ABMT's effects and its potential for preventing symptom relapse. Overall, the future of ABMT research should prioritize enhancing design styles, embracing personalization, addressing heterogeneity, and investigating long-term effects to maximize the effectiveness of ABMT in reducing mental health problems.

Conclusions

This systematic review and meta-analysis have shed light on the effectiveness of ABMT in addressing mental health symptoms. The findings reveal that active ABMT shows promise in reducing attentional biases, with a moderate overall effect size. However, its impact on directly alleviating anxiety and depression symptoms appears limited, as indicated by smaller and nonsignificant effect sizes within these subgroups. Interestingly, the placebo ABMT results emphasize the influence of belief and expectation in treatment outcomes, highlighting the importance of rigorous study designs to distinguish genuine effects from placebos. Moreover, moderator variables, such as the choice of threat stimuli, design style, and stimulus array type, emerge as critical factors influencing treatment efficacy, underscoring the need for personalized interventions. These findings provide valuable insights for tailoring and optimizing ABMT interventions for individuals with mental health problems.

Authors' Contributions

This study was conducted collaboratively by BB, MO, HB, YM, RO, and SM. BB, MO, and HB played pivotal roles in the study section, specifically in filtering and data extraction. YM investigated the risk of bias of the studies selected with BB. SM provided essential verification of the information and analysis methods used, while RO offered invaluable guidance in presenting the study. SM and RO made significant contributions, conducted a meticulous review, and approved the manuscript for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analyses) guidelines.
[DOCX File , 35 KB - mental_v11i1e56326_app1.docx]

Multimedia Appendix 2
Risk of bias.
[DOCX File , 22 KB - mental_v11i1e56326_app2.docx]

Multimedia Appendix 3
Sensitivity analysis.
[DOCX File , 27 KB - mental_v11i1e56326_app3.docx]

Multimedia Appendix 4
Funnel plots.
[DOCX File , 55 KB - mental_v11i1e56326_app4.docx]

Multimedia Appendix 5
Moderator analysis.
[DOCX File , 28 KB - mental_v11i1e56326_app5.docx]

Multimedia Appendix 6
Raw data.
[DOCX File , 43 KB - mental_v11i1e56326_app6.docx]

References

1. Myruski S, Cho H, Bikson M, Dennis-Tiwary TA. Transcranial direct current stimulation (tDCS) augments the effects of gamified, mobile attention bias modification. Front Neuroergon 2021 May 19;2:652162 [FREE Full text] [doi: 10.3389/fnrgo.2021.652162] [Medline: 38235222]
2. Firth J, Torous J, Carney R, Newby J, Cosco TD, Christensen H, et al. Digital technologies in the treatment of anxiety: recent innovations and future directions. Curr Psychiatry Rep 2018 May 19;20(6):44 [FREE Full text] [doi: 10.1007/s11920-018-0910-2] [Medline: 29779065]
3. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022 Feb;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
4. Moroz N, Moroz I, D'Angelo MS. Mental health services in Canada: barriers and cost-effective solutions to increase access. Healthc Manage Forum 2020 Nov;33(6):282-287. [doi: 10.1177/0840470420933911] [Medline: 32613867]
5. Di Carlo F, Sociali A, Picutti E, Pettorruso M, Vellante F, Verrastro V, et al. Telepsychiatry and other cutting-edge technologies in COVID-19 pandemic: bridging the distance in mental health assistance. Int J Clin Pract 2021 Jan;75(1):10 [FREE Full text] [doi: 10.1111/ijcp.13716] [Medline: 32946641]
6. Ye J. Advancing mental health and psychological support for health care workers using digital technologies and platforms. JMIR Form Res 2021 Jun 30;5(6):e22075 [FREE Full text] [doi: 10.2196/22075] [Medline: 34106874]
7. Subu MA, Wati DF, Netrida N, Priscilla V, Dias JM, Abraham MS, et al. Types of stigma experienced by patients with mental illness and mental health nurses in Indonesia: a qualitative content analysis. Int J Ment Health Syst 2021 Oct 18;15(1):77 [FREE Full text] [doi: 10.1186/s13033-021-00502-x] [Medline: 34663399]
8. Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV). Washington, DC: American Psychiatric Association; 1994.
9. Lundkvist-Houndoumadi I, Thastum M. Anxious children and adolescents non-responding to CBT: clinical predictors and families' experiences of therapy. Clin Psychol Psychother 2017 Jan;24(1):82-93. [doi: 10.1002/cpp.1982] [Medline: 26514088]
10. James AC, Reardon T, Soler A, James G, Creswell C. Cognitive behavioural therapy for anxiety disorders in children and adolescents. Cochrane Database Syst Rev 2020 Nov 16;11(11):CD013162 [FREE Full text] [doi: 10.1002/14651858.CD013162.pub2] [Medline: 33196111]
11. Hollon SD, Beck AT. Cognitive and cognitive-behavioral therapies. In: Lambert MJ, editor. Bergin and Garfield's Handbook of Psychotherapy and Behavior Change. Hoboken, NJ: John Wiley & Sons; 2013:393-442.
12. Martell CR, Kanter JW. Behavioral activation in the context of "third wave" therapies. In: Herbert JD, Forman EM, editors. Acceptance and Mindfulness in Cognitive Behavior Therapy: Understanding and Applying the New Therapies. Hoboken, NJ: John Wiley & Sons; 2011.
13. Wieser MJ, Keil A. Attentional threat biases and their role in anxiety: a neurophysiological perspective. Int J Psychophysiol 2020 Jul;153:148-158 [FREE Full text] [doi: 10.1016/j.ijpsycho.2020.05.004] [Medline: 32428525]
14. Bar-Haim Y. Research review: attention bias modification (ABM): a novel treatment for anxiety disorders. J Child Psychol Psychiatry 2010 Aug;51(8):859-870. [doi: 10.1111/j.1469-7610.2010.02251.x] [Medline: 20456540]
15. MacLeod C, Clarke PJ. The attentional bias modification approach to anxiety intervention. Clin Psychol Sci 2015 Jan 06;3(1):58-78. [doi: 10.1177/2167702614560749]
16. Price RB, Beltz AM, Woody ML, Cummings L, Gilchrist D, Siegle GJ. Neural connectivity subtypes predict discrete attentional bias profiles among heterogeneous anxiety patients. Clin Psychol Sci 2020 May 01;8(3):491-505 [FREE Full text] [doi: 10.1177/2167702620906149] [Medline: 33758682]
17. Halder S, Mahato AK. Cognitive behavior therapy for children and adolescents: challenges and gaps in practice. Indian J Psychol Med 2019;41(3):279-283 [FREE Full text] [doi: 10.4103/IJPSYM.IJPSYM_470_18] [Medline: 31142932]
18. Sawyer MC, Nunez DE. Cognitive-behavioral therapy for anxious children: from evidence to practice. Worldviews Evid Based Nurs 2014 Feb;11(1):65-71. [doi: 10.1111/wvn.12024] [Medline: 24479740]
19. Haim YB. Increasing psychological health and performance in soldiers applying advanced eye-tracking-based attention bias modification. Tel Aviv University. 2020 Oct. URL: https://apps.dtic.mil/sti/trecms/pdf/AD1125511.pdf [accessed 2024-08-04]
20. Bar-Haim Y, Lamy D, Pergamin L, Bakermans-Kranenburg MJ, van IJzendoorn MH. Threat-related attentional bias in anxious and nonanxious individuals: a meta-analytic study. Psychol Bull 2007 Jan;133(1):1-24. [doi: 10.1037/0033-2909.133.1.1] [Medline: 17201568]
21. Beevers CG, Hsu KJ, Schnyer DM, Smits JA, Shumake J. Change in negative attention bias mediates the association between attention bias modification training and depression symptom improvement. J Consult Clin Psychol 2021 Oct;89(10):816-829 [FREE Full text] [doi: 10.1037/ccp0000683] [Medline: 34807657]

22. Hsu KJ, Shumake J, Caffey K, Risom S, Labrada J, Smits JA, et al. Efficacy of attention bias modification training for depressed adults: a randomized clinical trial. Psychol Med 2021 Mar 26;52(16):3865-3873. [doi: 10.1017/s0033291721000702]
23. Charvet L, George A, Cho H, Krupp LB, Dennis-Tiwary TA. Mobile attention bias modification training is a digital health solution for managing distress in multiple sclerosis: a pilot study in pediatric onset. Front Neurol 2021 Jul 28;12:719090 [FREE Full text] [doi: 10.3389/fneur.2021.719090] [Medline: 34393986]
24. Dennis-Tiwary TA, Denefrio S, Gelber S. Salutary effects of an attention bias modification mobile application on biobehavioral measures of stress and anxiety during pregnancy. Biol Psychol 2017 Jul;127:148-156 [FREE Full text] [doi: 10.1016/j.biopsycho.2017.05.003] [Medline: 28478138]
25. Wang K, Varma DS, Prosperi M. A systematic review of the effectiveness of mobile apps for monitoring and management of mental health symptoms or disorders. J Psychiatr Res 2018 Dec;107:73-78. [doi: 10.1016/j.jpsychires.2018.10.006] [Medline: 30347316]
26. Vermeir JF, White MJ, Johnson D, Crombez G, Van Ryckeghem DM. Gamified web-delivered attentional bias modification training for adults with chronic pain: protocol for a randomized, double-blind, placebo-controlled trial. JMIR Res Protoc 2022 Jan 27;11(1):e32359 [FREE Full text] [doi: 10.2196/32359] [Medline: 35084344]
27. Silang KA, Sohal PR, Bright KS, Leason J, Roos L, Lebel C, et al. eHealth interventions for treatment and prevention of depression, anxiety, and insomnia during pregnancy: systematic review and meta-analysis. JMIR Ment Health 2022 Feb 21;9(2):e31116 [FREE Full text] [doi: 10.2196/31116] [Medline: 35188471]
28. Qin M, Chen B, Sun S, Liu X. Effect of mobile phone app-based interventions on quality of life and psychological symptoms among adult cancer survivors: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2022 Dec 19;24(12):e39799 [FREE Full text] [doi: 10.2196/39799] [Medline: 36534460]
29. Serrano-Ripoll MJ, Zamanillo-Campos R, Fiol-DeRoque MA, Castro A, Ricci-Cabello I. Impact of smartphone app-based psychological interventions for reducing depressive symptoms in people with depression: systematic literature review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2022 Jan 27;10(1):e29621 [FREE Full text] [doi: 10.2196/29621] [Medline: 35084346]
30. Cumpston M, Li T, Page MJ, Chandler J, Welch VA, Higgins JP, et al. Updated guidance for trusted systematic reviews: a new edition of the Cochrane Handbook for Systematic Reviews of Interventions. Cochrane Database Syst Rev 2019 Oct 03;10(10):ED000142 [FREE Full text] [doi: 10.1002/14651858.ED000142] [Medline: 31643080]
31. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA Statement. Open Med 2009;3(3):e123-e130 [FREE Full text] [Medline: 21603045]
32. Zhang M, Ying J, Song G, Fung DS, Smith H. Attention and cognitive bias modification apps: review of the literature and of commercially available apps. JMIR Mhealth Uhealth 2018 May 24;6(5):e10034 [FREE Full text] [doi: 10.2196/10034] [Medline: 29793899]
33. Zhang M, Ying J, Song G, Fung DS, Smith H. Gamified cognitive bias modification interventions for psychiatric disorders: review. JMIR Ment Health 2018 Oct 25;5(4):e11640 [FREE Full text] [doi: 10.2196/11640] [Medline: 30361194]
34. Johnson N, Phillips M. Rayyan for systematic reviews. J Electron Resour Librariansh 2018 May 02;30(1):46-48. [doi: 10.1080/1941126x.2018.1444339]
35. Higgins JP, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011 Oct 18;343:d5928 [FREE Full text] [doi: 10.1136/bmj.d5928] [Medline: 22008217]
36. Racine JS. RStudio: a platform-independent IDE for R and Sweave. J Appl Econ 2012 Jan;27(1):167-172. [doi: 10.2307/41337225]
37. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. Stat Med 2001 Mar 30;20(6):825-840. [doi: 10.1002/sim.650] [Medline: 11252006]
38. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Cambridge, MA: Academic Press; 2013.
39. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634 [FREE Full text] [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]
40. Mathur MB, VanderWeele TJ. Sensitivity analysis for publication bias in meta-analyses. J R Stat Soc Ser C Appl Stat 2020 Nov;69(5):1091-1119 [FREE Full text] [doi: 10.1111/rssc.12440] [Medline: 33132447]
41. Dennis TA, O'Toole L. Mental health on the go: effects of a gamified attention bias modification mobile application in trait anxious adults. Clin Psychol Sci 2014 Sep 01;2(5):576-590 [FREE Full text] [doi: 10.1177/2167702614522228] [Medline: 26029490]
42. Enock PM, Hofmann SG, McNally RJ. Attention bias modification training via smartphone to reduce social anxiety: a randomized, controlled multi-session experiment. Cogn Ther Res 2014 Mar 4;38(2):200-216. [doi: 10.1007/s10608-014-9606-z]
43. Dennis-Tiwary TA, Egan LJ, Babkirk S, Denefrio S. For whom the bell tolls: neurocognitive individual differences in the acute stress-reduction effects of an attention bias modification game for anxiety. Behav Res Ther 2016 Feb;77:105-117 [FREE Full text] [doi: 10.1016/j.brat.2015.12.008] [Medline: 26745621]

44. Yang R, Cui L, Li F, Xiao J, Zhang Q, Oei TP. Effects of cognitive bias modification training via smartphones. Front Psychol 2017 Aug 14;8:1370 [FREE Full text] [doi: 10.3389/fpsyg.2017.01370] [Medline: 28855880]
45. Teng MH, Hou YM, Chang SH, Cheng HJ. Home-delivered attention bias modification training via smartphone to improve attention control in sub-clinical generalized anxiety disorder: a randomized, controlled multi-session experiment. J Affect Disord 2019 Mar 01;246:444-451. [doi: 10.1016/j.jad.2018.12.118] [Medline: 30599367]
46. Flaudias V, Zerhouni O, Chakroun-Baggioni N, Pires S, Schmitt A, de Chazeron I, et al. Reducing attentional bias in individuals with alcohol use disorders with a tablet application: a randomized controlled trial pilot study. Alcohol Alcohol 2020 Feb 07;55(1):51-55. [doi: 10.1093/alcalc/agz080] [Medline: 31724703]
47. Niles AN, Woolley JD, Tripp P, Pesquita A, Vinogradov S, Neylan TC, et al. Randomized controlled trial testing mobile-based attention-bias modification for posttraumatic stress using personalized word stimuli. Clin Psychol Sci 2020 Jul;8(4):756-772 [FREE Full text] [doi: 10.1177/2167702620902119] [Medline: 34414018]
48. Chelliah A, Robinson O. Efficacy of attention bias modification via smartphones in a large population sample. R Soc Open Sci 2022 Aug 10;9(8):211629 [FREE Full text] [doi: 10.1098/rsos.211629] [Medline: 35958083]
49. Flaudias V, Zerhouni O, Chakroun-Baggioni N, de Chazeron I, Llorca PM, Brousse G. Evaluation of a smartphone application on the reduction of attentional bias toward alcohol among students. Front Psychol 2022 Feb 10;13:790030 [FREE Full text] [doi: 10.3389/fpsyg.2022.790030] [Medline: 35222189]
50. Robinson JD, Cui Y, Linares Abrego P, Engelmann JM, Prokhorov AV, Vidrine DJ, et al. Sustained reduction of attentional bias to smoking cues by smartphone-delivered attentional bias modification training for smokers. Psychol Addict Behav 2022 Nov;36(7):906-919 [FREE Full text] [doi: 10.1037/adb0000805] [Medline: 35025555]
51. Hakamata Y, Lissek S, Bar-Haim Y, Britton JC, Fox NA, Leibenluft E, et al. Attention bias modification treatment: a meta-analysis toward the establishment of novel treatment for anxiety. Biol Psychiatry 2010 Dec 01;68(11):982-990 [FREE Full text] [doi: 10.1016/j.biopsych.2010.07.021] [Medline: 20887977]
52. Heeren A, Mogoaşe C, McNally RJ, Schmitz A, Philippot P. Does attention bias modification improve attentional control? A double-blind randomized experiment with individuals with social anxiety disorder. J Anxiety Disord 2015 Jan;29:35-42 [FREE Full text] [doi: 10.1016/j.janxdis.2014.10.007] [Medline: 25465885]
53. Hang Y, Zhang G, Wang C, Zhang N, Liu G. Attention bias modification for anxiety disorders in children and adolescents: a systematic review and meta-analysis. Psychiatry Res 2021 Jun;300:113896. [doi: 10.1016/j.psychres.2021.113896] [Medline: 33799199]
54. Hang Y, Xu L, Wang C, Zhang G, Zhang N. Can attention bias modification augment the effect of CBT for anxiety disorders? A systematic review and meta-analysis. Psychiatry Res 2021 May;299:113892. [doi: 10.1016/j.psychres.2021.113892] [Medline: 33799125]
55. Lazarov A, Suarez-Jimenez B, Abend R, Naim R, Shvil E, Helpman L, et al. Bias-contingent attention bias modification and attention control training in treatment of PTSD: a randomized control trial. Psychol Med 2019 Oct;49(14):2432-2440 [FREE Full text] [doi: 10.1017/S0033291718003367] [Medline: 30415648]
56. Mogg K, Waters AM, Bradley BP. Attention bias modification (ABM): review of effects of multisession ABM training on anxiety and threat-related attention in high-anxious individuals. Clin Psychol Sci 2017 Jul;5(4):698-717 [FREE Full text] [doi: 10.1177/2167702617696359] [Medline: 28752017]
57. Möbius M, Ferrari GR, van den Bergh R, Becker ES, Rinck M. Eye-tracking based attention bias modification (ET-ABM) facilitates disengagement from negative stimuli in dysphoric individuals. Cognit Ther Res 2018;42(4):408-420 [FREE Full text] [doi: 10.1007/s10608-018-9889-6] [Medline: 29962560]
58. Amir N, Beard C, Burns M, Bomyea J. Attention modification program in individuals with generalized anxiety disorder. J Abnorm Psychol 2009 Feb;118(1):28-33 [FREE Full text] [doi: 10.1037/a0012589] [Medline: 19222311]
59. Bar-Haim Y, Kerem A, Lamy D, Zakay D. When time slows down: the influence of threat on time perception in anxiety. Cogn Emot 2010;24(2):255-263. [doi: 10.1080/02699930903387603]
60. Price DD, Finniss DG, Benedetti F. A comprehensive review of the placebo effect: recent advances and current thought. Annu Rev Psychol 2008;59:565-590. [doi: 10.1146/annurev.psych.59.113006.095941] [Medline: 17550344]
61. Xia HS, Li YX, Zhang QY, Zhong DL, Liu XB, Gou XY, et al. Attention bias modification for depression: a systematic review and meta-analysis. Front Psychiatry 2023 Mar 10;14:1098610 [FREE Full text] [doi: 10.3389/fpsyt.2023.1098610] [Medline: 36970284]
62. Tong HL, Quiroz JC, Kocaballi AB, Fat SC, Dao KP, Gehringer H, et al. Personalized mobile technologies for lifestyle behavior change: a systematic review, meta-analysis, and meta-regression. Prev Med 2021 Jul;148:106532. [doi: 10.1016/j.ypmed.2021.106532] [Medline: 33774008]
63. Charles M, Wellington CE, Mokrysz C, Freeman TP, O'Ryan D, Curran HV. Attentional bias and treatment adherence in substitute-prescribed opiate users. Addict Behav 2015 Jul;46:100-105 [FREE Full text] [doi: 10.1016/j.addbeh.2015.03.017] [Medline: 25838001]
64. Zhang M, Heng S, Song G, Fung DS, Smith HE. Co-designing a mobile gamified attention bias modification intervention for substance use disorders: participatory research study. JMIR Mhealth Uhealth 2019 Oct 03;7(10):e15871 [FREE Full text] [doi: 10.2196/15871] [Medline: 31584003]

65. Zhang M, Fung DS, Smith H. A literature review of attentional biases amongst individuals with substance dependency: individual differences and modulating factors. Psychiatry Int 2020 Dec 04;1(2):125-134. [doi: 10.3390/psychiatryint1020013]

Abbreviations

ABMT: attentional bias modification training
CBT: cognitive behavioral therapy
PRISMA: Preferred Reporting Items for Systematic Review and Meta-Analyses
PTSD: posttraumatic stress disorder

Edited by J Torous; submitted 12.01.24; peer-reviewed by K Cho, WK Hou; comments to author 25.04.24; revised version received 07.06.24; accepted 05.07.24; published 02.09.24.

Please cite as:
Banire B, Orr M, Burns H, McGowan Y, Orji R, Meier S
Smartphone-Delivered Attentional Bias Modification Training for Mental Health: Systematic Review and Meta-Analysis
JMIR Ment Health 2024;11:e56326
URL: https://mental.jmir.org/2024/1/e56326
doi:10.2196/56326
PMID:39222349

Review

Empathic Conversational Agent Platform Designs and Their Evaluation in the Context of Mental Health: Systematic Review

Ruvini Sanjeewa[1], BSc; Ravi Iyer[1], PhD; Pragalathan Apputhurai[1], PhD; Nilmini Wickramasinghe[2], PhD; Denny Meyer[1], PhD

[1]School of Health Sciences, Swinburne University of Technology, Hawthorn, Australia

[2]School of Computing, Engineering and Mathematical Sciences, La Trobe University, Bundoora, Australia

Corresponding Author:
Ruvini Sanjeewa, BSc
School of Health Sciences
Swinburne University of Technology
PO Box 218
John Street
Hawthorn, 3122
Australia
Phone: 61 422587030
Email: rsanjeewa@swin.edu.au

Abstract

Background: The demand for mental health (MH) services in the community continues to exceed supply. At the same time, technological developments make the use of artificial intelligence–empowered conversational agents (CAs) a real possibility to help fill this gap.

Objective: The objective of this review was to identify existing empathic CA design architectures within the MH care sector and to assess their technical performance in detecting and responding to user emotions in terms of classification accuracy. In addition, the approaches used to evaluate empathic CAs within the MH care sector in terms of their acceptability to users were considered. Finally, this review aimed to identify limitations and future directions for empathic CAs in MH care.

Methods: A systematic literature search was conducted across 6 academic databases to identify journal articles and conference proceedings using search terms covering 3 topics: "conversational agents," "mental health," and "empathy." Only studies discussing CA interventions for the MH care domain were eligible for this review, with both textual and vocal characteristics considered as possible data inputs. Quality was assessed using appropriate risk of bias and quality tools.

Results: A total of 19 articles met all inclusion criteria. Most (12/19, 63%) of these empathic CA designs in MH care were machine learning (ML) based, with 26% (5/19) hybrid engines and 11% (2/19) rule-based systems. Among the ML-based CAs, 47% (9/19) used neural networks, with transformer-based architectures being well represented (7/19, 37%). The remaining 16% (3/19) of the ML models were unspecified. Technical assessments of these CAs focused on response accuracies and their ability to recognize, predict, and classify user emotions. While single-engine CAs demonstrated good accuracy, the hybrid engines achieved higher accuracy and provided more nuanced responses. Of the 19 studies, human evaluations were conducted in 16 (84%), with only 5 (26%) focusing directly on the CA's empathic features. All these papers used self-reports for measuring empathy, including single or multiple (scale) ratings or qualitative feedback from in-depth interviews. Only 1 (5%) paper included evaluations by both CA users and experts, adding more value to the process.

Conclusions: The integration of CA design and its evaluation is crucial to produce empathic CAs. Future studies should focus on using a clear definition of empathy and standardized scales for empathy measurement, ideally including expert assessment. In addition, the diversity in measures used for technical assessment and evaluation poses a challenge for comparing CA performances, which future research should also address. However, CAs with good technical and empathic performance are already available to users of MH care services, showing promise for new applications, such as helpline services.

(JMIR Ment Health 2024;11:e58974) doi:10.2196/58974

KEYWORDS

conversational agents; chatbots; virtual assistants; empathy; emotionally aware; mental health; mental well-being

Introduction

Background

An escalation in mental health (MH) diagnoses in the community, inadequate facilities, and a MH care workforce that does not meet demand are placing extraordinary pressures on an already strained system [1]. This service gap creates a significant opportunity for MH care interventions, enhanced using recent advances in modern technologies. Conversational agent (CA) platforms using artificial intelligence (AI) via machine learning (ML) techniques have emerged within the MH care domain, providing additional functionalities and support to address this gap [2]. Examples of CAs that use ML include Woebot, providing cognitive behavioral therapy [3]; Wysa, providing MH support by checking depressive symptoms [4]; Saarthi, trained to provide personalized and empathic support to patients via therapeutic techniques [5]; and Empathetic Research IoT Network, a chatbot that provides access to MH resources for students in need [6]. However, the lack of acceptance of CAs in the MH domain remains a barrier to the uptake of these innovations, and the lack of empathy often displayed by CAs contributes to end-user mistrust [7].

Empathy in patient care has been defined by the World Health Organization as an understanding of the patient's experiences, concerns, and perspectives, combined with a capacity to communicate this understanding and an intention to help [8]. Counselor empathy is an essential feature that enhances therapeutic outcomes for patients and can be measured via therapeutic alliance [9,10]. The same is true for CA-human interactions, where empathy exhibited by a CA system helps build rapport, encouraging users to more frequently engage with the CA system [11]. Contextual awareness, which allows CAs to respond to a user's current emotional situation when suggesting appropriate interventions, also facilitates empathic CA communication [12]. Both trustworthiness of the CA (as perceived by the user) and contextual awareness of the user's situation (as detected by the CA) are, therefore, important considerations when building an empathic CA. Empathy serves to enhance the bidirectional interaction between the CA and the end user [13].

Assessment of the effectiveness of CA platforms has received little attention in the MH care sector [14]. For the impact of these systems to be fully realized, these platforms need to meet the requirements of end users, which suggests a key role for lived experience and coproduction. The validity and reliability of these new digital technologies also need to be reviewed by MH care decision-makers and professionals to ensure successful integration in the sector [15]. In addition, evaluations need to assess the ability of such platforms to reduce symptoms of mental illness [16] while also enhancing user well-being and ensuring that patients feel understood [13]. However, any such evaluation needs to be conducted in the context of the role envisaged for the CA, considering the success of the bidirectional interaction described earlier.

While there are existing reviews exploring the efficacy of CAs designed for MH care [10,17,18], to our knowledge, this is the first review to specifically examine how these empathic CAs are designed and evaluated. A comprehensive systematic review and meta-analysis of AI-based CAs for promoting MH was conducted by Li et al [17], with a focus on the intervention and technical characteristics of effective CAs. The effectiveness of the CA designs was captured through user feedback. The meta-analysis explored the role of the CA, AI techniques, and delivery platforms that contributed to the success of these designs. In a similar review, Gaffney et al [18] targeted CA interventions for treating MH problems, with a specific focus on user experience outcomes as measures of efficacy. Another such study explored the evidence of effectiveness with regard to improving symptoms of MH conditions [19]. A critical finding of this review was that empathic response and personalization were significant facilitators of efficacy in these systems. However, the incorporation of this crucial empathy component within CAs has not been studied in any depth within the MH sector. Existing reviews have tended to focus on the inability of CAs to respond to unexpected user inputs rather than their ability to demonstrate empathy [19].

Objectives

This review aimed to assess the types of CA designs found in the MH care sector that are specifically tailored to convey empathy. It also aimed to describe the methods used to evaluate these empathic designs from a technical and implementation perspective. Therefore, this review considered how empathy has been engineered and the limitations identified with its use by a CA from a human perspective. There were three objectives: (1) to identify existing empathic CA design architectures within the MH care sector and to assess their technical performance in detecting and responding to user emotions appropriately; (2) to describe the approaches used to evaluate empathic CAs within the MH care sector in terms of their acceptability to users; and (3) to identify limitations and future directions for empathic CAs in MH care.

Methods

Database Search

A systematic literature search was conducted across 6 academic databases (Web of Science; Scopus; EBSCOhost: Academic Search Complete; CINAHL Complete; Computers and Applied Sciences Complete; and IEEE Xplore) for journal articles and conference proceedings from January 1, 2010, to September 30, 2023. The period of data capture dates from the time when AI-informed CA technology emerged as a distinct area of research [20], and conference proceedings were included to ensure that the most recent studies could be included.

The search terms covered 3 topics: "conversational agents," "mental health," and "empathy." Possible keywords were broadened using synonyms for each topic, pilot searching of existing literature, and discussion among research team members. Boolean operators combined different keywords and their synonyms to establish the final search strategy. Wildcards were included (eg, empath* = empathic). Medical Subject Heading terms were used where appropriate. An example of the search syntax is available in Multimedia Appendix 1 [4-6,21-36].

Eligibility Criteria

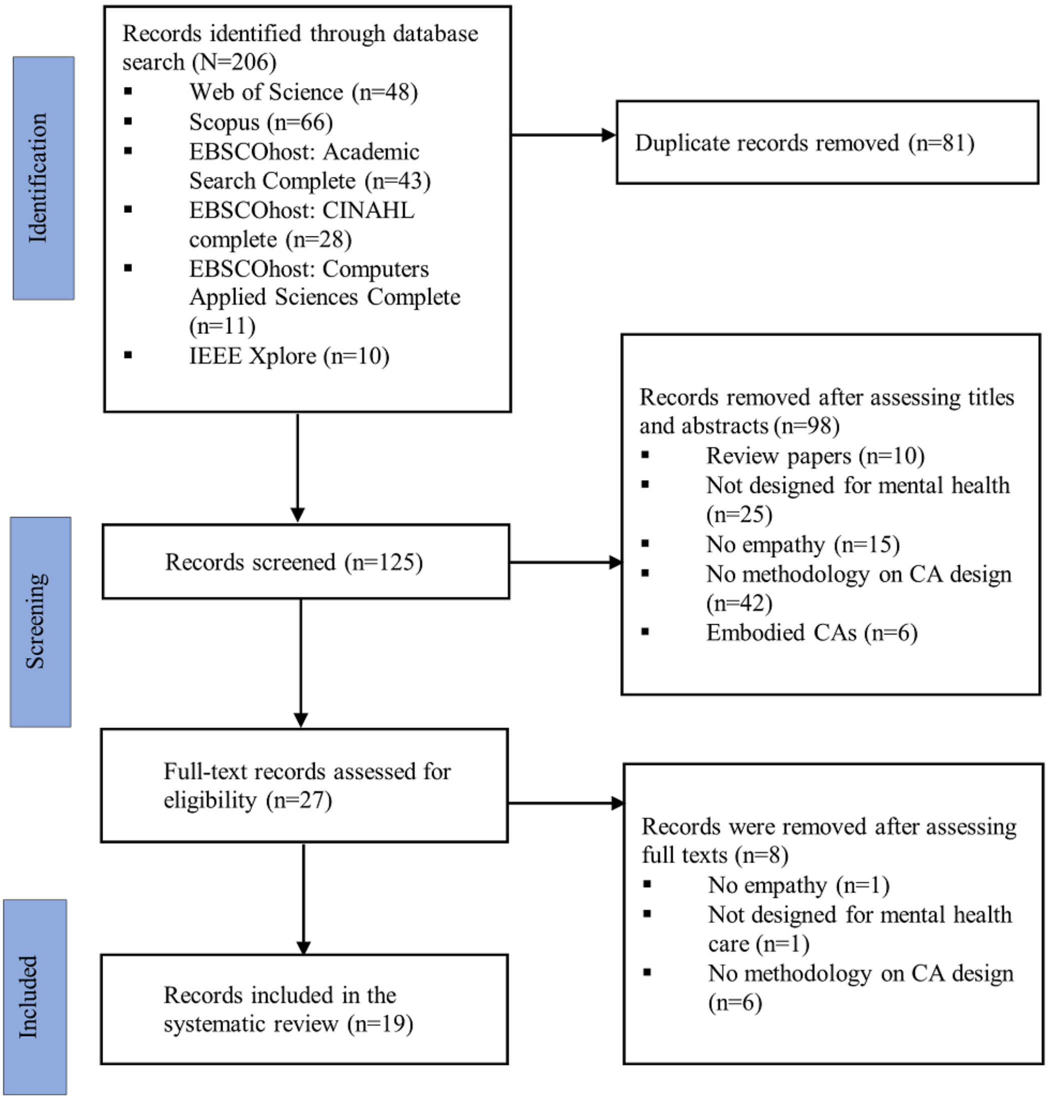
Publications discussing CA interventions for the MH care domain were eligible for the review. There were no restrictions on research design (eg, observational designs and narrative review). This review considered both textual and vocal modes of interaction with the CA. Publications were included if they referred to CA empathy or related terms (eg, emotional intelligence, emotional awareness, and compassion). Publications that did not feature a methodology section that detailed CA design, types of data sets, and participants were excluded. Systematic reviews, scoping reviews, and meta-analyses were excluded. Publications that used data inputs other than text and vocal cues (eg, facial recognition) were also excluded. Multimedia Appendix 1 provides the full-text screening checklist.

Screening

Eligible references were exported to the EndNote (version 20; Clarivate) software [37], where duplicates were removed. The first author (RS) conducted the title and abstract search, mapping against the eligibility criteria. A full-text screening was then performed by the first author and by 2 other authors, DM and RI, independently. Any disagreements on full-text screening were discussed, and an agreement was reached before proceeding. Figure 1 illustrates the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart describing the screening process. PRISMA checklist is reported in Multimedia Appendix 2.

Data including details on the study designs, how empathy was evaluated, and the types of CA architectures used were extracted to obtain a summary of all findings (Multimedia Appendix 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) procedure applied. CA: conversational agent.



Quality Assessment

The Joanna Briggs Institute critical appraisal tool was used to assess the methodological quality of the papers shortlisted while also considering the extent to which each study addressed the possibility of bias in design, conduct, and analysis [38]. This appraisal tool was specifically designed for the assessment of the variety of study designs encountered in this systematic review. Decisional criteria were answered with *yes*, *no*, *unclear*, *or not applicable*. The proportion of *yes* responses relative to the total number of assessment questions was used for quality assessment purposes. Separate quality assessments were conducted for publications that included a description of the implementation as well as the design of the CA platform and for publications that included only a description of the design.

Risk of Bias

Risk of bias was assessed using the revised Cochrane risk-of-bias tool for randomized trials This included risks of bias due to randomization, deviations from the intended intervention, missing data, the measurement of outcomes, and the selection of results. The risk of bias in nonrandomized studies of interventions tool was used to evaluate the nonrandomized studies.

Results

Overview

A total of 19 studies met all the inclusion criteria. The study characteristics are summarized in Table 1.

Table 1. Study characteristics.

Study	CA[a]	Training database	Aim of the study	Evaluation measures for detecting and responding to user emotions	Mode of exchange	Analysis model for generating empathic responses
Jiang et al [21], 2022	Replika	14 Chinese female users (aged 19-26 years)	Explore types of mediated empathy that occur in human-AI[b] interactions	In-depth interviews and survey results: user ratings of empathy	Text and voice	Transformer architecture
Brocki et al [22], 2023	Serena	Trained on "Pushshift" Reddit data set and tested on psychotherapy transcript	Help improve outcomes of counseling by lowering barriers to access	Survey results: user ratings of engagement and helpfulness	Text	Transformer architecture
Persons et al [6], 2021	ERIN[c]	15 undergraduate students	Help users with finding resources about sensitive issues	Survey results: user ratings for experience	Text	Rule-based architecture
Trappey et al [23], 2022	Virtual reality empathy-centric counseling CA	120 university students	Provide complementary support for students who were troubled	Survey results: user ratings of stress levels, life impact, and psychological sensitivity	Voice and text	Transformer architecture
Ghandeharioun et al [24], 2019	EMMA[d]	39 participants	Delivery of just-in-time MH[e] interventions	Survey results: user ratings of preference; behavioral metrics: user engagement	Text	Hybrid architecture
Meng and Dai [25], 2021	AI CA	278 participants from Midwestern University	Check whether the CA's emotional support was effective in reducing people's stress and worry	Survey results: user ratings of stress, worry, and perceived support	Text	Transformer architecture
Goel et al [26], 2021	Empathic CA with an attention mechanism	Trained with the Facebook AI Empathic Dialogue data set	Support users express their feelings and anxious thoughts	None	Text	Neural network architecture
Adikari et al [27], 2022	Empathic CA	Data set from Cancer Chat Canada	Provide empathic patient-centered MH care	Behavioral metrics for user engagement	Text	Hybrid architecture
Inkster et al [4], 2018	Wysa	129 users with self-reported symptoms of depression	Evaluation of the effectiveness and engagement levels of Wysa	Survey results for symptom assessment	Text	Unspecified ML[f] architecture
Beredo and Ong [28], 2022	Vhope	Senior high school and college students (aged 17-20 years)	Help the students maintain their well-being	Response ratings provided by experts	Text	Hybrid architecture
Rathnayaka et al [29], 2022	Bunji	Australian mobile users on Google Play Store	Remote health monitoring	Survey results for symptom and mood assessment	Text	Unspecified ML architecture
Morris et al [30], 2018	Koko	37,169 individuals who signed up for the Koko platform	A corpus-based approach to simulate expressed empathy	Response ratings provided by users	Text	Hybrid architecture
Ghandeharioun et al [31], 2019	A behavioral change CA	39 participants (n=7, 18% were female, and n=32, 82% were male)	Conduct experience sampling	Survey results: user ratings of likability and CA intelligence	Text	Rule-based architecture
Saha et al [32], 2022	Empathic CA	Data set: conversations between the support seekers who were depressed	Generate empathic and motivational responses	Response ratings by users for fluency, adaptability, and motivation	Text	Transformer architecture

Study	CA[a]	Training database	Aim of the study	Evaluation measures for detecting and responding to user emotions	Mode of exchange	Analysis model for generating empathic responses
Agnihotri et al [33], 2021	Topic-driven and affective CA	Data set: "ScenarioSA" with affective state labels	Tackle the emotional and contextual relevance for mental well-being	Response ratings for emotional relevance	Text	Transformer architecture
Rani et al [5], 2023	Saarthi	None	None	None	Text	Unspecified ML architecture
Alazraki et al [34], 2021	An empathic AI coach	23 participants recruited through crowd working websites	Achieve a high level of engagement during web-based therapy sessions	Survey results: user ratings of empathy and expert ratings of fluency	Text	Hybrid architecture
Gundavarapu et al [35], 2022	A CA companion	Data set: created using sources such as Wikipedia	Provide emotional support, without judgment	None	Text	Neural network architecture
Mishra et al [36], 2023	Counseling CA	A novel conversational data set	Provide MH and legal counseling	Survey results: user ratings of empathy	Text	Transformer architecture

[a]CA: conversational agent.

[b]AI: artificial intelligence.

[c]ERIN: Empathetic Research IoT Network.

[d]EMMA: Emotion-aware mHealth agent.

[e]MH: mental health.

[f]ML: machine learning.

Of the 19 studies, 6 (32%) were conducted in the United States and 6 (32%) in India. In addition, 1 (5%) study each from Australia, Canada, China, the Philippines, Poland, Switzerland, and the United Kingdom were also included. The year of publication is summarized in Multimedia Appendix 3, indicating a sharp rise in the number of publications since 2022. Most studies, 14 (74%) out of 19, described both design and human evaluations. The types of study designs among the 19 studies included are 9 (47%) cross-sectional studies, 5 (26%) randomized controlled trials (RCTs), 4 (21%) quasi-experimental designs, and 1 (5%) qualitative study. Only 5 (26%) of the 19 studies referred to an explicit definition of empathy, as summarized in Textbox 1.

Textbox 1. Definitions of empathy.

Studies and definition of empathy

- Jiang et al [21], 2022
  - Empathy processing is a situation-specific, cognitive-affective state or process with the projection of oneself into another's feelings, actions, and experiences.
- Trappey et al [23], 2022
  - Roger's [39] definition of empathy:
    - Level 1: responding to an individual's explicitly expressed meaning and feelings with a simple repetition of basic understanding.
    - Level 2: responding to the implicit, half-expressed, or implied feelings of the person with corresponding emotional words to acknowledge them and bring their true feelings to the surface.
    - Level 3: recognizing the individual's confusing and contradictory feelings that subconsciously obscure what they really care about, capturing the core of the emotion, and then responding to the patient's desire with affirmations.
    - Level 4: when the person is suppressing their feelings or not expressing their feelings in the conversation, guessing their intentions from what they are describing, capturing the core of the emotion, and responding to it directly or indirectly in a way that is acceptable to the person.
- Rathnayaka et al [29], 2022
  - Empathic engagement means, "making the impression of a credible and trustworthy conversation partner that can hear you out and offer a detached point of view on things."
- Saha et al [32], 2022
  - Empathy or empathic interactions refer to the ability to feel the emotions and experiences of others [40].
- Alazraki et al [34], 2021
  - Definition of empathy by Barrett-Lennard [41]:
    - First phase: where the listener sympathizes and resonates with what is being expressed by the speaker.
    - Second phase: where the listener compassionately responds to the speaker. Third phase: where the speaker assimilates the listener's response.

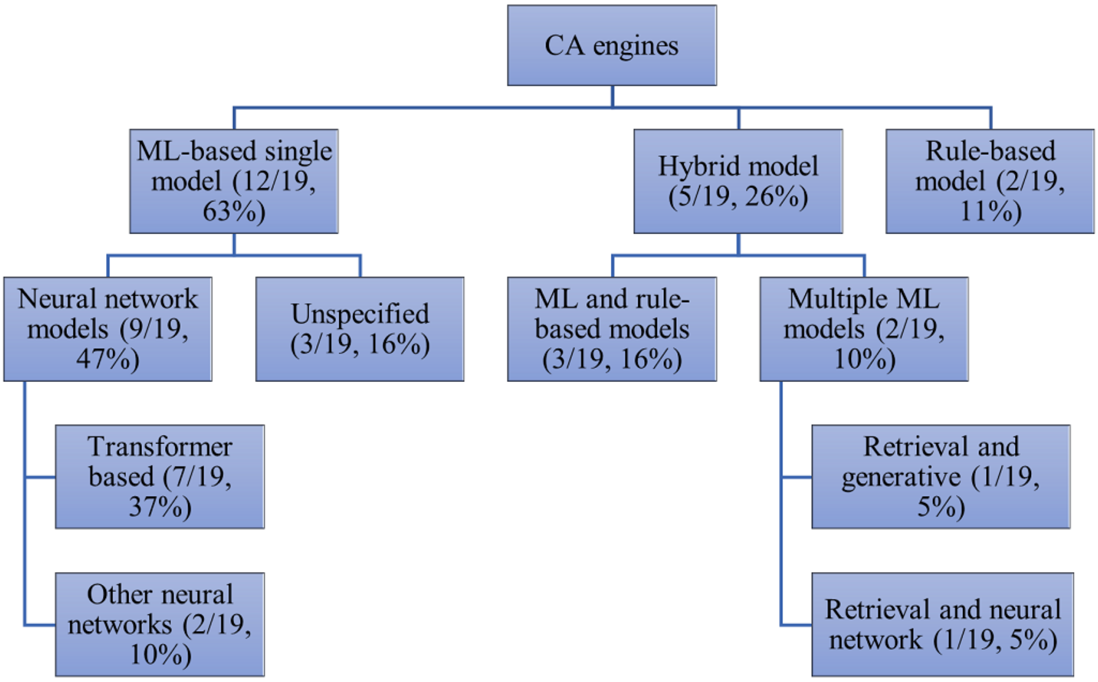
Keywords used to identify a CA varied across studies from "chatbot" (9/19, 47%) to "conversational agent" (6/19, 32%) to "dialog system" (2/19, 11%) to "virtual assistant" (1/19, 5%) to "conversational AI agent" (1/19, 5%). The mode of interaction chosen by most of the CA designs, 17 (89%) out of 19, was text (eg, live chat, symptom checker, and text-based counseling), with voice interactions being used in interactive avatar and counseling roles in 2 (11%) studies.

In the *Technical Design of the CAs* section, we consider the technical designs used for these CAs and their performance in detecting and responding to user emotions before discussing how human-user evaluations were conducted and the conclusions reached from these evaluations.

Technical Design of the CAs

The types of CA architectures (or engines) considered by the authors included a mix of recent technologies, as summarized in Figure 2, with ML-based architectures used in 12 (63%) out of 19 cases. The transformer-based engine, which learns meaning from context, was used in 7 of the 19 (37%) studies, sometimes in the form of a large language model (LLM). A minority of the papers, 3 (16%) out of 19, did not specify the type of engine used within the design. Hybrid or ensemble models use several models in parallel to improve the accuracy of the overall CA design. A more detailed breakdown of the CA engine types with explanations is shown in Multimedia Appendix 4. Figures S1 and S2 in Multimedia Appendix 4 also illustrate how a single engine and a hybrid engine work with user input to provide an empathic response.

Figure 2. Types of conversational agent (CA) architectures. ML: machine learning.



Transformer-based engines included Bidirectional Encoder Representations Transformer (BERT), Sentence-BERT, Robustly Optimized-BERT, Generative Pre-trained Transformer 2, and sequence-2-sequence models. Other neural network architecture–based CA designs were incorporated in 2 (11%) of the 19 papers [26,35].

Of the 19 publications, 5 (26%) considered hybrid models. Of these hybrid models, 2 applied a ML model to capture user emotion and then applied a rule-based algorithm to generate appropriate responses in dialogue management [24-27]. For example, EMMA gathered mobile sensor data to infer user mood and then assigned users to appropriate wellness interventions [24]. Once assigned, the CA then responded with emotionally expressive responses selected at random from a set of prescripted phrases using a rule-based approach [27]. In another example, VHope, an internet-based therapist, used a hybrid model containing a retrieval model that deciphered user input combined with a generative model to elicit empathic responses [28].

Among the 19 papers, the 3 (16%) papers using unspecified architectures commenced with natural language processing (NLP) before using various ML approaches. In one example, continuous emotional support via remote MH care monitoring and personalized assistance was provided [29]. MH monitoring was performed by scheduling activities that were meaningful to each user, sending out reminders as encouragement, and forwarding satisfaction surveys to receive feedback.

Overall, 2 (11%) of the 19 publications implemented CA design approaches based on rule-based NLP architectures. For example, a mobile phone–based CA measured the level of emotion in user input and then selected an appropriate empathic response from a set of predefined scripts using a rule-based decision tree [31]. In the *Summary of the Results of the Assessment of the Technical Design of CAs in Terms of Classification Accuracy* section, we will discuss the technical performance of the CAs reviewed.

Summary of the Results of the Assessment of the Technical Design of CAs in Terms of Classification Accuracy

The accuracy of the designs in detecting and responding to user emotions appropriately is summarized in Table 2. Technical evaluations of the CA designs usually involved comparisons with a “gold standard,” using data not previously used for training the CA.

Table 2. Measures used for evaluating the technical performance of CA[a] designs.

Type of CA assessment	Assessment of user emotions or CA responses	Accuracy measure
Classification of sentiment and issues	User emotions	• Mathews correlation coefficient=0.857 [23]
Classification of valence and arousal	User emotions	• Accuracy of valence=80.4% [24] • Accuracy of arousal=50.4% [24]
Classification of recommended resources (for patients)	User emotions	• F1-score=0.87 [27]
Classification of objections during conversations	User emotions	• Accuracy=99.2% [4] • Specificity=99.7% [4] • Precision=74.7% [4] • Recall=62.1% [4]
Performance of the topic classifier	User emotions	• Accuracy=95% [33] • Precision=0.954 [33] • Recall=0.947 [33] • F1-score=0.95 [33]
Classification for empathy function	User emotions	• Accuracy=80.18% [34] • F1-score=80.66% [34] • W-ACC[b]=0.977 [36] • Macro F1-score=0.972 [36]
Prediction of valence and arousal	User emotions	• Accuracy of valence=82.2% [24] • Accuracy of arousal=65.7% [24]

Accuracy of the response generation	CA responses	• BLEU[c] score=0.126 [26]
		• BLEU-1 score (focused on a single word)=0.161 [32]
		• Perplexity score=50.90 [32]
		• ROUGE-L[d] score=0.124 [32]
		• Embedding-based metrics:
		  • Average=0.733 [32]
		  • Extrema=0.377 [32]
		  • Greedy=0.478 [32]

Emotion prediction	User emotions	• Accuracy
		  • Correctly predict the next emotion as positive or negative=79% [27]
		  • Proportion of correct emotion out of all emotions predicted=63% [27]

Performance of the language model	CA responses	• Perplexity score=9.977 [28] • Perplexity score=1.91 [23] • Response length=18.71 [23]
Emotion recognition	User emotions	• Accuracy=94.96% [34] • F1-score=95.10% [34]

[a]CA: conversational agent.

[b]W-ACC: weighted accuracy.

[c]BLEU: Bilingual Evaluation Understudy.

[d]ROUGE-L: Recall Oriented Understudy for Gisting Evaluation–Longest Common Sequence.

A technical evaluation of empathic CA performance was conducted in 17 (89%) of the 19 papers reviewed; however, only 10 (53%) papers reported these results. These studies conducted comprehensive assessments where technical performance was measured in terms of recognition, classification, prediction, and response generation abilities during interactions with end users. The assessments were centered around the ability of the CA to discern user emotions correctly and to respond appropriately. Of the 19 papers, 4 (21%) focused on the CA responses during the technical assessments, while the rest of the studies (n=15, 79%) considered user emotions. A variety of measures were used for each such assessment, highlighting the diversity in evaluation methodologies across studies. These metrics are categorized in detail under the type of CA performance in Multimedia Appendix 5 [4-6,23-34,36].

In general, the performances of the CA designs were satisfactory. The highest classification accuracy for user emotions was reported by ML-based CAs. In one of these studies, a Robustly Optimized-BERT transformer model, which was built integrating 3 classifiers for politeness, counseling strategy, and empathic feedback, achieved good results overall. This empathy classifier achieved excellent performance with a weighted accuracy score of 0.977 and an F_1-score of 0.972 [36]. In a second study, a topic-driven classification model used a Generative Pre-trained Transformer 2 model for generating controlled responses, and the classification model accomplished relatively high scores of accuracy (95%), precision (0.954), and recall (0.947) and an F_1-score of 0.95 [33].

However, high accuracy and a more nuanced response generation were consistently apparent in all the CAs using hybrid architectures [24,27,28,30,34], suggesting that hybrid models lead to enhanced performance in tasks requiring complex understanding of user emotions and the generation of contextual responses.

Human Evaluation of CAs

Most of the reviewed studies, 16 (84%) out of 19, conducted a human evaluation of the implemented CA designs. Acceptability by end users was evaluated in terms of user experience, satisfaction, and levels of engagement. A detailed summary of the human evaluations of these designs is presented in Multimedia Appendix 5.

The human evaluation was performed by only CA users in most cases (13/16, 81%), while experts in the field of MH contributed to the process of assessing the CA in the remaining studies (3/16, 19%). Table 3 summarizes the empathy measures used in these papers.

Table 3. Measurement of empathy in CAs[a].

Study and year	The method of empathy measurement	How was empathy measured?	Who did the evaluation?	Evaluation results

Jiang et al [21], 2022	• Self-reports:
	  • In-depth interview responses
	  • Multiple response ratings
		Using the RoPE[b] scale (binary responses) and QCAE[c]	• Replika users provided the empathy ratings	• Perceived cognitive empathy was higher than perceived affective empathy

Beredo and Ong [28], 2022	• Self-reports: • Response ratings	Affect criterion or empathy was measured using a binary scale of 0 (no) to 1 (yes)	• Evaluated by 3 experts who studied and practiced psychology	• Responses were rated 79% empathic
Alazraki et al [34], 2021	• Self-reports: • Multiple response ratings	Multiple ratings to evaluate the perceived level of empathy, with ratings ranging from strongly disagree to strongly agree on a 5-point Likert scale	• Evaluated by users • 2 separate clinicians specialized in MH[d] also evaluated the chatbot personas	• When interacting with the Kai persona, 75% of users agreed that the bot was empathic • Interaction with other study personas achieved a 56% empathic rating
Mishra et al [36], 2023	• Self-reports: • Response ratings	A single 5-point Likert scale	• 6 evaluators rated each dialogue interaction for empathy • Empathy ratings by evaluators cross-validated for quality by government-run institutions	• Average empathy rating=57%
Agnihotri et al [33], 2021	• Self-reports: • Response ratings	Emotional relevance is rated using a single 5-point Likert scale	• Evaluated by 3 human annotators—male nonnative English speakers from a technical university with an average age of 21 years	When an empathic response generator was used, emotional relevance=61.4% • When a topic classifier was added, emotional relevance=43%

[a]CA: conversational agent.

[b]RoPE: Robot's Perceived Empathy.

[c]QCAE: Questionnaire of Cognitive and Affective Empathy.

[d]MH: mental health.

Alazraki et al [34] conducted a cross-sectional study with 23 volunteers and 2 clinicians who engaged with a web-based chatbot platform using 4 prescripted conversations of different CA personas. An anonymous web-based questionnaire collected participant feedback regarding the level of empathy displayed by the chatbot, engagement levels, and the ability of the chatbot to identify emotions in the participant. The survey results revealed that 75% of users agreed that the CA persona Kai was empathic, 63% found it engaging, and 75% rated it as useful. In contrast, Beredo and Ong [28] asked 3 psychologists to provide feedback on chatbot user logs. Empathy was measured using the affect criterion, a measure of the ability of the CA to read and respond to users with empathy, along with performance and humanlike characteristics. On the basis of expert feedback, 67% of the CA responses were relevant, 78% seemed human, and 70% were empathic.

In an RCT, a group of 39 participants were randomly allocated to a treatment group interacting with the emotion-aware chatbot EMMA, while a control group (n=39) was assigned to an emotionally nonexpressive chatbot, with 2 weeks of monitoring in each case [24]. The participants engaging with EMMA showed higher frequency of interactions and responded quicker than the control group. The feedback of the users was useful in understanding how empathy was perceived during the study.

The only qualitative experimental study involved an AI-based chatbot, Replika, designed to improve resilience and user well-being [21]. The author followed an ethnographic approach for their study of empathy, asking users to download the Replika application and write down reflective notes on their conversations with Replika. The results of this study expand the empathy theories within human conversations to human-AI interactions through variations in cognitive empathy, affective empathy, and empathic responses. A list of technical terms used in the paper is further explained in Multimedia Appendix 6.

Risk of Bias and Quality Assessment Results

The included RCTs showed a low risk of bias on the revised Cochrane risk-of-bias tool. Of the 14 nonrandomized studies included in the review, all showed a moderate to high risk of bias. A total of 5 (36%) studies [27,32-34,36] were moderately biased, and 1 (7%) study [28] was seriously biased according to the risk of bias in nonrandomized studies of interventions tool. The Joanna Briggs Institute quality assessment results were generally low when only the design component of the studies was assessed, with 32% (6/19) of the papers receiving a score of 0. However, an overall moderate quality was seen in publications when both the design and implementation stages were appraised. Multimedia Appendix 7 [4-6,21-36] shows the quality assessment results.

Discussion

Principal Findings

The study and use of CA technology have been the subject of extensive research across many fields, such as education, customer service, and health care. Moreover, there are reviews focusing on AI-based CAs, their effectiveness, and their impact in the realm of MH care [17,18,42]. While these reviews offer significant insights into AI-based CA designs in MH care, the importance of empathy is not central. Although these reviews suggest the need for empathy in CA innovations in MH care, they do not consider CA designs specifically aimed at generating and evaluating empathy. To address this gap, this review compares various empathic CA designs, their effectiveness in detecting and responding to user emotions, and their acceptability to users.

CA Designs

This review has found that most researchers used an ML-based transformer engine for designing empathic CAs, achieving excellent classification and prediction results. Surprisingly, several researchers used rule-based architectures and retrieval engines. While lacking the sophistication of transformer-based engines in terms of comprehension, rule-based approaches were able to efficiently identify keywords and themes, ensuring that consumer needs were addressed within a limited number of categories. Rule-based systems are comparatively easy to design and implement, allowing for a trade-off between classification accuracy and economic feasibility. However, rule-based systems tend to generate more predictable, inflexible, and repetitive responses compared to advanced LLM engines and, therefore, might be more suitable for providing simple information to managers and MH care workers, rather than responding to end users requiring more nuanced responses.

Hybrid architecture seems best suited to the detection of user emotion followed by the retrieval of a suitable response. Therefore, having >1 model appears to facilitate a more robust model output. This is supported by the superior accuracies achieved by hybrid architectures in the classification and prediction tasks. The hybrid model of Adikari et al [27] achieved the highest accuracy of 87% (F_1-score=0.87) in recommending a resource based on the concerns expressed by the patients. However, the highest accuracy in emotion recognition (95% accuracy in identifying sadness, anger, fear, and happiness) was obtained by Alazraki et al [34]. The combined features of high accuracy and improved user experience probably make these the best performing CAs within the review.

While the use of such robust LLMs has significantly improved language-based CA technology, it is important to recognize that these models are not without disadvantages [43]. These models have been found to perpetuate biases with regard to gender, race, and MH conditions present in the training data [44,45]. Such biases can strengthen gender stereotypes and reduce response accuracy when dealing with users from diverse cultural backgrounds, potentially causing harm to users. Such issues may have serious impacts on user trust, the credibility of the empathic CA, and user well-being. Such biases can be mitigated by ensuring that the training data sets represent diverse gender categories, races, and cultural backgrounds and that advanced technical approaches are used to detect and minimize any such biases in the training data [46-48].

Ethical and privacy concerns associated with these LLMs are critical [49,50]. Following ethical guidelines centered around transparency, accountability, and adherence are pivotal to user privacy, while measures to maintain data security through strict access controls and regular security checks also need to be in place. Privacy should be a core component of CA designs, with limitations placed on personal data collection whenever possible [7]. These strategies are especially important for an empathic CA design dealing with users seeking MH care. Any breaches of privacy and ethical guidelines pose a high risk to user mental well-being as well as users' trust in and acceptance of these new technologies [51]. The AI safety guidelines established by the European Union provide a key foundation for the creation of secure and ethical experiences for users [50].

Due to the complexity of LLMs and the many parameters involved, some models can have high latency in response time, which can cause potential challenges for a real-time CA dealing with vulnerable users waiting for a response. However, the use of parallel processing, optimization techniques, and hardware that supports the requirements of these AI models has facilitated a decrease in execution times [52].

Human Evaluations of CAs

Among the reviewed publications, human evaluation of chatbots was common. However, only 26% (5/19) of the studies used an RCT design to assess the CA platform. Random assignment to the treatment arm is known to reduce bias while improving the reliability of the experimental results. Any confounding factors are, therefore, likely to be controlled for in an RCT, making it important to overcome the practical difficulties these designs present in this context. RCTs provide the opportunity to observe user experiences with the CA designs over time. Ideally, future studies should consider RCT designs for their human evaluations, and ideally, the long-term effects of the CA can be examined over an extended timeline.

Previous experiences with CAs could be an important confounding factor. On the basis of these experiences, expectations of users regarding CA performance may affect actual engagement with the CA. Previous bad experiences may make it less likely that a user will try to engage fully with a CA, resulting in a less favorable evaluation and satisfaction levels [53]. Another confounding factor could be the rate at which the user likes to communicate. If the CA cannot automatically adapt its speed of response to that preferred by the user, it is likely that this will also impact evaluation results [54].

The human evaluations of CAs in this review focused on their ability to portray empathy, satisfy user needs, provide useful and contextually informed responses, and facilitate user engagement. Most CAs were evaluated as satisfactory by end users. However, among the 19 papers reviewed, only 5 (26%) papers provided quantitative evaluations of CA empathy, and only 5 (26%) papers provided a definition of empathy.

Because empathy has been defined in numerous ways in the literature, it is important that in future studies users are given a framework that guides their perceptions of empathy. Future research on empathic CA designs would, therefore, benefit from a clear and well-established definition of empathy, such as that provided by the World Health Organization [8]. Ideally, standardized scales for perceived empathy should be used to enhance the reliability, comparability, and validity of survey results. In this review, other self-report measures were used as surrogates for empathy, with considerable variation in the types of scales used. However, self-report scales are subjective and prone to bias, with different meanings based on users' lived experiences [55]. Ideally, the impact of the CA on MH outcomes should also be assessed. Only 2 (11%) of the 19 papers in this review [4,29] used the Patient Health Questionnaire as their measure of MH outcomes, while 2 (11%) other papers considered stress levels in their evaluation [23,25].

Furthermore, the human evaluations were mostly conducted by study participants. Experts and professionals in the field of MH care were rarely consulted. There is a need for greater consultation with focus groups and user groups to ensure that the CA design best reflects the needs of all stakeholders [22]. Future research in this area should also consider an iterative design framework, incorporating the co-design and coevaluation of prototypes involving all stakeholders [22].

In summary, there were deficiencies in all the human evaluations included in this review. Only 5 (26%) of the 19 papers in this review included a direct evaluation of CA empathy in the design, while the rest (n=14, 74%) were more concerned with general user satisfaction. Only 2 (40%) of 5 these studies used multiple rating scales to measure the level of empathy portrayed by a CA, and only 1 (20%) of 5 these studies [34] considered evaluations by both users and clinicians. However, there were 4 studies that did consider the impact of the CA on MH outcomes.

Future Opportunities

A significant limitation of the CAs reviewed was the use of only textual input in all but 2 (11%) of the 19 studies where voice data were included, thus losing a valuable opportunity to leverage alternative and powerful forms of data input for evaluating empathy.. A range of vocal characteristics have been associated with the detection of suicide risk and psychological distress, which suggests that vocal characteristics might provide a natural extension for the detection of levels of empathy [56,57]. The omission of voice data is surprising given that empathy is communicated predominately through vocal cues. However, textual information is not without its advantages. As we have shown in this review, NLP approaches have been used to successfully detect and convey empathy by CAs. A novel approach would be to leverage both streams of information to identify vocal characteristics indicative of different levels of empathy in addition to textual cues. Characteristics of vocal and textual cues that are associated with empathy could be combined to create a CA design to attend to users of MH care facilities such as helpline services, patient triage, and emergency services [21,23].

Creating a CA design that accurately portrays empathy and adjusts the level of empathy to match the emotional status of patients is a significant challenge. Effective vocal interaction often faces hurdles due to technical issues in voice analysis, including the smooth processing and interpretation of data. These challenges are compounded by poor audio quality [58]; the presence of overlapping psychological states in users; and linguistic variability influenced by culture, age, gender, and accents [59-61]. The use of high-quality audio devices to capture user voice [62, 63] and the use of training data sets reflecting diverse human demographic features are two challenges in algorithm development aiming to provide effective vocal interaction in CAs in real time.

The integration of an empathic CA with voice analysis capabilities into crisis helpline services could benefit users and the service provider. Attending to callers during peak hours for the collection of demographic information, triage, and risk assessment of callers using their voice patterns are some of the possible roles that CAs could fulfill. The involvement of CAs in these capacities could help reduce caller wait times, streamline processes, and ensure 24-hour service availability while providing a nonjudgmental and sensitive interaction for users within a safe environment. Improved empathy portrayal by the CA would help enhance user engagement and CA acceptability, helping reduce the gap between the demand and supply of available crisis helpline services.

Summary

This review confirms that empathy is an important characteristic for CA implementation for MH care. It highlights the strengths of the ML-based architectures when it comes to CA design and provides evidence of both technical and human assessments of CA performance. The need for improvement in measures used for detecting the level of empathy exhibited by CAs is manifest. The importance of AI safety regarding ethical and privacy concerns is a neglected area and should be considered as a priority for future designs. The promise of empathic CA applications that use vocal inputs and outputs is another area warranting further research, with opportunities for crisis helpline services.

Limitations of the Review

The studies included in this review presented a mix of methods, which made it challenging to compare and analyze the results. This relates to the diversity in the CA designs included, along with the different data formats obtained through human evaluations, such as survey results, response ratings, and interview feedback. The methods used to assess the accuracy of the technical designs were also varied, and a lack of empathy definitions and standard measures for perceived empathy made study comparisons difficult.

The quality rating of the studies emphasized the need for the complete reporting of CA designs as well as rigorous evaluation. Deficiencies in these areas meant that the quality ratings for several papers were low. Evaluation guidelines were often missing, which made it challenging to appraise the performance of these systems. Classification accuracy and the accuracy of the responses generated were assessed using a variety of methods, further complicating this comparison.

Conclusions

The objective of this systematic review was to identify the existing architectures of empathic CA designs and the types of CA design assessments used in MH care. A further aim was to determine how CA empathy is evaluated and to examine the limitations and future ideas for CAs in this specific context. More than half of the selected papers used the latest technologies in CA architectures, including designs developed using ML-based transformer engines (eg, LLMs). Evaluations of technical capabilities were conducted in most of the papers and demonstrated good levels of accuracy.

This review suggests that a hybrid design is ideally used for the design of an empathic CA, allowing an initial assessment of user emotion before any CA response is developed. This review indicates that human feedback is required to assess the extent to which the CA is successful in demonstrating empathy. It is recommended that well-validated scales be used for this purpose. Further research on the portrayal of empathy in CAs for MH care would benefit by involving cocreation activities, explicit definitions of empathy, and effective evaluation of empathy using standardized empathy scales, as well as by using vocal features associated with empathy in addition to textual cues.

Despite its limitations, this review demonstrates that it is possible to design AI-empowered CAs that evoke empathy within MH care applications, with many of these CAs being rated as satisfactory by human users. This suggests that such CAs could prove beneficial in a range of settings, such as crisis helpline services, gathering data on user characteristics and emotions, and in postvention follow-up, helping to bridge the gap between the existing supply and demand for MH services.

Acknowledgments

This study was funded by Swinburne University of Technology.

Authors' Contributions

RS, DM, and RI contributed to the study selection process. RS and DM conducted the quality assessment of the included studies. RS, DM, RI, PA, and NW were involved in the concept, design, revisions, and final approval of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Screening process and study characteristics.
[DOCX File , 42 KB - mental_v11i1e58974_app1.docx]

Multimedia Appendix 2

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist 2020.
[DOCX File , 32 KB - mental_v11i1e58974_app2.docx]

Multimedia Appendix 3
Evolution of conversational agent (year by year).
[DOCX File , 56 KB - mental_v11i1e58974_app3.docx]

Multimedia Appendix 4
Detailed summary of conversational agent types.
[DOCX File , 49 KB - mental_v11i1e58974_app4.docx]

Multimedia Appendix 5
Results of conversational agent evaluations.
[DOCX File , 99 KB - mental_v11i1e58974_app5.docx]

Multimedia Appendix 6
Dictionary of technical terms.
[DOCX File , 16 KB - mental_v11i1e58974_app6.docx]

Multimedia Appendix 7
Risk of bias and quality assessment.
[DOCX File , 16 KB - mental_v11i1e58974_app7.docx]

References

1. Mental health atlas 2020. World Health Organization. 2021 Oct 8. URL: https://www.who.int/publications/i/item/9789240036703 [accessed 2024-08-10]
2. Schick A, Feine J, Morana S, Maedche A, Reininghaus U. Validity of chatbot use for mental health assessment: experimental study. JMIR Mhealth Uhealth 2022 Oct 31;10(10):e28082 [FREE Full text] [doi: 10.2196/28082] [Medline: 36315228]
3. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 06;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
4. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth 2018 Nov 23;6(11):e12106 [FREE Full text] [doi: 10.2196/12106] [Medline: 30470676]
5. Rani K, Vishnoi H, Mishra M. A mental health chatbot delivering cognitive behavior therapy and remote health monitoring using NLP and AI. In: Proceedings of the International Conference on Disruptive Technologies. 2023 Presented at: ICDT 2023; May 11-12, 2023; Greater Noida, India. [doi: 10.1109/icdt57929.2023.10150665]
6. Persons B, Jain P, Chagnon C, Djamasbi S. Designing the empathetic research IoT network (ERIN) chatbot for mental health resources. In: Proceedings of the HCI in Business, Government and Organizations: 8th International Conference, HCIBGO 2021, Held as Part of the 23rd HCI International Conference, HCII 2021. 2021 Presented at: HCI 2021; July 24-29, 2021; Virtual Event. [doi: 10.1007/978-3-030-77750-0_41]
7. Li L, Peng W, Rheu MM. Factors predicting intentions of adoption and continued use of artificial intelligence chatbots for mental health: examining the role of UTAUT model, stigma, privacy concerns, and artificial intelligence hesitancy. Telemed J E Health 2024 Mar 01;30(3):722-730. [doi: 10.1089/tmj.2023.0313] [Medline: 37756224]
8. Hojat M. A definition and key features of empathy in patient care. In: Hojat M, editor. Empathy in Health Professions Education and Patient Care. Cham, Switzerland: Springer; 2016.
9. Koulouri T, Macredie RD, Olakitan D. Chatbots to support young adults' mental health: an exploratory study of acceptability. ACM Trans Interact Intell Syst 2022 Jul 20;12(2):1-39. [doi: 10.1145/3485874]
10. He Y, Yang L, Qian C, Li T, Su Z, Zhang Q, et al. Conversational agent interventions for mental health problems: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2023 Apr 28;25:e43862 [FREE Full text] [doi: 10.2196/43862] [Medline: 37115595]
11. Kraus M, Seldschopf P, Minker W. Towards the development of a trustworthy chatbot for mental health applications. In: Proceedings of the 27th International Conference on MultiMedia Modeling. 2021 Presented at: MMM 2021; June 22-24, 2021; Prague, Czech Republic. [doi: 10.1007/978-3-030-67835-7_30]
12. Kallivalappil N, D'souza K, Deshmukh A, Kadam C, Sharma N. Empath.ai: a context-aware chatbot for emotional detection and support. In: Proceedings of the 14th International Conference on Computing Communication and Networking Technologies. 2023 Presented at: ICCCNT 2023; July 6-8, 2023; Delhi, India. [doi: 10.1109/icccnt56998.2023.10306584]

13. Lin S, Lin L, Hou C, Chen B, Li J, Ni S. Empathy-based communication framework for chatbots: a mental health chatbot application and evaluation. In: Proceedings of the 11th International Conference on Human-Agent Interaction. 2023 Presented at: HAI '23; December 4-7, 2023; Gothenburg, Sweden. [doi: 10.1145/3623809.3623865]
14. Boucher EM, Harake NR, Ward HE, Stoeckl SE, Vargas J, Minkel J, et al. Artificially intelligent chatbots in digital mental health interventions: a review. Expert Rev Med Devices 2021 Dec;18(sup1):37-49 [FREE Full text] [doi: 10.1080/17434440.2021.2013200] [Medline: 34872429]
15. Sweeney C, Potts C, Ennis E, Bond R, Mulvenna MD, O'neill S, et al. Can chatbots help support a person's mental health? Perceptions and views from mental healthcare professionals and experts. ACM Trans Comput Healthcare 2021 Jul 15;2(3):1-15. [doi: 10.1145/3453175]
16. Daley K, Hungerbuehler I, Cavanagh K, Claro HG, Swinton PA, Kapps M. Preliminary evaluation of the engagement and effectiveness of a mental health chatbot. Front Digit Health 2020 Nov 30;2:576361 [FREE Full text] [doi: 10.3389/fdgth.2020.576361] [Medline: 34713049]
17. Li H, Zhang R, Lee YC, Kraut RE, Mohr DC. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med 2023 Dec 19;6(1):236 [FREE Full text] [doi: 10.1038/s41746-023-00979-5] [Medline: 38114588]
18. Gaffney H, Mansell W, Tai S. Conversational agents in the treatment of mental health problems: mixed-method systematic review. JMIR Ment Health 2019 Oct 18;6(10):e14166 [FREE Full text] [doi: 10.2196/14166] [Medline: 31628789]
19. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and opinions of patients about mental health chatbots: scoping review. J Med Internet Res 2021 Jan 13;23(1):e17828 [FREE Full text] [doi: 10.2196/17828] [Medline: 33439133]
20. Behrens S. The history and evolution of conversational AI. Fabric. 2021 Apr 17. URL: https://gyant.com/the-history-and-evolution-of-conversational-ai/ [accessed 2023-03-13]
21. Jiang Q, Zhang Y, Pian W. Chatbot as an emergency exist: mediated empathy for resilience via human-AI interaction during the COVID-19 pandemic. Inf Process Manag 2022 Nov;59(6):103074 [FREE Full text] [doi: 10.1016/j.ipm.2022.103074] [Medline: 36059428]
22. Brocki L, Dyer GC, Gładka A, Chung NC. Deep learning mental health dialogue system. In: Proceedings of the IEEE International Conference on Big Data and Smart Computing. 2023 Presented at: BigComp 2023; February 13-16, 2023; Jeju, Republic of Korea. [doi: 10.1109/bigcomp57234.2023.00097]
23. Trappey AJ, Lin AP, Hsu KY, Trappey CV, Tu KL. Development of an empathy-centric counseling chatbot system capable of sentimental dialogue analysis. Processes 2022 May 08;10(5):930. [doi: 10.3390/pr10050930]
24. Ghandeharioun A, McDuff D, Czerwinski M, Rowan K. EMMA: an emotion-aware wellbeing chatbot. In: Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction. 2019 Presented at: ACII 2019; September 3-6, 2019; Cambridge, UK. [doi: 10.1109/acii.2019.8925455]
25. Meng J, Dai Y. Emotional support from AI chatbots: should a supportive partner self-disclose or not? J Comput Mediat Commun 2021 May 19;26(4):207-222. [doi: 10.1093/jcmc/zmab005]
26. Goel R, Vashisht S, Dhanda A, Susan S. An empathetic conversational agent with attentional mechanism. In: Proceedings of the International Conference on Computer Communication and Informatics. 2021 Presented at: ICCCI 2021; January 27-29, 2021; Coimbatore, India. [doi: 10.1109/iccci50826.2021.9402337]
27. Adikari A, de Silva D, Moraliyage H, Alahakoon D, Wong J, Gancarz M, et al. Empathic conversational agents for real-time monitoring and co-facilitation of patient-centered healthcare. Future Gener Comput Syst 2022 Jan;126:318-329. [doi: 10.1016/j.future.2021.08.015]
28. Beredo JL, Ong EC. A hybrid response generation model for an empathetic conversational agent. In: Proceedings of the International Conference on Asian Language Processing. 2022 Presented at: IALP 2022; October 27-28, 2022; Singapore, Singapore. [doi: 10.1109/ialp57159.2022.9961311]
29. Rathnayaka P, Mills N, Burnett D, De Silva D, Alahakoon D, Gray R. A mental health chatbot with cognitive skills for personalised behavioural activation and remote health monitoring. Sensors (Basel) 2022 May 11;22(10):3653 [FREE Full text] [doi: 10.3390/s22103653] [Medline: 35632061]
30. Morris RR, Kouddous K, Kshirsagar R, Schueller SM. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. J Med Internet Res 2018 Jun 26;20(6):e10148 [FREE Full text] [doi: 10.2196/10148] [Medline: 29945856]
31. Ghandeharioun A, McDuff D, Czerwinski M, Rowan K. Towards understanding emotional intelligence for behavior change chatbots. In: Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction. 2019 Presented at: ACII 2019; September 3-6, 2019; Cambridge, UK. [doi: 10.1109/acii.2019.8925433]
32. Saha T, Gakhreja V, Das AS, Chakraborty S, Saha S. Towards motivational and empathetic response generation in online mental health support. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022 Presented at: SIGIR '22; July 11-15, 2022; Madrid, Spain. [doi: 10.1145/3477495.3531912]
33. Agnihotri M, Rao SB, Jayagopi DB, Hebbar S, Rasipuram S, Maitra A, et al. Towards generating topic-driven and affective responses to assist mental wellness. In: Proceedings of the 25th International Conference on Pattern Recognition. 2021 Presented at: ICPR 2021; January 10-15, 2021; Milan, Italy. [doi: 10.1007/978-3-030-68790-8_11]

34. Alazraki L, Ghachem A, Polydorou N, Khosmood F, Edalat A. An empathetic AI coach for self-attachment therapy. In: Proceedings of the IEEE Third International Conference on Cognitive Machine Intelligence. 2021 Presented at: CogMI 2021; December 13-15, 2021; Atlanta, GA. [doi: 10.1109/cogmi52975.2021.00019]
35. Gundavarapu MR, Koundinya GS, Sai TB, Sree GK. Empathic chatbot: emotional astuteness for mental health well-being. In: Proceedings of the 7th International Conference on Computing in Engineering & Technology. 2022 Presented at: ICCET 2022; February 25-27, 2022; Virtual Event. [doi: 10.1007/978-981-19-2719-5_65]
36. Mishra K, Priya P, Ekbal A. Help me heal: a reinforced polite and empathetic mental health and legal counseling dialogue system for crime victims. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. 2023 Presented at: AAAI 2023; February 7-14, 2023; Washington, DC. [doi: 10.1609/aaai.v37i12.26685]
37. Gotschall T. EndNote 20 desktop version. J Med Libr Assoc 2021 Jul 01;109(3):520-522 [FREE Full text] [doi: 10.5195/jmla.2021.1260] [Medline: 34629985]
38. Moola S, Munn Z, Tufanaru C, Aromataris E, Sears K, Sfetcu R, et al. Systematic reviews of etiology and risk. In: Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z, editors. JBI Manual for Evidence Synthesis. Adelaide, Australia: Joanna Briggs Institute; 2024.
39. Rogers CR. Empathic: an unappreciated way of being. Couns Psychol 2016 Sep 04;5(2):2-10. [doi: 10.1177/001100007500500202]
40. Elliott R, Bohart AC, Watson JC, Murphy D. Therapist empathy and client outcome: an updated meta-analysis. Psychotherapy (Chic) 2018 Dec;55(4):399-410 [FREE Full text] [doi: 10.1037/pst0000175] [Medline: 30335453]
41. Barrett-Lennard GT. The empathy cycle: refinement of a nuclear concept. J Counsel Psychol 1981;28(2):91-100. [doi: 10.1037//0022-0167.28.2.91]
42. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry 2019 Jul;64(7):456-464 [FREE Full text] [doi: 10.1177/0706743719828977] [Medline: 30897957]
43. Xu X, Yao B, Dong Y, Gabriel S, Yu H, Hendler J, et al. Mental-LLM: leveraging large language models for mental health prediction via online text data. Proc ACM Interact Mob Wearable Ubiquitous Technol 2024 Mar 06;8(1):1-32. [doi: 10.1145/3643540]
44. Salutari F, Ramos J, Rahmani HA, Linguaglossa L, Lipani A. Quantifying the bias of transformer-based language models for African American English in masked language modeling. In: Proceedings of the 27th Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2023 Presented at: PAKDD 2023; May 25-28, 2023; Osaka, Japan. [doi: 10.1007/978-3-031-33374-3_42]
45. Li B, Peng H, Sainju R, Yang J, Yang L, Liang Y, et al. Detecting gender bias in transformer-based models: a case study on BERT. arXiv Preprint posted online on October 15, 2021. [doi: 10.48550/arXiv.2110.15733]
46. Kamboj P, Kumar S, Goyal V. Measuring and mitigating gender bias in contextualized word embeddings. In: Proceedings of the IEEE International Conference on Blockchain and Distributed Systems Security. 2023 Presented at: ICBDS 2023; October 6-8, 2023; New Raipur, India. [doi: 10.1109/icbds58040.2023.10346586]
47. Meade N, Poole-Dayan E, Reddy S. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. arXiv Preprint posted online on October 16, 2021 [FREE Full text] [doi: 10.18653/v1/2022.acl-long.132]
48. Gira M, Zhang R, Lee K. Debiasing pre-trained language models via efficient fine-tuning. In: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion. 2022 Presented at: LTEDI 2022; May 27, 2022; Dublin, Ireland. [doi: 10.18653/v1/2022.ltedi-1.8]
49. Kumar A, Murthy SV, Singh S, Ragupathy S. The ethics of interaction: mitigating security threats in LLMs. arXiv Preprint posted online on January 22, 2024.
50. Khowaja SA, Khuwaja P, Dev K, Wang W, Nkenyereye L. ChatGPT needs SPADE (sustainability, PrivAcy, digital divide, and ethics) evaluation: a review. Cogn Comput 2024 May 05. [doi: 10.1007/s12559-024-10285-1]
51. Coghlan S, Leins K, Sheldrick S, Cheong M, Gooding P, D'Alfonso S. To chat or bot to chat: ethical issues with using chatbots in mental health. Digit Health 2023 Jun 22;9:20552076231183542 [FREE Full text] [doi: 10.1177/20552076231183542] [Medline: 37377565]
52. Agrawal A, Kedia N, Panwar A, Mohan J, Kwatra N, Gulavani BS, et al. Taming throughput-latency tradeoff in LLM inference with Sarathi-serve. arXiv Preprint posted online on March 4, 2024 [FREE Full text] [doi: 10.48550/arXiv.2403.02310]
53. Santhanam S, Karduni A, Shaikh S. Studying the effects of cognitive biases in evaluation of conversational agents. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020 Presented at: CHI '20; April 25-30, 2020; Honolulu, HI. [doi: 10.1145/3313831.3376318]
54. Gnewuch U, Morana S, Adam MT, Maedche A. Opposing effects of response time in human–chatbot interaction. Bus Inf Syst Eng 2022 May 30;64:773-791. [doi: 10.1007/s12599-022-00755-x]
55. Prince SA, Cardilli L, Reed JL, Saunders TJ, Kite C, Douillette K, et al. A comparison of self-reported and device measured sedentary behaviour in adults: a systematic review and meta-analysis. Int J Behav Nutr Phys Act 2020 Mar 04;17(1):31 [FREE Full text] [doi: 10.1186/s12966-020-00938-3] [Medline: 32131845]

56. Iyer R, Nedeljkovic M, Meyer D. Using voice biomarkers to classify suicide risk in adult telehealth callers: retrospective observational study. JMIR Ment Health 2022 Aug 15;9(8):e39807 [FREE Full text] [doi: 10.2196/39807] [Medline: 35969444]
57. Iyer R, Nedeljkovic M, Meyer D. Using vocal characteristics to classify psychological distress in adult helpline callers: retrospective observational study. JMIR Form Res 2022 Dec 19;6(12):e42249 [FREE Full text] [doi: 10.2196/42249] [Medline: 36534456]
58. Schaeffler F, Jannetts S, Beck J. Reliability of clinical voice parameters captured with smartphones — measurements of added noise and spectral tilt. In: Proceedings of the 20th Annual Conference of the International Speech Communication Association. 2019 Presented at: INTERSPEECH 2019; September 15-19, 2019; Graz, Austria. [doi: 10.21437/interspeech.2019-2910]
59. Laukka P, Elfenbein HA, Thingujam NS, Rockstuhl T, Iraki FK, Chui W, et al. The expression and recognition of emotions in the voice across five nations: a lens model analysis based on acoustic features. J Pers Soc Psychol 2016 Nov;111(5):686-705. [doi: 10.1037/pspi0000066] [Medline: 27537275]
60. Markova D, Richer L, Pangelinan M, Schwartz DH, Leonard G, Perron M, et al. Age- and sex-related variations in vocal-tract morphology and voice acoustics during adolescence. Horm Behav 2016 May;81:84-96. [doi: 10.1016/j.yhbeh.2016.03.001] [Medline: 27062936]
61. Israelsson A, Seiger A, Laukka P. Blended emotions can be accurately recognized from dynamic facial and vocal expressions. J Nonverbal Behav 2023 May 17;47(3):267-284. [doi: 10.1007/s10919-023-00426-9]
62. Busquet F, Efthymiou F, Hildebrand C. Voice analytics in the wild: validity and predictive accuracy of common audio-recording devices. Behav Res Methods 2024 Mar 30;56(3):2114-2134 [FREE Full text] [doi: 10.3758/s13428-023-02139-9] [Medline: 37253958]
63. Padmapriya J, Sasilatha T, R K, Aagash G, Bharathi V. Voice extraction from background noise using filter bank analysis for voice communication applications. In: Proceedings of the Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks. 2021 Presented at: ICICV 2021; February 4-6, 2021; Tirunelveli, India. [doi: 10.1109/icicv50876.2021.9388453]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations Transformer
CA: conversational agent
LLM: large language model
MH: mental health
ML: machine learning
NLP: natural language processing
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial

Edited by J Torous; submitted 30.03.24; peer-reviewed by Y Okoro, P Upadhyaya, BG Teferra, M Bagewadi Ellur; comments to author 06.05.24; revised version received 01.07.24; accepted 02.07.24; published 09.09.24.

Please cite as:
Sanjeewa R, Iyer R, Apputhurai P, Wickramasinghe N, Meyer D
Empathic Conversational Agent Platform Designs and Their Evaluation in the Context of Mental Health: Systematic Review
JMIR Ment Health 2024;11:e58974
URL: https://mental.jmir.org/2024/1/e58974
doi:10.2196/58974
PMID:

Review

Digital Mental Health Interventions for Alleviating Depression and Anxiety During Psychotherapy Waiting Lists: Systematic Review

Sijia Huang[1], MA; Yiyue Wang[1], BS; Gen Li[2], PhD; Brian J Hall[2,3,4], PhD; Thomas J Nyman[1,2,5], PhD

[1]Faculty of Arts and Sciences, New York University Shanghai, Shanghai, China

[2]Center for Global Health Equity, New York University Shanghai, Shanghai, China

[3]School of Global Public Health, New York University, New York, NY, United States

[4]Health, Behavior and Society, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States

[5]School of Psychology and Clinical Language Sciences, University of Reading, Reading, United Kingdom

Corresponding Author:
Thomas J Nyman, PhD
School of Psychology and Clinical Language Sciences
University of Reading
Earley
Reading, RG6 6ET
United Kingdom
Phone: 44 1189875123
Email: t.nyman@reading.ac.uk

Related Article:

This is a corrected version. See correction statement: https://mental.jmir.org/2024/1/e67281

Abstract

Background: Depression and anxiety have become increasingly prevalent across the globe. The rising need for treatment and the lack of clinicians has resulted in prolonged waiting times for patients to receive their first session. Responding to this gap, digital mental health interventions (DMHIs) have been found effective in treating depression and anxiety and are potentially promising pretreatments for patients who are awaiting face-to-face psychotherapy. Nevertheless, whether digital interventions effectively alleviate symptoms for patients on waiting lists for face-to-face psychotherapy remains unclear.

Objective: This review aimed to synthesize the effectiveness of DMHIs for relieving depression and anxiety symptoms of patients on waiting lists for face-to-face therapy. This review also investigated the features, perceived credibility, and usability of DMHIs during waiting times.

Methods: In this systematic review, we searched PubMed, PsycINFO, Cochrane, and Web of Science for research studies investigating the effectiveness of DMHIs in reducing either depression or anxiety symptoms among individuals waiting for face-to-face psychotherapy. The search was conducted in June 2024, and we have included the studies that met the inclusion criteria and were published before June 6, 2024.

Results: Of the 9267 unique records identified, 8 studies met the eligibility criteria and were included in the systematic review. Five studies were randomized controlled trials (RCTs), and 3 studies were not. Among the RCTs, we found that digital interventions reduced depression and anxiety symptoms, but the majority of interventions were not more effective compared to the control groups where participants simply waited or received a self-help book. For the non-RCTs, the interventions also reduced symptoms, but without control groups, the interpretation of the findings is limited. Finally, participants in the included studies perceived the digital interventions to be credible and useful, but high dropout rates raised concerns about treatment adherence.

Conclusions: Due to the lack of effective interventions among the reviewed studies, especially among the RCTs, our results suggest that waiting list DMHIs are not more effective compared to simply waiting or using a self-help book. However, more high-quality RCTs with larger sample sizes are warranted in order to draw a more robust conclusion. Additionally, as this review revealed concerns regarding the high dropout rate in digital interventions, future studies could perhaps adopt more personalized and human-centered functions in interventions to increase user engagement, with the potential to increase treatment adherence and effectiveness.

(JMIR Ment Health 2024;11:e56650) doi:10.2196/56650

KEYWORDS

digital health; digital technology; digital intervention; digital interventions; waiting list; digital mental health intervention; DMHI; digital mental health interventions; DMHIs; digital mental health; mental health intervention; mental health interventions; mental health; mental illness; mental disease; mental diseases; mental illnesses; depression; depressed; major depressive disorder; MDD; depressive disorder; depressive; anxiety; anxious; self-guided; self-guidance; self-mediated; self-mediation; systematic review; systematic reviews; mood disorder; therapy; tele-therapy; web-based therapy

Introduction

Poor mental health is an increasing global challenge [1,2], with the prevalence of depression and anxiety reaching a peak during the COVID-19 pandemic [3-5]. Several meta-analyses have concluded that psychotherapies, such as cognitive behavioral therapy (CBT), acceptance and commitment therapy, and problem-solving therapy (PST), are as efficacious at treating depression and anxiety as pharmacotherapy, and that their effects are more enduring [6-8]. Compared to pharmacotherapy, psychotherapy is often the preferred treatment for depression and anxiety by most patients, as well as clinicians [8]. According to the latest Mental Health Gap Action Program guideline by the World Health Organization [9], psychotherapy, instead of medication, is usually recommended as a frontline treatment for common mental disorders.

However, access to psychotherapy treatment remains limited worldwide, even in high-income countries [1]. The insufficient number of clinicians has led to prolonged waiting times before face-to-face psychotherapy can begin. For example, in Germany, waiting times are on average 12.5 weeks for an initial consultation and 23.4 weeks for the first therapy session [10]. In the Netherlands, patients wait at least 6 weeks for the first treatment session [11]. In Hong Kong, while urgent and semiurgent cases wait for 1-4 weeks, stable cases, which are 76% of the total cases, wait for an average of 40 weeks [12]. A prolonged time spent on waiting lists for therapy can have negative impacts on the well-being of patients, demonstrated by increased symptoms of depression and anxiety, remission, deterioration in life quality, and even an increase in mortality [10,13-16]. This delay is also significant for most patients as their symptoms do not usually dissipate naturally with time [10,14].

Although waiting list interventions (ie, interventions that are implemented during the weeks or months before face-to-face treatment) are potentially important for patients who are waiting for treatment, few studies have investigated their impact on reducing depression and anxiety. However, Grünzig et al [17] conducted a systematic review of low-intensity interventions to reduce depressive symptoms before outpatient psychotherapy and found limited evidence of their effectiveness in reducing depressive symptoms. Among the reviewed studies, the interventions were a mix of face-to-face sessions, web-based training for self-help strategies, and supervised bibliotherapy with feedback via email or telephone [17]. Due to the wide range of intervention formats, Grünzig et al [17] also found that acceptance varied between the interventions.

Responding to this treatment gap, digital mental health interventions (DMHIs) are potentially a solution to the waiting list problem because they can be implemented digitally and with minimal input from a therapist [18]. Emerging evidence suggests that DMHIs are effective in improving a wide range of mental health conditions, including depression and anxiety [19,20]. Moreover, DMHIs have adopted psychotherapies, like CBT and acceptance and commitment therapy, and these have been found to be as effective as face-to-face therapies [21,22]. Due to the advantages of DMHIs, such as being more scalable, more accessible, and more cost-effective compared to face-to-face therapies [23,24], DMHIs offer the possibility to provide patients with a pre–face-to-face therapy intervention while they are on a waiting list. In addition, DMHIs can be used in a stepped care model as the frontline treatment, which steps up to more intense and advanced care if patients do not improve with the DMHIs.

However, although evidence shows that DMHIs are effective in treating depression and anxiety, it is unknown whether DMHIs are effective for patients on waiting lists. Unlike face-to-face psychotherapies that are usually delivered in a relatively standard procedure, DMHIs encompass diverse features, such as being available on multiple digital devices (website, computer, mobile apps) and offering many delivery formats (email, text message, virtual reality, games, self-help, or guided) [25-27]. It is unclear which of these features are critical for the creation and implementation of effective interventions. The typical factors undermining the effectiveness of DMHIs include low user engagement and low perceived credibility and usability; both of which are associated with poor treatment adherence and subsequently lower effectiveness [28-30]. In contrast, guided DMHIs have been found to be associated with higher user engagement [31], higher treatment adherence [32], and higher effectiveness compared to self-guided DMHIs [21,22]. In addition to user engagement, perceived credibility and usability are positively correlated with treatment outcomes [33]. Hence, it is important to investigate the perceived credibility and usability of DMHIs, as well as user engagement for patients on waiting lists in order to assess their effectiveness.

To our knowledge, there are no previous systematic reviews that synthesized the effectiveness of DMHIs for relieving symptoms of depression and anxiety for patients on waiting lists for face-to-face psychotherapy. Since depression and anxiety are the most prevalent mental health disorders globally [34], this review focused on DMHI's impacts on alleviating depression and anxiety. This review also investigated the features, perceived credibility, and usability of DMHIs during waiting times.

Methods

Overview

This systematic review was conducted following the guidelines of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses; see Multimedia Appendix 1 for PRISMA checklist) statement [35]. The study protocol was uploaded to the Open Science Framework prior to running the initial search and updates after the search have been documented.

Search Strategy and Selection Criteria

Searches were conducted in the following 4 web-based databases: PubMed, PsycINFO, Cochrane, and Web of Science. Backward and forward citation searching was manually conducted using the Web of Science. The search was carried out in June 2024. Studies published before June 6, 2024, and that met the inclusion criteria were included in the review.

The search terms used were a combination of the population (ie, patients on the waiting lists for psychotherapy), treatment (ie, intervention), and outcome (ie, depression or anxiety, or both). The detailed search terms for each domain are shown in Table 1. In addition, the language was limited to English, and the publication type was limited to peer-reviewed publications in scientific journals with full-text access.

To fulfill the objectives of this review, the inclusion criteria were (1) randomized controlled trials (RCTs), or clinical trials, or pilot studies, or feasibility studies (2) of any kind of digital intervention (3) for individuals on waiting lists (4) for psychotherapy (5) with clear psychological outcome measurements that included depressive or anxiety symptoms.

Table 1. Search terms.

Concept	Search terms[a]
Concept 1: population	"before psychotherapy" OR "before therapy" OR "pretreatment" OR "pretherapy" OR "before outpatient psychotherapy" OR "waiting time" OR "waiting list"
Concept 2: intervention	"digital intervention" OR "digital mental health" OR "DMH" OR "eMental health" OR "mobile mental health" OR "mobile psychiatry" OR "technology" OR "online" OR "mobile" OR "phone" OR "app" OR "web" OR "internet" OR "computer*" OR "ehealth" OR "mhealth" OR "guided" OR "self-help"
Concept 3: outcome	"depress*" OR "anxiety" OR "MDD" OR "anxious" OR "GAD" OR "generalized anxiety disorder"

[a]Search terms are based on the inclusion criteria (see eligibility criteria), targeting the population, intervention, outcome, and study type. The search terms were combinations of the keywords listed under categories 1, 2, and 3.

Screening and Data Extraction

Three of the coauthors (SH, GL, and TJN) independently completed the title and abstract screening of the retrieved study records, as well as the full-text reviews of the eligible studies. The web-based platform Covidence (SaaS Enterprise) was used in managing the screening process. The corresponding author resolved any screening disparities.

Data were extracted through Covidence following the same template for each study across three domains: (1) study characteristics (ie, study title, authors, publication year, country, study type, health care settings, recruitment, sample size and demographics, and inclusion and exclusion criteria); (2) intervention characteristics (ie, therapy type, components, duration, and outcome measurements); and (3) study results (ie, primary outcome: the reduction in depression and anxiety symptoms, secondary outcome: user engagement, intervention credibility, and usability).

Data Synthesis

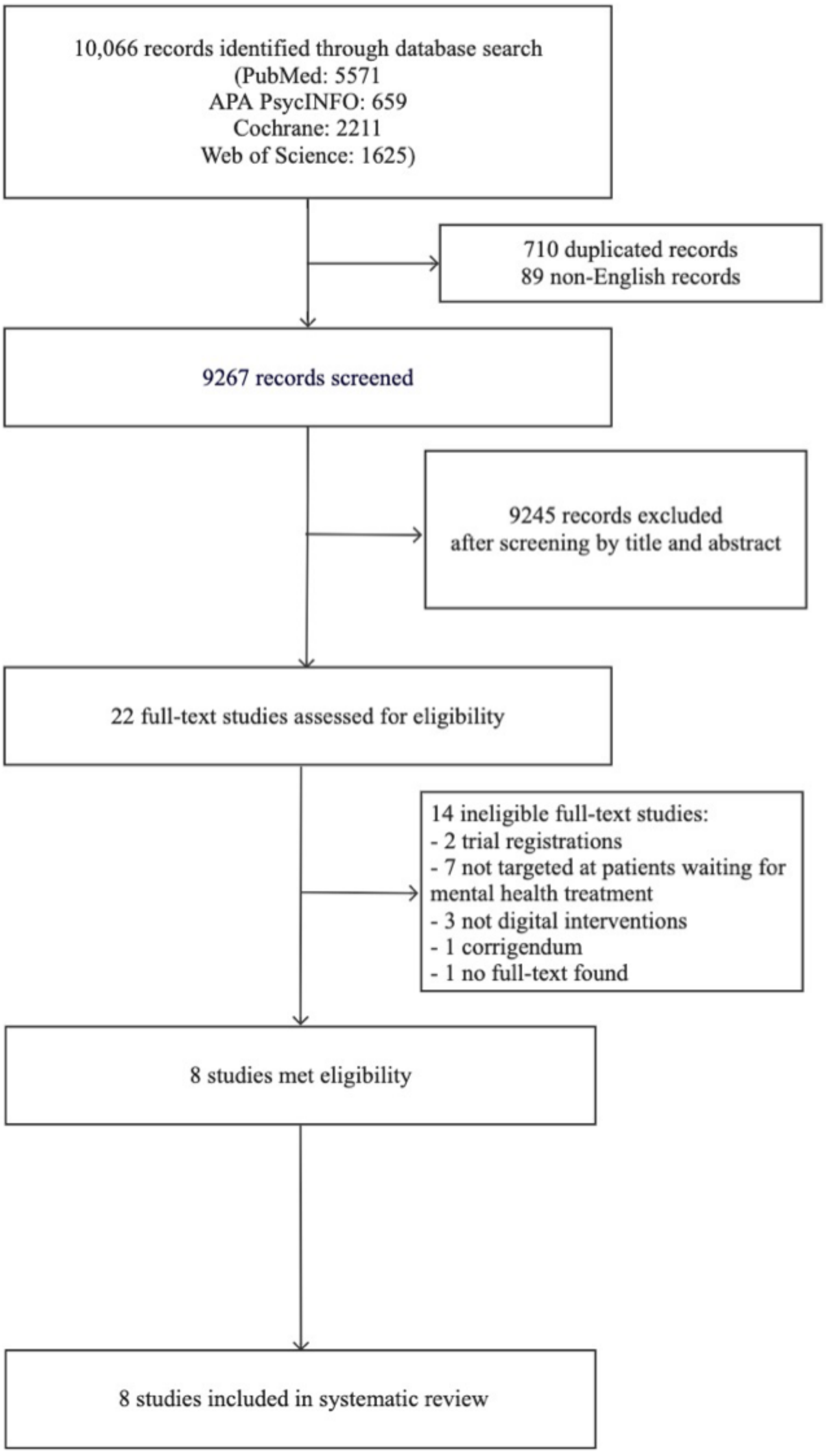
The extracted data were first synthesized by the study characteristics and the features of interventions including sample size, recruitment methods, types of psychotherapy adapted, delivery formats, and outcome measures. Next, we synthesized the effectiveness of the digital intervention by analyzing the reduction of depression and anxiety symptoms, separately for different study types (ie, RCTs and non-RCTs). Finally, reports about user engagement and perceived credibility and usability were synthesized.

Results

Overview

Of the 10,066 study records retrieved from the databases, and after removing duplicates and non-English studies, a total of 9267 unique titles and abstracts were screened. Following the screening, a final 22 studies remained for full-text screening (Figure 1). After scanning the full texts, 8 studies met eligibility criteria and were included in the review.

Figure 1. PRISMA flow diagram outlining data set and study identification. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.



Study Characteristics

The study characteristics are shown in Table 2. In all the studies, the participants were informed about the digital intervention by their health care service providers and were recruited if they were interested in participating and passed screening. Five studies used an RCT design, two studies were feasibility studies, and 1 study was a nonrandomized experimental study. Studies by Kenter et al [11] and Kolovos et al [36] used the same data collection for analysis and were therefore discussed together in the current review.

Table 2. Study characteristics.

Reference, study type, and location	Setting	Recruitment	Sample size	Demographics	Inclusion and exclusion criteria
Krämer et al (2021) [37], RCT[a], Germany	Outpatient clinics	Individuals were informed about this study and provided with a screening questionnaire.	136	Age: mean 36.3 (SD 11.9); 78% female; 58% in a relationship	Included depressive patients on the waiting lists (age 18 years and older; CES-D[b]>22; BSI[c] item 9≤1). Excluded patients indicating suicidal ideation at baseline.
Villemaire-Krajden and Myhr (2019) [38], RCT, Canada	A university clinical teaching unit in a tertiary care hospital	Participants were recruited from referrals for CBT[d] and instructions were sent by emails.	67	Age: mean 38.21 (SD 12.71); 67% female; 43% in a relationship	Unspecified, but included outpatients with all types of depressive disorders, anxiety, bipolar, insomnia, psychosis, adjustment disorder, personality disorders, and autism spectrum disorder.
Kenter et al (2016) [11], RCT, The Netherlands	Outpatient clinics	Recruited patients after their registration for regular face-to-face treatments. Screened by telephone.	269	Age: mean 38 (SD 11.4); 53.9% female; 24.2% in a relationship	Included outpatients waiting for face-to-face treatment at clinics with major depressive disorder diagnosed by *DSM-IV*[e] (age 18 years and older, waiting time >8 weeks). Excluded patients with suicidal ideation and antidepressant medication.
Kolovos et al (2016) [36], RCT, The Netherlands	Outpatient clinics	Recruited patients after their registration for regular face-to-face treatments. Screened by telephone.	269	Age: mean 38 (SD 11.4); 53.9% female; 24.2% in a relationship	Included outpatients waiting for face-to-face treatment at clinics with major depressive disorder diagnosed by *DSM-IV* (age 18 years and older, waiting time >8 weeks). Excluded patients with suicidal ideation and antidepressant medication.
Twomey et al (2014) [39], RCT, Ireland	A public health care provider in Dublin	Participants were invited by telephone, and instructions were sent by email.	149	Age: mean 35.3 (SD 10.3); 73.8% female; 41.6% in a relationship	Included adults on waiting lists with symptoms of depression, anxiety, or stress (as shown by their initial referral). Excluded those with psychosis or cognitive impairment.
Hentati et al (2022) [40], feasibility study, Sweden	A routine psychiatric care unit	Participants were informed by the health care personnel at the psychiatric care unit and self-registered digitally.	12	Age: mean 34.3 (SD 11.1); 50% female; 75% in a relationship	Included patients with symptoms of depression or anxiety on the waiting list (age 18 years and older; PHQ-9[f] or GAD-7[g]≥5; MADRS-S[h] item 9<4). Excluded patients with severe suicidal ideation, and psychiatric or somatic difficulties.
Duffy et al (2019) [41], feasibility study, United Kingdom	Outpatient clinics	Participants were informed of the study through their clinician and discussed their participation with clinicians during the assessment appointment.	123	Age: 28% between 17 and 24, 46% between 25 and 44, 26% between 45 and 64, and 1% between 65 and 80; 69% female	Included clients with severe presentations of anxiety or depression, or both, requiring high-intensity treatment. Excluded clients requiring low-intensity treatment and those with substance abuse.
Whitfield et al (2006) [42], nonrandomized experimental study, United Kingdom	Outpatient clinics	Referred participants were invited to a screening appointment.	20	Age: mean 38.05 (SD 12.98); 45% female	Included patients with problems of depression and depression with anxiety (aged between 16 and 65, BDI-II[i]<2). Excluded patients with current active suicidal intent and psychosis.

[a]RCT: randomized control trial.

[b]CES-D: Center for Epidemiologic Studies Depression Scale [43].

[c]BSI: brief symptom inventory [44].

[d]CBT: cognitive behavioral therapy.

[e]*DSM-IV: Diagnostic and Statistical Manual of Mental Disorders* (4th Edition).

[f]PHQ-9: Patient Health Questionnaire-9 [45].

[g]GAD-7: Generalized Anxiety Disorder-7 [46].

[h]MADRS-S: Montgomery-Åsberg Depression Rating Scale - Self Assessment [47].

[i]BDI-II: Beck Depression Inventory Version II [48].

Among the 5 RCTs, only the intervention group in Krämer et al [37] reported a significant decrease in depressive symptoms, as well as improvements in psychological symptoms and quality of life, compared to the control group. In the other 4 RCTs, 3 RCTs provided the control groups with a self-help book and did not find the internet-based interventions more effective in reducing depression and anxiety symptoms than the control groups, and 1 RCT did not provide any intervention for the control group [11,36,38,39]. The 3 non-RCT studies, which did not have a control group, reported mixed findings. Two studies found that the effectiveness of the digital interventions was beneficial and reduced depression and anxiety symptoms during the waiting time and at follow-up [41,42]; however, after the intervention in the remaining study, only a few participants reached the threshold for clinical improvement in their symptoms of depression or anxiety [40].

Quality Assessment

We used the RoB 2 tool (a revised Cochrane risk of bias tool for randomized trials) to assess the risk of bias of the included RCTs. The RoB 2 evaluates five domains, including the randomization process, deviations from intended interventions (effect of assignment to intervention), missing outcome data, measurement of the outcome, and selection of the reported result [49]. In addition, the ROBINS-I (Risk of Bias in Nonrandomized Studies-of Interventions) tool by Cochrane was used to assess the risk of bias of the included non-RCTs. The ROBINS-I evaluates seven domains including confounding, selection of participants, classification of interventions, deviations from intended interventions, missing data, measurement of the outcomes, and selection of the reported results [50].

Out of the 5 included RCTs, 4 RCTs had low risk and 1 RCT had some concern with regard to the randomization process (due to significant baseline differences between the two groups). Two studies had low risk and three studies had some concerns regarding low intervention adherence. Regarding missing data, 2 studies had low risk and 3 studies had moderate to high risk due to dropouts. All 5 studies had some concerns due to it being unclear if the assessment method was double-blinded or not. Finally, all studies had a low risk of bias regarding the selection of the reported result. A detailed quality assessment for each study is presented in Table S1 in Multimedia Appendix 2.

The 3 included non-RCTs all had moderate risks. A common serious risk factor among all studies was the consideration of confounders. None of the studies reported potential confounders, which contributed to the potential bias of the study results. In addition, intervention adherence was a concern in all studies. Due to the moderate to high dropout rates, the handling of the missing values was problematic in two of the three studies. Finally, similar to RCTs, the 3 non-RCTs also only included self-reported outcome measurements, and it was unclear if outcome assessors were aware of the intervention received by study participants. A detailed quality assessment for each study is presented in Table S2 in Multimedia Appendix 2.

Randomized Controlled Trials

The intervention characteristics of RCTs are shown in Tables 3 and 4. Krämer et al [37] found that the "GET.ON Mood Enhancer" that they used was effective in reducing depressive symptoms. The intervention group improved significantly after the intervention compared to the control group ($F_{2,121.5}$=3.91; P<.05; see mean, SD, and between-group effect size in Table 5). The control group remained on the waiting list and did not receive any intervention during the study period. These differences were also maintained at the 5-month follow-up, where the depressive symptoms were significantly reduced in the intervention group compared to the control group. In addition, Krämer et al [37] reported a significant difference in the decrease of both psychological symptoms ($F_{2,119.4}$=4.37; P<.05) and mental health quality of life ($F_{2,119.7}$=3.36; P<.05), indicating that the intervention group improved more than controls.

Table 3. Intervention characteristics of RCT[a] studies.

Reference	Therapy	Delivery formats	Components	Duration and frequency
Krämer et al (2021) [37]	CBT[b]	Internet-based platform "GET.ON Mood Enhancer"	Six consecutive modules of 45 minutes, written semistandardized feedback by an e-coach (trained psychologist), homework assignments, a digital mood diary, and an optional text message coach	7 weeks
Villemaire-Krajden and Myhr (2019) [38]	CBT	Internet-based website "Good Days Ahead" accessible via computer	Nine lessons of 30 minutes, explanations for the CBT model and its application, vignettes presenting case examples, quizzes, exercises, introduction to a few behavioral techniques such as activity monitoring and scheduling	9 weekly lessons
Kenter et al (2016) [11]	PST[c]	Internet-based platform, email reminders if not finishing a session	Five sessions with structured homework assignments, adopting a structured 6-step approach: identifying the problem, finding solutions, selecting one solution, creating a plan to solve the problem with this solution, executing the plan, and evaluating the plan, plus web-based feedback by a coach	5 weekly sessions
Kolovos et al (2016) [36]	PST	Internet-based platform, email reminders if not finishing a session	Five sessions with structured homework assignments, adopting a structured 6-step approach: identifying the problem, finding solutions, selecting one solution, creating a plan to solve the problem with this solution, executing the plan, and evaluating the plan, plus web-based feedback by a coach	5 weekly sessions
Twomey et al (2014) [39]	CBT	Internet-based platform "MoodGYM," weekly automated email reminders	A brief introductory session and five 20- to 40-minute sessions, containing written information, animations, human-centered exercises, and quizzes	32 days

[a]RCT: randomized controlled trial.

[b]CBT: cognitive behavioral therapy.

[c]PST: problem-solving therapy.

Table 4. Control group and outcome measures of RCT[a] studies.

Reference	Control group	Measurement		
		Depression	Anxiety	Others
Krämer et al (2021) [37]	Access to treatment as usual was not restricted, participants may be seeing a general practitioner, psychiatrist, psychotherapist, using self-help, or other inpatient or outpatient practitioner	PHQ-9[b]	Not measured	Psychological symptoms (BSI[c]), quality of life—mental and physical (SF-12[d]), need and motivation for psychotherapy, attitude toward F2F psychotherapy (ATSPPH[e]), attitude toward web-based interventions (adapted ATSPPH), negative effects (INEP[f]), outpatient psychotherapy history
Villemaire-Krajden and Myhr (2019) [38]	Provided with a workbook	BDI-II[g]	BAI[h]	Subjective well-being, problem or symptoms, life functioning difficulties, and risk or harm to self or others (CORE-OM[i])
Kenter et al (2016) [11]	Received a self-help book without any guidance or further instructions	CES-D[j]	HADS-A[k]	Insomnia (ISI[l]), self-rated health (EQ-5D [51]), perceived control (PMS[m])
Kolovos et al (2016) [36]	Received a self-help book without any guidance or further instructions	CES-D	Not measured	Response to treatment and quality of life (EQ-5D)
Twomey et al (2014) [39]	No intervention at all	DASS[n]-Depression	DASS-Anxiety	DASS-Stress, work and social functioning (WSAS[o])

[a]RCT: randomized controlled trial.
[b]PHQ-9: Patient Health Questionnaire-9 [45].
[c]BSI: brief symptom inventory [44].
[d]SF-12: 12-Item Short-Form Health Survey [52].
[e]ATSPPH: Attitude Toward Seeking Professional Psychological Help Scale [53].
[f]INEP: inventory for the assessment of negative effects of psychotherapy [54].
[g]BDI-II: Beck Depression Inventory Version II [48].
[h]BAI: Beck Anxiety Inventory [55].
[i]CORE-OM: Clinical Outcomes in Routine Evaluation Outcome Measure [56].
[j]CES-D: Center for Epidemiologic Studies Depression Scale [43].
[k]HADS-A: Hospital Anxiety and Depression Scale Anxiety subscale [57].
[l]ISI: Insomnia Severity Index questionnaire [58].
[m]PMS: Pearlin Mastery Scale [59].
[n]DASS-21: Depression, Anxiety, and Stress Scale-21 [60].
[o]WSAS: Work and Social Adjustment Scale [61].

Table 5. Statistical results of reviewed studies.

Reference	Depression symptom, mean (SD)		Anxiety symptom, mean (SD)		Between-group effect size: interaction effect, time × group
	Intervention	Control	Intervention	Control	
Krämer et al (2021) [37]	Before treatment (ITT[a]): 14.6 (0.5); posttreatment (ITT): 10.5 (0.7); five-month follow-up (ITT): 9.9 (0.7)	Before treatment (ITT): 15.7 (0.5); posttreatment (ITT): 13.3 (0.6); five-month follow-up (ITT): 12.7 (0.7)	Not measured	Not measured	Posttreatment: P=.004, Cohen d=0.55; five month follow up: P=.005, Cohen d=0.52
Villemaire-Krajden and Myhr (2019) [38]	Before treatment: 25.50 (11.94); posttreatment: 24.14 (12.50)	Before treatment: 29.27 (11.85); posttreatment: 28.59 (14.70)	Before treatment: 22.33 (14.38); posttreatment: 20.00 (13.42)	Before treatment: 19.86 (10.61); posttreatment: 18.32 (11.65)	P>.05, partial eta squared=0 for both depressive and anxiety symptoms
Kenter et al (2016) [11] and Kolovos et al (2016) [36]	Before treatment: 37.0 (11.6); posttreatment: 27.0 (15.1)	Before treatment: 35.2 (12.1); posttreatment: 25.9 (14.9)	Before treatment: 12.4 (3.9); posttreatment: 10.5 (5.4)	Before treatment: 12.6 (4.6); posttreatment: 10.0 (5.5)	Depression: Cohen d=0.07; Anxiety: Cohen d=0.09
Twomey et al (2014) [39]	Before treatment: 18.79 (11.46); posttreatment: 14.29 (10.62)	Before treatment: 16.26 (9.71); posttreatment: 16.26 (10.52)	Before treatment: 12.93 (8.06); posttreatment: 11.71 (8.91)	Before treatment: 10.32 (7.14); posttreatment: 11.79 (9.56)	Depression: Cohen d=0.19; Anxiety: Cohen d=0.01
Hentati et al (2022) [40]	No data provided	N/A[b]	No data provided	N/A	N/A
Duffy et al (2019) [41]	Before-posttreatment: decrease 3.6 points	N/A	Before-posttreatment: decrease 3.2 points	N/A	N/A
Whitfield et al (2006) [42]	Before treatment (ITT): 28.15 (11.41); posttreatment (ITT): 20.00 (10.41); three-month follow-up (ITT): 18.95 (10.41)	N/A	Before treatment (ITT): 20.30 (11.23); posttreatment (ITT): 14.55 (7.82); three-month follow-up (ITT): 9.90 (8.47)	N/A	N/A

[a]ITT: intention-to-treat.

[b]Not applicable.

Villemaire-Krajden and Myhr [38] investigated the effectiveness of using computerized CBT to decrease symptoms of distress in outpatients during a waiting period for services. They did not find significant interactions between the intervention and control group, with a very small between-group effect size (Table 5). This indicates that compared to the control group, the intervention group did not score significantly better with regard to well-being, symptom severity, functioning, or motivation for CBT. In addition, depressive symptoms, anxiety symptoms, and life functioning difficulties were not found to have changed significantly between groups from baseline to the postintervention assessment. However, for both groups, the well-being (measured by Clinical Outcomes in Routine Evaluation Outcome Measure) of the participants significantly increased ($F_{1,37}$=6.31; P<.05), and the symptoms significantly decreased ($F_{1,37}$=8.74; P<.05) over time.

Both Kenter et al [11] and Kolovos et al [36] investigated the effectiveness of a web-based PST, self-help intervention in reducing depression symptoms. The intervention examined by Kenter et al [11] was effective in reducing depression and anxiety symptoms, but in comparison with the control group (who were provided with a self-help book), it was not more effective. They used intention-to-treat (ITT) analysis and found a significant reduction in depressive symptoms in both the intervention group (see Table 5, within-group effect size: Cohen d=0.75) and the control group (within-group effect size: Cohen d=0.69) from baseline to posttreatment. Despite the moderate within-group effect size for both groups, the between-group effect size was small (Table 5). The between-group difference posttreatment was not significant in the regression analysis. Similarly, for anxiety, Kenter et al [11] reported moderate within-group effect sizes (intervention group: Cohen d=0.41; control group: Cohen d=0.52) and small between-group effect size (Table 5), and they found no significant difference between the intervention group and the control group. The clinical outcomes in Kolovos et al [36] were the same as Kenter et al [11] but Kolovos et al [36] focused on estimating the cost-effectiveness of the intervention in comparison with enhanced usual care. Kolovos et al [36] found that from a societal perspective, the intervention was not cost-effective in comparison with the enhanced usual care.

Twomey et al [39] examined the effectiveness of a self-help CBT intervention ("MoodGYM") on reducing general psychological distress, stress, depression, anxiety, and impaired daily functioning symptoms. Twomey et al [39] found symptom improvement for the treated individuals with a within-group effect size of Cohen d=0.4 for depression and Cohen d=0.14 for anxiety. Nevertheless, similar to the studies conducted by Kenter et al [11] and Kolovos et al [36] they did not find significant differences between the intervention group and the simply waiting control group (depression: $F_{1,64}$=2.99; P>.05; anxiety: $F_{1,64}$=1.72; P>.05) [39]. At posttreatment, small between-group effect sizes were found for both depression and anxiety (Table 5), suggesting that "MoodGYM" was not more effective than the control condition.

Non-RCTs

Intervention characteristics of non-RCTs are shown in Table 6. Hentati et al [40] tested the effectiveness of a self-guided PST intervention. The results showed a 16% and 22% median for symptom improvement in depression and anxiety, respectively, from screening to posttreatment. However, from pretreatment to posttreatment, the improvement in depression was reduced by 3%, while anxiety symptoms were reduced by 3%. Hentati et al [40] concluded that there was insufficient evidence for the effectiveness of this digital intervention.

Table 6. Intervention characteristics of non-RCT[a] studies.

Reference	Therapy	Delivery formats	Components	Duration and frequency	Measurement		
					Depression	Anxiety	Others
Hentati et al (2022) [40]	CBT[b]	Internet-based platform accessible via computer and mobile devices	Psychoeducational texts, treatment rationale, examples of problems and suggestions of solutions, illustrative pictures, instructions, and problem-solving exercises	4 weeks	PHQ-9[c]	GAD-7[d]	Treatment credibility (CEQ[e]), usability (SUS[f]), behavioral engagement, suicidal ideation (MADRS-S[g]), negative effects (NEQ[h]), post-study questionnaire on user experience
Duffy et al (2019) [41]	CBT	Internet-based platform "SilverCloud"	Eight modules including tools such as self-monitoring and thought recording, behavioral activation, cognitive restructuring, and challenging core beliefs	47 days, one review every 10 to 12 days.	PHQ-9	GAD-7	Work and social functioning (WSAS[i])
Whitfield et al (2006) [42]	CBT	Computerized CD Rom	Six sessions of 45-60 minutes, including text, cartoon illustrations, animations, human-centered questions, sound, and video, as well as the offer of short support sessions by a self-help support psychiatric nurse	6 weekly sessions	BDI-II[j]	BAI[k]	Hopelessness (BHS[l]), social adaptation self-evaluation (SASS[m])

[a]RCT: randomized controlled trial.

[b]CBT: cognitive behavioral therapy.

[c]PHQ-9: Patient Health Questionnaire-9 [45].

[d]GAD-7: Generalized Anxiety Disorder-7 [46].

[e]CEQ: Credibility/Expectancy Questionnaire [62].

[f]SUS: System Usability Scale [63].

[g]MADRS-S: Montgomery-Åsberg Depression Rating Scale - Self Assessment [47].

[h]NEQ: Negative Effects Questionnaire [64].

[i]WSAS: Work and Social Adjustment Scale [61].

[j]BDI-II: Beck Depression Inventory Version II [48].

[k]BAI: Beck Anxiety Inventory [55].

[l]BHS: Beck Hopelessness Scale [65].

[m]SASS: Social Adaptation Self-evaluation Scale [66].

Duffy et al [41] examined the outcomes of using iCBT as a prequel for patients requiring high-intensity treatment. The results indicated that 58% exhibited reliable improvement from baseline to iCBT exit and around 20% of the sample achieved reliable recovery in advance of starting face-to-face therapy. This was demonstrated by the decrease in depression (within-group effect size: Cohen d=0.61) and anxiety (within-group effect size: Cohen d=0.69). In addition, the decrease in the WSAS score, which estimates the severity of work and social adjustment impairment, followed a similar pattern with a substantial reduction (mean 2.4, SD 8.7 points) in the severity score at iCBT exit (within-group effect size: Cohen d=0.31).

Finally, Whitfield et al [42] investigated the effectiveness of the compact disc CBT self-help intervention. They adopted an ITT analysis where the last observation for the participant was carried forward if data were missing due to questionnaire nonresponse. The ITT analysis could possibly reduce any potential bias by accounting for the situation of those who had dropped out of the study. Although they did not calculate the effect sizes, the mean BDI-II scores for depression symptoms significantly decreased (t_{19}=4.91; P<.001), and the mean BAI score for anxiety symptoms significantly decreased as well (t_{19}=2.51; P=.02; Table 5). Furthermore, Whitfield et al [42] found the effectiveness of the intervention persisted at the 3-month follow-up.

Summary

To summarize, there is some evidence of the effectiveness of digital interventions to improve depression and anxiety symptoms in individuals waiting for psychotherapy. Seven of the eight studies reported digital interventions that led to a reduction in depressive and anxiety symptoms, with, on average, a moderate within-group effect size. However, only one study found that the waiting list intervention was effective in comparison with the control group. Moreover, due to the small between-group effect sizes in four of the five RCTs and the small sample sizes in the non-RCTs, it is difficult to conclude that the waiting list interventions were effective.

The secondary outcomes are shown in Table S3 in Multimedia Appendix 2. In brief, the reviewed studies showed that the user engagement of the self-guided digital interventions was low, but the guided digital interventions appeared more engaging compared to the unguided ones. The low engagement and high dropout rates may have contributed to the interventions being less effective. The interventions were generally perceived as credible and useful, with moderate to high user satisfaction, suggesting that digital interventions have the potential to be adapted as a pretreatment for patients on waiting lists.

Discussion

Principal Findings

This systematic review identified 8 studies of digital interventions aimed at reducing the depression and anxiety symptoms of patients waiting for face-to-face psychotherapy. We found that waiting list interventions were effective in reducing symptoms of depression and anxiety, but not more effective compared to simply waiting for treatment or control groups who used a self-help book. Although the digital interventions were perceived as credible and useful, low user engagement was a major concern for treatment adherence and effectiveness.

Effectiveness and Features of the Waiting List Digital Interventions

We found that the DMHIs implemented with patients on waiting lists reduced depression and anxiety symptoms with moderate effect sizes. These effect sizes were similar to those of prior meta-analyses on the effectiveness of DMHIs in general by Firth et al [22] and Moshe et al [28]. Nevertheless, this review did not find significant differences between the intervention group and the control group. Several factors might have reduced symptoms in the control groups. First, in the RCTs conducted by Kenter et al [11] and Kolovos et al [36], the control group was provided with a self-help book instead of merely waiting; the aim was to increase the participation rate. Even with small effect sizes, previous studies have shown that self-help books were effective in reducing depression symptoms [8]. In addition, some spontaneous improvements were found among the waiting control group, as demonstrated in Twomey et al [39]. This is in line with previous studies that show that spontaneous improvement does occur among some patients with depression and those with anxiety [16,67].

Concerning the features of the DMHIs, there was no difference between DMHIs implemented for patients specifically on waiting lists and general DMHIs. In addition, across the reviewed 8 studies, we did not find evidence of a specific technological platform (ie, using phone vs computer) or a specific psychological therapy (ie, CBT vs PST) having better effectiveness in DMHIs, which is consistent with previous studies [8,28].

While DMHIs have the potential to provide more scalable, more accessible, and more cost-efficient treatments compared to face-to-face therapies [23,24], it is important to acknowledge that DMHIs may not be suitable to address some patients' need for face-to-face psychotherapy, especially for those with severe symptoms [11,36]. Due to these concerns, Twomey et al [39] and Krämer et al [37] recommended that digital interventions should not be provided as a substitute for face-to-face psychotherapy but were better suited to being an additional or complementary treatment option that could bridge the waiting time. This is consistent with Cornish's [68] stepped-care model, where patients are able to receive timely digital interventions on the same day that they seek care, balancing out the treatment intensity and resources available.

User Engagement

As identified by earlier studies, guided interventions are more engaging compared to unguided ones [31,32], and general DMHIs implemented during the waiting time exhibited lower user engagement and treatment adherence compared to face-to-face psychotherapies [28-30]. In this review, Krämer et al [37] show that there was a lower dropout rate and higher treatment adherence compared to the other self-guided studies. This might be explained by the participants' demographics. For example, Twomey et al [39] reported that although 85% of the males dropped out of the study, the number of females that dropped out was only 8%. In contrast, Kenter et al [11] and Kolovos et al [36] reported that 64.7% of the participants who did not complete the study were female. They also found that participants who were younger, less educated, and had lower incomes were more likely to drop out [11,36]. In addition, as Farrington et al [69] suggest, technical difficulties experienced by certain age groups (ie, older people) and incompatibility between different digital device systems may also have been barriers to user engagement.

Another factor for low user engagement might be related to how the digital intervention was promoted. In all the studies reviewed, the interventions were presented to the participants as an additional or temporary treatment rather than a replacement for the face-to-face psychotherapy they were waiting for. Promoting the interventions in this way might have lowered the expectancy of participants as regards the effect of the treatment and they may also have perceived it as less important. For instance, Hentati et al [40] reported that the participants were not convinced that the digital intervention would result in any major reduction in their symptoms although they perceived the intervention as credible. Kenter et al [11] found that the completion rate was positively associated with participants' expectancy and perceived credibility of the intervention. Therefore, changing the way digital interventions are promoted may increase the engagement rate.

Limitations and Future Studies

Due to the understudied nature of the topic, the first limitation of this systematic review was the restricted number of randomized control studies (n=5), which reduces the possibility of drawing a robust conclusion concerning the effectiveness of waiting list interventions. In addition, according to the quality assessment, the reviewed studies had some shared concerns about using self-report data and missing values due to dropouts. In future studies, once a sufficient number of high-quality studies with adequate sample sizes are available, conducting a meta-analysis will be crucial to provide a more comprehensive synthesis of the evidence in this area. Second, various instruments were adopted to measure depression and anxiety symptoms, posing challenges when controlling for baseline differences and when comparing the effectiveness of the studies. Moreover, the psychological outcomes were self-reported in most studies, and these might be less sensitive to symptom changes compared to a clinician's assessment [37]. Third, this study was limited to patients with depression and those with anxiety with relatively stable conditions. Patients with severe suicidal ideation and psychosis were excluded, even though these groups are at a higher risk for symptom deterioration while waiting for psychotherapy. Ketner et al [11] reported one case of suicide in the control group, suggesting that future waiting list interventions could target individuals with less stable conditions. Additionally, when considering future studies, we recommend the use of multiple research protocols for RCTs in the field of digital interventions for patients on waiting lists that target a wider range of symptoms (ie, phobic disorders and eating disorders) [70-72]. Moreover, we recommend that researchers adopt more therapies (ie, exposure therapy) and include larger samples [73,74]. Finally, it is also important for future studies to consider the use of digital interventions in different populations. We acknowledge that this review only included English studies, which may have missed valuable studies conducted in non-English speaking countries. The reviewed studies in this study were also conducted in high-income Western countries. It is unclear whether resource-limited areas may benefit more from low-cost, web-based interventions for mental health issues.

Conclusions

As a potential solution to the problem of prolonged waiting lists for psychotherapy, this systematic review examined the effectiveness of DMHIs in reducing anxiety and depression symptoms of patients on waiting lists for psychotherapy. Our results showed that among the RCTs, DMHIs were overall not more effective when compared with simply waiting or the control groups who used a self-help book. Among the non-RCTs, although the intervention reduced depression and anxiety symptoms, the study design with no control group made it difficult to conclude the effectiveness of the intervention. Waiting list DMHIs may prove to be an adequate treatment for some patients, but low user engagement remains a concern for treatment adherence and effectiveness. Despite patients rating DMHIs as credible and useful, it is nevertheless evident that as yet DMHIs might not fully replace face-to-face psychotherapies for all patients.

Authors' Contributions

All authors designed the study protocol together, contributed to writing, as well as reviewed and revised the final manuscript. SH, YW, GL, and TJN conducted the literature search, data extraction, and synthesized the data. SH and YW made Figure 1. SH and TJN made all the tables. TJN supervised the research project.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 1076 KB - mental_v11i1e56650_app1.pdf]

Multimedia Appendix 2
Quality assessments and secondary outcomes.
[DOCX File , 22 KB - mental_v11i1e56650_app2.docx]

References

1. Patel V, Saxena S, Frankish H, Boyce N. Sustainable development and global mental health—a Lancet commission. Lancet 2016;387(10024):1143-1145. [doi: 10.1016/S0140-6736(16)00208-7] [Medline: 27025320]
2. GBD 2019 DiseasesInjuries Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the global burden of disease study 2019. Lancet 2020;396(10258):1204-1222 [FREE Full text] [doi: 10.1016/S0140-6736(20)30925-9] [Medline: 33069326]
3. Kola L, Kohrt BA, Hanlon C, Naslund JA, Sikander S, Balaji M, et al. COVID-19 mental health impact and responses in low-income and middle-income countries: reimagining global mental health. Lancet Psychiatry 2021;8(6):535-550 [FREE Full text] [doi: 10.1016/S2215-0366(21)00025-0] [Medline: 33639109]
4. COVID-19 Mental Disorders Collaborators. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet 2021;398(10312):1700-1712 [FREE Full text] [doi: 10.1016/S0140-6736(21)02143-7] [Medline: 34634250]
5. Torales J, O'Higgins M, Castaldelli-Maia JM, Ventriglio A. The outbreak of COVID-19 coronavirus and its impact on global mental health. Int J Soc Psychiatry 2020;66(4):317-320. [doi: 10.1177/0020764020915212] [Medline: 32233719]
6. DeRubeis RJ, Siegle GJ, Hollon SD. Cognitive therapy versus medication for depression: treatment outcomes and neural mechanisms. Nat Rev Neurosci 2008;9(10):788-796 [FREE Full text] [doi: 10.1038/nrn2345] [Medline: 18784657]
7. Bai Z, Luo S, Zhang L, Wu S, Chi I. Acceptance and commitment therapy (ACT) to reduce depression: a systematic review and meta-analysis. J Affect Disord 2020;260:728-737. [doi: 10.1016/j.jad.2019.09.040] [Medline: 31563072]
8. Cuijpers P, Berking M, Andersson G, Quigley L, Kleiboer A, Dobson KS. A meta-analysis of cognitive-behavioural therapy for adult depression, alone and in comparison with other treatments. Can J Psychiatry 2013;58(7):376-385 [FREE Full text] [doi: 10.1177/070674371305800702] [Medline: 23870719]
9. Mental Health Gap Action Programme (mhGAP) Guideline for Mental, Neurological and Substance Use Disorders [Internet]. Geneva: World Health Organization; 2023.
10. Ruesch M, Helmes A, Bengel J. Cognitive behavioral group therapy for patients with physical diseases and comorbid depressive or adjustment disorders on a waiting list for individual therapy: results from a randomized controlled trial. BMC Psychiatry 2017;17(1):340 [FREE Full text] [doi: 10.1186/s12888-017-1494-9] [Medline: 29017481]
11. Kenter RMF, Cuijpers P, Beekman A, van Straten A. Effectiveness of a web-based guided self-help intervention for outpatients with a depressive disorder: short-term results from a randomized controlled trial. J Med Internet Res 2016;18(3):e80 [FREE Full text] [doi: 10.2196/jmir.4861] [Medline: 27032449]
12. Hongkong Hospital Authority. URL: https://www.ha.org.hk/visitor/sopc_waiting_time.asp?id=7&lang=ENG [accessed 2023-12-13]
13. Gagliardi AR, Yip CYY, Irish J, Wright FC, Rubin B, Ross H, et al. The psychological burden of waiting for procedures and patient-centred strategies that could support the mental health of wait-listed patients and caregivers during the COVID-19 pandemic: a scoping review. Health Expect 2021;24(3):978-990 [FREE Full text] [doi: 10.1111/hex.13241] [Medline: 33769657]
14. Reichert A, Jacobs R. The impact of waiting time on patient outcomes: evidence from early intervention in psychosis services in England. Health Econ 2018;27(11):1772-1787 [FREE Full text] [doi: 10.1002/hec.3800] [Medline: 30014544]
15. Furukawa TA, Noma H, Caldwell DM, Honyashiki M, Shinohara K, Imai H, et al. Waiting list may be a nocebo condition in psychotherapy trials: a contribution from network meta-analysis. Acta Psychiatr Scand 2014;130(3):181-192. [doi: 10.1111/acps.12275] [Medline: 24697518]
16. Whiteford HA, Harris MG, McKeon G, Baxter A, Pennell C, Barendregt JJ, et al. Estimating remission from untreated major depression: a systematic review and meta-analysis. Psychol Med 2013;43(8):1569-1585. [doi: 10.1017/S0033291712001717] [Medline: 22883473]
17. Grünzig SD, Bengel J, Göhner W, Violetta Krämer L. Low-intensity interventions to reduce depressive symptoms before outpatient psychotherapy—a systematic literature review. Psychother Psychosom Med Psychol 2019;69(6):212-223. [doi: 10.1055/a-0630-2397] [Medline: 30045413]
18. Dao KP, De Cocker K, Tong HL, Kocaballi AB, Chow C, Laranjo L. Smartphone-delivered ecological momentary interventions based on ecological momentary assessments to promote health behaviors: systematic review and adapted checklist for reporting ecological momentary assessment and intervention studies. JMIR Mhealth Uhealth 2021;9(11):e22890 [FREE Full text] [doi: 10.2196/22890] [Medline: 34806995]
19. Saramago P, Gega L, Marshall D, Nikolaidis GF, Jankovic D, Melton H, et al. Digital interventions for generalized anxiety disorder (GAD): systematic review and network meta-analysis. Front Psychiatry 2021;12:726222 [FREE Full text] [doi: 10.3389/fpsyt.2021.726222] [Medline: 34938209]
20. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019;21(7):e12869 [FREE Full text] [doi: 10.2196/12869] [Medline: 31333198]
21. Pauley D, Cuijpers P, Papola D, Miguel C, Karyotaki E. Two decades of digital interventions for anxiety disorders: a systematic review and meta-analysis of treatment effectiveness. Psychol Med 2023;53(2):567-579 [FREE Full text] [doi: 10.1017/S0033291721001999] [Medline: 34047264]

22. Firth J, Torous J, Carney R, Newby J, Cosco TD, Christensen H, et al. Digital technologies in the treatment of anxiety: recent innovations and future directions. Curr Psychiatry Rep 2018;20(6):44 [FREE Full text] [doi: 10.1007/s11920-018-0910-2] [Medline: 29779065]
23. Hollis C, Morriss R, Martin J, Amani S, Cotton R, Denis M, et al. Technological innovations in mental healthcare: harnessing the digital revolution. Br J Psychiatry 2015;206(4):263-265 [FREE Full text] [doi: 10.1192/bjp.bp.113.142612] [Medline: 25833865]
24. Murray E, Hekler EB, Andersson G, Collins LM, Doherty A, Hollis C, et al. Evaluating digital health interventions: key questions and approaches. Am J Prev Med 2016;51(5):843-851 [FREE Full text] [doi: 10.1016/j.amepre.2016.06.008] [Medline: 27745684]
25. Patel S, Akhtar A, Malins S, Wright N, Rowley E, Young E, et al. The acceptability and usability of digital health interventions for adults with depression, anxiety, and somatoform disorders: qualitative systematic review and meta-synthesis. J Med Internet Res 2020;22(7):e16228 [FREE Full text] [doi: 10.2196/16228] [Medline: 32628116]
26. Hollis C, Falconer CJ, Martin JL, Whittington C, Stockton S, Glazebrook C, et al. Annual research review: digital health interventions for children and young people with mental health problems—a systematic and meta-review. J Child Psychol Psychiatry 2017;58(4):474-503. [doi: 10.1111/jcpp.12663] [Medline: 27943285]
27. Alkhaldi G, Hamilton FL, Lau R, Webster R, Michie S, Murray E. The effectiveness of technology-based strategies to promote engagement with digital interventions: a systematic review protocol. JMIR Res Protoc 2015;4(2):e47 [FREE Full text] [doi: 10.2196/resprot.3990] [Medline: 25921274]
28. Moshe I, Terhorst Y, Philippi P, Domhardt M, Cuijpers P, Cristea I, et al. Digital interventions for the treatment of depression: a meta-analytic review. Psychol Bull 2021;147(8):749-786. [doi: 10.1037/bul0000334] [Medline: 34898233]
29. Weightman M. Digital psychotherapy as an effective and timely treatment option for depression and anxiety disorders: implications for rural and remote practice. J Int Med Res 2020;48(6):300060520928686 [FREE Full text] [doi: 10.1177/0300060520928686] [Medline: 32527170]
30. Fuhr K, Schröder J, Berger T, Moritz S, Meyer B, Lutz W, et al. The association between adherence and outcome in an internet intervention for depression. J Affect Disord 2018;229:443-449. [doi: 10.1016/j.jad.2017.12.028] [Medline: 29331706]
31. Musiat P, Johnson C, Atkinson M, Wilksch S, Wade T. Impact of guidance on intervention adherence in computerised interventions for mental health problems: a meta-analysis. Psychol Med 2022;52(2):229-240. [doi: 10.1017/S0033291721004621] [Medline: 34802474]
32. van Ballegooijen W, Cuijpers P, van Straten A, Karyotaki E, Andersson G, Smit JH, et al. Adherence to internet-based and face-to-face cognitive behavioural therapy for depression: a meta-analysis. PLoS One 2014;9(7):e100674 [FREE Full text] [doi: 10.1371/journal.pone.0100674] [Medline: 25029507]
33. Smeets RJEM, Beelen S, Goossens MEJB, Schouten EGW, Knottnerus JA, Vlaeyen JWS. Treatment expectancy and credibility are associated with the outcome of both physical and cognitive-behavioral treatment in chronic low back pain. Clin J Pain 2008;24(4):305-315. [doi: 10.1097/AJP.0b013e318164aa75] [Medline: 18427229]
34. GBD 2015 DiseaseInjury IncidencePrevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet 2016;388(10053):1545-1602 [FREE Full text] [doi: 10.1016/S0140-6736(16)31678-6] [Medline: 27733282]
35. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
36. Kolovos S, Kenter R, Bosmans J, Beekman A, Cuijpers P, Kok R, et al. Economic evaluation of internet-based problem-solving guided self-help treatment in comparison with enhanced usual care for depressed outpatients waiting for face-to-face treatment: a randomized controlled trial. J Affect Disord 2016;200:284-292. [doi: 10.1016/j.jad.2016.04.025] [Medline: 27155071]
37. Krämer LV, Grünzig SD, Baumeister H, Ebert DD, Bengel J. Effectiveness of a guided web-based intervention to reduce depressive symptoms before outpatient psychotherapy: a pragmatic randomized controlled trial. Psychother Psychosom 2021;90(4):233-242. [doi: 10.1159/000515625] [Medline: 33946072]
38. Villemaire-Krajden R, Myhr G. Evaluating the use of a computerized CBT program for outpatients on a waitlist in a university CBT unit. J Psychiatr Pract 2019;25(4):268-278. [doi: 10.1097/PRA.0000000000000396] [Medline: 31291207]
39. Twomey C, O'Reilly G, Byrne M, Bury M, White A, Kissane S, et al. A randomized controlled trial of the computerized CBT programme, MoodGYM, for public mental health service users waiting for interventions. Br J Clin Psychol 2014;53(4):433-450. [doi: 10.1111/bjc.12055] [Medline: 24831119]
40. Hentati A, Forsell E, Ljótsson B, Lindefors N, Kraepelien M. A self-guided and monitored digital problem-solving intervention for patients with symptoms of depression or anxiety on the waiting list for treatment in routine psychiatric care: feasibility study. BJPsych Open 2022;8(2):e43 [FREE Full text] [doi: 10.1192/bjo.2022.14] [Medline: 35130998]
41. Duffy D, Enrique A, Connell S, Connolly C, Richards D. Internet-delivered cognitive behavior therapy as a prequel to face-to-face therapy for depression and anxiety: a naturalistic observation. Front Psychiatry 2019;10:902 [FREE Full text] [doi: 10.3389/fpsyt.2019.00902] [Medline: 31998149]

42. Whitfield G, Hinshelwood R, Pashely A, Campsie L, Williams C. The impact of a novel computerized CBT CD rom (overcoming depression) offered to patients referred to clinical psychology. Behav Cogn Psychother 2005 Nov 22;34(1):1-11. [doi: 10.1017/S135246580500250X]
43. Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. Appl Psychol Meas 2016;1(3):385-401. [doi: 10.1177/014662167700100306]
44. Franke GH, Derogatis LR. Brief Symptom Inventory von L.R. Derogatis ; BSI ; (Kurzform der SCL-90-R) ; deutsche Version. Germany: Beltz Test; 2000.
45. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
46. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
47. Svanborg P, Asberg M. A new self-rating scale for depression and anxiety states based on the comprehensive psychopathological rating scale. Acta Psychiatr Scand 1994;89(1):21-28. [doi: 10.1111/j.1600-0447.1994.tb01480.x] [Medline: 8140903]
48. Beck AT, Steer RA, Brown GK. Manual for the Beck Depression Inventory-II. San Antonio, TX: Psychological Corporation; 1996.
49. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898 [FREE Full text] [doi: 10.1136/bmj.l4898] [Medline: 31462531]
50. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016;355:i4919 [FREE Full text] [doi: 10.1136/bmj.i4919] [Medline: 27733354]
51. EuroQol Group. EuroQol—a new facility for the measurement of health-related quality of life. Health Policy 1990;16(3):199-208. [doi: 10.1016/0168-8510(90)90421-9] [Medline: 10109801]
52. Ware J, Kosinski M, Keller SD. A 12-item short-form health survey: construction of scales and preliminary tests of reliability and validity. Med Care 1996;34(3):220-233. [doi: 10.1097/00005650-199603000-00003] [Medline: 8628042]
53. Fischer EH, Farina A. Attitudes toward seeking professional psychologial help: a shortened form and considerations for research. J Coll Stud Dev 1995;36(4):368-373. [doi: 10.1037/t05375-000]
54. Ladwig I, Rief W, Nestoriuc Y. What are the risks and side effects of psychotherapy?—Development of an inventory for the assessment of negative effects of psychotherapy (INEP). Verhaltenstherapie 2014;24(4):252-263 [FREE Full text] [doi: 10.1159/000367928]
55. Beck AT, Epstein N, Brown G, Steer RA. An inventory for measuring clinical anxiety: psychometric properties. J Consult Clin Psychol 1988;56(6):893-897. [doi: 10.1037//0022-006x.56.6.893] [Medline: 3204199]
56. Evans C, Mellor-Clark J, Margison F, Barkham M, Audin K, Connell J, et al. CORE: clinical outcomes in routine evaluation. J Ment Health 2000;9(3):247-255. [doi: 10.1080/jmh.9.3.247.255]
57. Spinhoven P, Ormel J, Sloekers PP, Kempen GI, Speckens AE, Van Hemert AM. A validation study of the hospital anxiety and depression scale (HADS) in different groups of dutch subjects. Psychol Med 1997;27(2):363-370. [doi: 10.1017/s0033291796004382] [Medline: 9089829]
58. Morin C. Insomnia: Psychological Assessment And Management: Psychological Assessment and Management (Treatment Manuals for Practitioners). New York: Guilford press; 1993.
59. Pearlin LI, Schooler C. The structure of coping. J Health Soc Behav 1978;19(1):2-21. [doi: 10.2307/2136319] [Medline: 649936]
60. Lovibond PF, Lovibond SH. The structure of negative emotional states: comparison of the depression anxiety stress scales (DASS) with the beck depression and anxiety inventories. Behav Res Ther 1995;33(3):335-343. [doi: 10.1016/0005-7967(94)00075-u] [Medline: 7726811]
61. Mundt JC, Marks IM, Shear MK, Greist JM. The work and social adjustment scale: a simple measure of impairment in functioning. Br J Psychiatry 2002;180:461-464. [doi: 10.1192/bjp.180.5.461] [Medline: 11983645]
62. Borkovec TD, Nau SD. Credibility of analogue therapy rationales. J Behav Ther Exp Psychiatry 1972;3(4):257-260. [doi: 10.1016/0005-7916(72)90045-6]
63. Lewis JR. The system usability scale: past, present, and future. Int J Hum Comput Interact 2018;34(7):577-590. [doi: 10.1080/10447318.2018.1455307]
64. Rozental A, Kottorp A, Forsström D, Månsson K, Boettcher J, Andersson G, et al. The negative effects questionnaire: psychometric properties of an instrument for assessing negative effects in psychological treatments. Behav Cogn Psychother 2019;47(5):559-572. [doi: 10.1017/S1352465819000018] [Medline: 30871650]
65. Beck AT, Weissman A, Lester D, Trexler L. The measurement of pessimism: the hopelessness scale. J Consult Clin Psychol 1974;42(6):861-865. [doi: 10.1037/h0037562] [Medline: 4436473]
66. Bosc M, Dubini A, Polin V. Development and validation of a social functioning scale, the social adaptation self-evaluation scale. Eur Neuropsychopharmacol 1997;7 Suppl 1:S57-70; discussion S71. [doi: 10.1016/s0924-977x(97)00420-3] [Medline: 9169311]

67. Bandelow B, Sagebiel A, Belz M, Görlich Y, Michaelis S, Wedekind D. Enduring effects of psychological treatments for anxiety disorders: meta-analysis of follow-up studies. Br J Psychiatry 2018;212(6):333-338. [doi: 10.1192/bjp.2018.49] [Medline: 29706139]
68. Cornish P. Stepped Care 2.0: A Paradigm Shift in Mental Health. Switzerland: Springer; 2020.
69. Farrington C, Aristidou A, Ruggeri K. mHealth and global mental health: still waiting for the mH2 wedding? Global Health 2014;10:17 [FREE Full text] [doi: 10.1186/1744-8603-10-17] [Medline: 24670011]
70. Kok RN, van Straten A, Beekman ATF, Cuijpers P. Short-term effectiveness of web-based guided self-help for phobic outpatients: randomized controlled trial. J Med Internet Res 2014;16(9):e226 [FREE Full text] [doi: 10.2196/jmir.3429] [Medline: 25266929]
71. Vollert B, von Bloh P, Eiterich N, Beintner I, Hütter K, Taylor CB, et al. Recruiting participants to an internet-based eating disorder prevention trial: impact of the recruitment strategy on symptom severity and program utilization. Int J Eat Disord 2020;53(5):476-484. [doi: 10.1002/eat.23250] [Medline: 32112593]
72. Kolar DR, Hammerle F, Jenetzky E, Huss M. Smartphone-enhanced low-threshold intervention for adolescents with anorexia nervosa (SELTIAN) waiting for outpatient psychotherapy: study protocol of a randomised controlled trial. BMJ Open 2017;7(10):e018049 [FREE Full text] [doi: 10.1136/bmjopen-2017-018049] [Medline: 29061627]
73. Levinson CA, Ralph-Nearman C, Brown ML, Gardner T, Cardi V, Treasure J, et al. A pilot randomized control trial of online exposure for eating disorders and mechanisms of change delivered after discharge from intensive eating disorder care: a registered report. Int J Eat Disord 2021;54(11):2066-2074 [FREE Full text] [doi: 10.1002/eat.23603] [Medline: 34480773]
74. Reins JA, Boß L, Lehr D, Berking M, Ebert DD. The more I got, the less I need? Efficacy of internet-based guided self-help compared to online psychoeducation for major depressive disorder. J Affect Disord 2019;246:695-705. [doi: 10.1016/j.jad.2018.12.065] [Medline: 30611913]

Abbreviations

CBT: cognitive behavioral therapy
DMHI: digital mental health intervention
ITT: intention-to-treat
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PST: problem-solving therapy
RCT: randomized controlled trial
ROBINS-I: Risk of Bias in Nonrandomized Studies-of Interventions

Edited by J Torous; submitted 22.01.24; peer-reviewed by MM Hossain, D Turnbull, C Navas-Ferrer; comments to author 25.05.24; revised version received 14.07.24; accepted 02.08.24; published 10.09.24.

Please cite as:
Huang S, Wang Y, Li G, Hall BJ, Nyman TJ
Digital Mental Health Interventions for Alleviating Depression and Anxiety During Psychotherapy Waiting Lists: Systematic Review
JMIR Ment Health 2024;11:e56650
URL: https://mental.jmir.org/2024/1/e56650
doi:10.2196/56650
PMID:39255015

Review

Digital Psychotherapies for Adults Experiencing Depressive Symptoms: Systematic Review and Meta-Analysis

Joanna Omylinska-Thurston[1], PsyD; Supritha Aithal[2], PhD; Shaun Liverpool[2], PhD; Rebecca Clark[1], BSc; Zoe Moula[3], PhD; January Wood[1], MSc; Laura Viliardos[1], PhD; Edgar Rodríguez-Dorans[4], PhD; Fleur Farish-Edwards[2], MSc; Ailsa Parsons[1], MSc; Mia Eisenstadt[5], PhD; Marcus Bull[6], MA; Linda Dubrow-Marshall[1], PhD; Scott Thurston[1], PhD; Vicky Karkou[2], PhD

[1]School of Health and Society, University of Salford, Manchester, United Kingdom
[2]Faculty of Health, Social Care and Medicine, Edge Hill University, Ormskirk, United Kingdom
[3]Faculty of Nursing, Midwifery and Palliative Care, King's College London, London, United Kingdom
[4]School of Health and Social Science, University of Edinburgh, Edinburgh, United Kingdom
[5]Evidence Based Practice Unit, University College London, London, United Kingdom
[6]Faculty of Life Sciences and Education, University of South Wales, Newport, United Kingdom

Corresponding Author:
Joanna Omylinska-Thurston, PsyD
School of Health and Society
University of Salford
Frederick Road Campus
Broad Street
Manchester, M6 6PU
United Kingdom
Phone: 44 0161 295 0000
Email: j.omylinska-thurston1@salford.ac.uk

Related Article:

This is a corrected version. See correction statement: https://mental.jmir.org/2024/1/e67439

Abstract

Background: Depression affects 5% of adults and it is a major cause of disability worldwide. Digital psychotherapies offer an accessible solution addressing this issue. This systematic review examines a spectrum of digital psychotherapies for depression, considering both their effectiveness and user perspectives.

Objective: This review focuses on identifying (1) the most common types of digital psychotherapies, (2) clients' and practitioners' perspectives on helpful and unhelpful aspects, and (3) the effectiveness of digital psychotherapies for adults with depression.

Methods: A mixed methods protocol was developed using PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. The search strategy used the Population, Intervention, Comparison, Outcomes, and Study Design (PICOS) framework covering 2010 to 2024 and 7 databases were searched. Overall, 13 authors extracted data, and all aspects of the review were checked by >1 reviewer to minimize biases. Quality appraisal was conducted for all studies. The clients' and therapists' perceptions on helpful and unhelpful factors were identified using qualitative narrative synthesis. Meta-analyses of depression outcomes were conducted using the standardized mean difference (calculated as Hedges *g*) of the postintervention change between digital psychotherapy and control groups.

Results: Of 3303 initial records, 186 records (5.63%; 160 studies) were included in the review. Quantitative studies (131/160, 81.8%) with a randomized controlled trial design (88/160, 55%) were most common. The overall sample size included 70,720 participants (female: n=51,677, 73.07%; male: n=16,779, 23.73%). Digital interventions included "stand-alone" or non–human contact interventions (58/160, 36.2%), "human contact" interventions (11/160, 6.8%), and "blended" including stand-alone and human contact interventions (91/160, 56.8%). What clients and practitioners perceived as helpful in digital interventions included support with motivation and accessibility, explanation of task reminders, resources, and learning skills to manage symptoms. What was perceived as unhelpful included problems with usability and a lack of direction or explanation. A total of 80 studies with 16,072 participants were included in the meta-analysis, revealing a moderate to large effect in favor of digital psychotherapies for depression (Hedges *g*=–0.61, 95% CI –0.75 to –0.47; *Z*=–8.58; *P*<.001). Subgroup analyses of the studies with different intervention delivery formats and session frequency did not have a statistically significant effect on the results (*P*=.48 and *P*=.97, respectively). However, blended approaches revealed a large effect size (Hedges *g*=–0.793), while interventions involving human contact (Hedges *g*=–0.42) or no human contact (Hedges *g*=–0.40) had slightly smaller effect sizes.

Conclusions: Digital interventions for depression were found to be effective regardless of format and frequency. Blended interventions have larger effect size than those involving human contact or no human contact. Digital interventions were helpful especially for diverse ethnic groups and young women. Future research should focus on understanding the sources of heterogeneity based on intervention and population characteristics.

Trial Registration: PROSPERO CRD42021238462; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=238462

(JMIR Ment Health 2024;11:e55500) doi:10.2196/55500

KEYWORDS

digital psychotherapies; depression; adults; systematic review; meta-analysis; mobile phone

Introduction

Background

Globally, depression affects >280 million people including 5% adults [1]. It is a leading cause of disability worldwide and a major contributor to the overall global burden of disease [2]. In the United Kingdom, for example, depression affects 3% to 6% of people [3] from diverse socioeconomic, educational, and cultural backgrounds.

In view of the prevalence of depression across different groups, the 2022 National Institute for Health and Care Excellence (NICE) guidelines for depression [3] have been revised to include a range of interventions as a first line of treatment, thus meeting diverse needs while widening clients' choices. The interventions include behavioral activation, exercise, mindfulness, cognitive behavioral therapy (CBT), counseling for depression, psychodynamic psychotherapy, and couples therapy. In addition, the NICE guidelines [3] recommended antidepressant medication but for severe depression only, placing an emphasis on clients' preferences and the role of psychological and physical interventions over drug treatment.

In the past, interventions for depression were provided predominantly face to face, and digital psychotherapies were available only to a limited degree. However, the outbreak of COVID-19 and consequent lockdowns, social distancing, and isolation rules have led to many people struggling with anxiety and depression [4]. The National Health Service had to adapt to provide therapy in a more flexible way, including digital delivery [5], which was in line with the policies outlined in the UK Government's Five Year Forward View for Mental Health [6] and the National Health Service Long Term Plan [7]. This has led to a need for more research in relation to the provision of digital psychotherapies.

The need for guidance in digital psychotherapies has given rise to recent systematic reviews on depression. However, the reviews are limited to formal diagnoses of unipolar depression [8] or relate to chronic health conditions [9-11]. The available reviews also focus on specific modes of delivery (eg, smartphone apps [12]) or specific client groups (eg, perinatal clients [13] or children and young people [14-17]). Another limitation of most of the published systematic reviews is a focus on CBT only [11,18-20] ignoring the fact that a range of interventions are being offered for depression [3]. A broader review of evidence in relation to a range of digital approaches and psychotherapeutic theories is needed, and this systematic review aims to address this gap.

Most recent meta-analyses highlighted that there is no difference between technology-based and in-person treatments for depression [21], and there are some indications [22] that this is reflected in clients' preferences for treatment: 55.5% of adults choose digital psychotherapy for depression. However, practitioners draw attention to the therapeutic relationship in digital psychotherapies, which is important to consider [21]. There is some evidence that working digitally does not reduce the quality of the therapeutic relationship [23], but therapists are often concerned that they do not have the same access to the clients' experience as in face-to-face interactions. Researchers argue that most therapeutic activity is grounded in the body involving body-to-body communication, attunement, and coregulation of feelings in the shared physical space, which cannot be replaced by a web-based treatment [24]. A preliminary literature search revealed no systematic reviews highlighting service users' perspectives on digital psychotherapies for depression. This would seem essential when discussing individual preferences and ethical considerations of web-based treatments especially in terms of safety and privacy [25]. This current systematic review aims to address this gap and present service users' perspectives within the existing research on digital psychotherapies for depression, highlighting their needs and preferences.

An additional problem in existing literature is a lack of clarity in the terminology used for the digital modes of delivery, leading to confusion in relation to what is effective and what is not [26]. Delivery can be, for example, asynchronous, synchronous, self-guided, with a therapist, or blended. Types of media can include telephone, videoconferencing, emails, websites, or apps. In the literature, terms are often used inconsistently. Therefore, the current systematic review will review the terminology and summarize current evidence using consistently defined terms.

Moreover, it is important to understand the specific factors impacting the effectiveness of digital interventions for depression. There are arguments that effectiveness depends on the duration of the intervention, baseline severity, adherence, and the level of human guidance [19]. This current systematic review will focus on these and other important factors within the data set, aiming to provide more specific guidance in relation to the digital psychotherapies for depression.

Research Questions

Taking the above issues into account, this systematic review will aim to answer the following 3 research questions that have not yet been addressed in other systematic reviews. The research questions will include:

1. What are the most common types of digital psychotherapeutic interventions for adults with depression?
2. What are the clients' and practitioners' perspectives on helpful and unhelpful factors in digital psychotherapeutic interventions for adults with depression?
3. What is the effectiveness of digital psychotherapeutic interventions for adults with depression?

Methods

Overview

A mixed methods systematic review protocol was developed using PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [27] and registered with PROSPERO (2021; CRD42021238462). Unlike traditional systematic reviews, this review combined both quantitative and qualitative studies. The intention was to maximize the findings not only by examining the effectiveness of digital psychotherapies but also by mapping the utility, impact, and the ability of those findings to inform policy and practice.

Search Strategy

The search strategy was developed using 3 key concepts: population, intervention, and context (population: people with depression, intervention: digital interventions, and context: psychotherapy) using Boolean operators and truncation marks (Textbox 1). The following data bases were searched up to February 12, 2024: CINAHL, PsycArticles, PsycINFO, PubMed, BASE, Academic Search Premier, and ProQuest Health Research Premium Collection.

Textbox 1. Search strings.

Step 1: "depress*" OR "dysthymi*" OR "adjustment disorder*" OR "mood disorder*" OR "affective disorder*" OR "affective symptom*"

AND

Step 2: "online" OR "remote" OR "tele-therap*" OR "digital" OR "e-mental health" OR "e-therap*" OR "mobile*" OR "internet-administered" OR "web-based" OR "app*" OR "digital*" OR "technolog*" OR "computer*" OR "tablet*" OR "m-health*" OR "mobile health" OR "e-health" OR "electronic health"

AND

Step 3: "Psychotherap*" OR "psychologic*" OR "therap*" OR "counselling" OR "counseling"

The search was limited to the 2010-to-2024 time frame in order to focus on the latest advances in digital psychotherapies.

Screening

Search results were independently screened by 4 reviewers (SA, SL, JO-T, and RC) at the title and abstract level. The full texts were then assessed for eligibility based on the predetermined criteria that were set using a combination of the Population, Intervention, Comparison, Outcomes and Study Design (PICOS) framework [28] (Table 1) and other factors such as context, time period, and the type of publication. Where data needed for eligibility assessment were missing, the authors were contacted to provide the information. Unclear or unresolved cases were discussed and moderated during the weekly team meetings.

Table 1. Eligibility criteria.

Criteria	Inclusion	Exclusion
Population	Participants with a mean age of ≥18 years, any gender, ethnicity, country, and with any severity and chronicity of depression. ≥75% of participants in the study should have a diagnosis of depression or self-report depression or low mood as a primary reason for being involved in the study. The diagnosis could involve MDD[a], dysthymic disorder, peripartum depression (previously postpartum depression), seasonal affective disorder, or premenstrual dysphoric disorder.	Studies where depression is not a primary outcome (comorbid with psychotic or other medical and mental health conditions) and acute phase of depression.
Intervention	Studies with digital psychotherapeutic intervention as the main intervention, including all forms of verbal psychotherapies and counseling (eg, humanistic, psychoanalytic or psychodynamic, cognitive or behavioral, and integrative); creative or arts psychotherapies (eg, dance, drama, art, music, and poetry); and any combination of the above delivered in any digital format (eg, websites, apps, telephone, videoconferencing, emails, etc)	Studies where ≥50% sessions are delivered nondigitally (eg, face-to-face consultation in the therapy room). S studies with advice, guidance, signposting, coaching, psychoeducation, and peer support. Studies focused on only screening, assessment, prevention, and follow- up.
Comparators	All types of comparators such as waiting list, treatment as usual, face-to-face psychological therapies, pharmacological interventions, physical interventions, or studies with no comparators	None
Outcomes	Depressive symptoms measured using any validated instruments (self-rated or observational tools) is considered as the primary outcome. In addition, views or perspectives of clients and practitioners on the processes and helpful and unhelpful factors or aspects of digital psychotherapies for adults with depression	None
Study design	Any type of empirical research with quantitative, qualitative, mixed, or arts-based approaches using surveys, pilot studies, intervention protocols, and quasi-experimental studies, RCTs[b], interviews, and other methods with people experiencing depression is considered.	Systematic reviews, secondary sources, opinion-based articles, editorials, policy reviews and statements, and commentaries. Unpublished masters or doctoral level dissertations, unpublished conference presentations, conference proceedings where full-length articles are not available, clinical case examples without explicit research methodology, and narrative articles.
Context	Psychotherapeutic interventions delivered "digitally" by qualified and registered therapists or web-based interventions or apps informed by psychotherapeutic approaches.	Nontherapeutic studies, educational videoconferences, workshops, and self-help programs that involve exclusively chats or support groups.
Time period	2010-2024	Before 2010 and unpublished ongoing studies
Publication type	Peer-reviewed	Editorials, conference presentations, and opinion-based articles

[a]MDD: major depressive disorder.

[b]RCT: randomized controlled trial.

Extraction

Studies involving both therapeutic processes and outcomes of digital psychotherapies were included to aid further understanding of the digital contribution to psychotherapies for depression. An extraction form based on a Microsoft Excel spreadsheet was developed to gather information for each research question. The review authors piloted 8 initial studies to refine the extraction form. A total of 10 authors independently extracted data from approximately 10 to 30 studies each, and 4 authors (SA, SL, JO-T, and RC) cross-checked and verified all the extracted data. Disagreements were resolved by discussion in weekly meetings, and when issues remained unclear, the members of the review team arbitrated. In the first instance, the team extracted demographic data related to the studies and population characteristics. To answer the first research question in relation to the most common types of digital interventions for depression, we used the Template for Intervention Description and Replication (TIDieR), which is a commonly used intervention description and replication checklist [29]. Furthermore, clients' and practitioners' perceptions on helpful and unhelpful factors of digital psychotherapies were gathered to answer the second research question. Finally, numeric data related to the primary outcomes of depression were documented for the third research question.

Quality Assessment

Randomized controlled trials (RCTs) were evaluated using the Cochrane risk-of-bias assessment tool [30] to identify risks such as selection bias (random sequence generation and allocation concealment), performance bias (blinding of participants and personnel), detection bias (blinding of outcome assessment), attrition bias (incomplete outcome data), and reporting bias (selective reporting). Studies with non-RCT designs (eg, controlled trials, pretest-posttest design, mixed methods, and qualitative studies) were evaluated using the Mixed Methods Appraisal Tool (MMAT; version 2018) [31]. The review authors working in pairs answered the questions that covered various aspects of the quality of execution and reporting of the studies.

Data Synthesis and Analysis

All the studies that met the inclusion criteria were considered for qualitative narrative synthesis [32] to present the characteristics of digital psychotherapies. The synthesis began with the "mapping" of the available relevant evidence against the specific research questions. Intervention-specific and person-specific factors influencing digital psychotherapy were explored and thematically analyzed using a modified behavior change model [33], which provides a useful classification of the barriers and facilitators in digital interventions [34]. Moreover, the quantitative data from studies with RCT components were analyzed to evaluate the pooled robustness of the digital psychotherapy outcomes.

Meta-analyses of depression outcomes were conducted using the standardized mean difference (SMD, calculated as Hedges *g*) of the postintervention change between digital psychotherapy and control groups to accommodate for expected methodological and intervention design variations [35]. Analyses were conducted for all depression outcomes combined, and the precision of the SMD was calculated for each trial by the 95% CI. A negative SMD implied better therapeutic effects over time in the digital psychotherapy group compared to the control group. All the analyses were performed using Comprehensive Meta-Analysis software. The pooled effect sizes were interpreted using the same rule for describing Cohen *d* effect sizes as applied to Hedges *g*. SMDs of ≤0.30, 0.30 to 0.60, and >0.60 were considered as small, moderate, and large effect sizes, respectively.

We intended to include crossover trials but only the first active treatment period. For studies with multiple arms, only those with the digital intervention and the control arms were included in the analysis [36]. If there were 2 digital intervention arms with a single control group, then the sample size of the control group was halved before the meta-analysis to avoid counting the same participants twice. When studies presented data from >1 depression measurement tool, we prioritized data in the following order: Beck Depression Inventory, Patient Health Questionnaire-9, and Hamilton Depression Rating Scale based on the most frequently used tools identified in this review.

Dealing With Missing Data

For meta-analysis where mean, SD, and sample size were missing from the end of intervention scores, we looked for alternate formats available on Comprehensive Meta-Analysis to compute the missing data. If we lacked sufficient information to extrapolate missing information, we contacted the study authors to obtain the missing data and the reason for the missing data. Where this was not possible, we excluded the study from the meta-analysis but used the descriptive information to answer other research questions and qualitative synthesis.

Assessment of Heterogeneity

We initially explored heterogeneity across studies using a visual inspection of forest plots (potential heterogeneity was considered where CIs were not overlapping). Furthermore, to assess the presence and extent of between-study variation I^2 statistic with 95% CIs (uncertainty), *Q* statistics were used. The *Q* test was performed to check if there was any variance in the true effect size across studies with an α criterion set at a low statistical power of .100 for a better possibility to reject the null hypothesis and identify if the effect sizes varied across studies. The I^2 test was performed to identify what proportion of the variance in observed effects shows variance in true effects rather than a sampling error.

Assessment of Reporting Bias

For evaluating the risk of publication bias, funnel plots for overall depression outcomes were visually inspected for asymmetry (ie, SMDs charted against their SE). As ≥10 studies were pooled, we formally tested funnel plot asymmetry using Egger test of the intercepts [37]. A positive intercept indicates that studies with smaller sample sizes tend to report more positive results than large-sample studies. When the test found notable asymmetry (*P*=.10), we reported primary outcomes based on a fixed effects model along with a random effects model. This strategy gave more weight to larger trials and helped to counterbalance a possible inflation of the therapeutic effect by discussing a more conservative effect estimate [38].

Subgroup Analysis

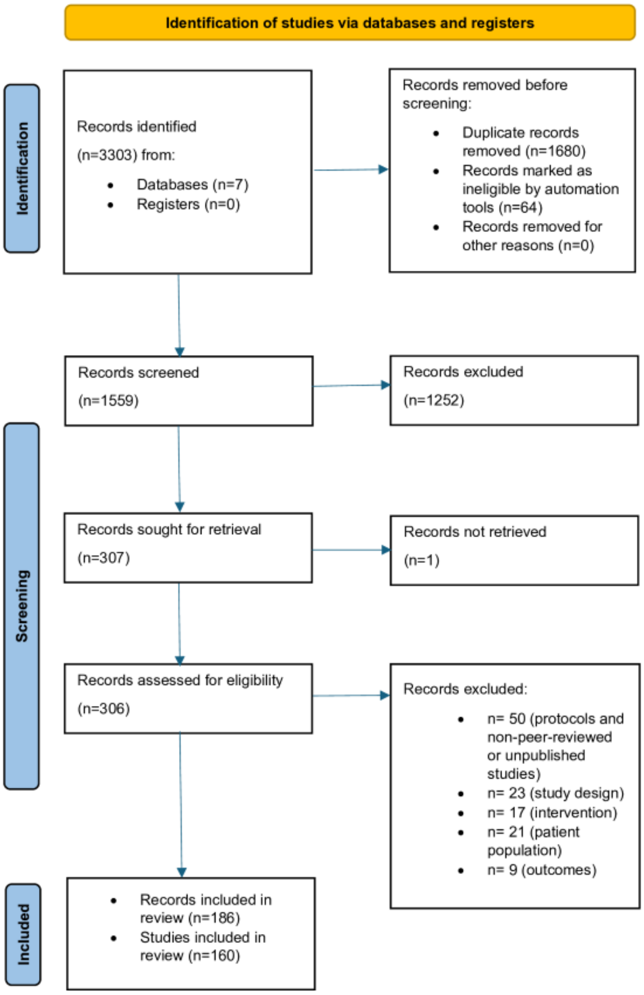
Subgroup meta-analyses were conducted to investigate between-study variability, explore the reasons for heterogeneity, and recognize intervention design components that may moderate observed efficacy. Subgroup analyses were based on a mixed effects model, which used a random effects model to generate within-subgroup variance and a fixed effects model to compare effects between subgroups [39]. Between-subgroup heterogeneity was tested using Cochrane *Q* statistic and was considered significant at the *P*=.05 level. The following moderating factors related to the intervention were included in our analysis plan: delivery format of the digital intervention (contact with human, no contact with human, and blended) and session frequency (once per week and more than once per week).

Results

Overview

As shown in the PRISMA flow diagram (Figure 1), the comprehensive search of 7 academic databases resulted in 3303 records. After duplicates were removed, the remaining 1559 (47.2%) records were screened at the title and abstract level, excluding 1252 (80.31%) records. Out of the remaining 307 articles, 306 (99.7%) were read in their entirety, while 1 (0.3%) study was not retrieved. A total of 23 (7.5%) records were excluded based on the study design, 21 (6.9%) in relation to population, 17 (5.5%) due to intervention, 9 (2.9%) based on outcomes, and 50 (16.3%) due to the type of publication. The remaining 186 (60.8%) records (Multimedia Appendix 1 [40-224]), corresponding to 160 studies, were included for data extraction. Of these, 25 (13.4%) records constituted sibling studies that formed 18 groups of studies using the same samples. To avoid double counting of data, only the main publication was considered for demographic data extraction and research questions. Data from sibling studies were mainly extracted in relation to clients' and practitioners' perspectives on helpful and unhelpful aspects of digital psychotherapies. In terms of meta-analysis of the overall efficacy of the digital psychotherapy for adults with depression, data from 80 RCTs were considered.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram for new systematic reviews that included searches of databases and registers only.



General Characteristics of Included Studies

Study Characteristics

Year of Publication

This systematic review covers the period between 2010 and February 2024. Most papers were published in 2018 (21/160, 13.1%) and 2020 (18/160, 11.2%). In the years 2023, 2021, 2013, and 2014, a total of 13 (8.1%) or 14 (8.8%) papers were published each year. The remaining years saw 3 to 10 papers yearly.

Country

A total of 31 countries were covered in the studies (n=160) including the United States (n=49, 30.6%), the United Kingdom (n=18, 11.2%), Australia (n=15, 9.4%), Germany (n=13, 8.1%), the Netherlands (n=12, 7.5%), Canada (n=9, 5.6%), and others (n=40, 25%). Most of the studies (n=116, 72.5%) were conducted in countries from the global north (eg, the United States, the United Kingdom, the Netherlands, Australia, Germany, and Canada). Furthermore, 2 (1.25%) studies included 2 sites, Australia and the United States, and 3 (1.8%) studies were multinational.

Methodological Approach

The studies (n=160) were conducted using a range of approaches but predominantly adopted quantitative approaches (n=131, 81%), out of which more than half of the studies (n=88, 55% were RCTs. A total of 15 (9.4%) studies were conducted using mixed methods designs where 3 (20%) studies had an RCT component. Only 13 (8.1%) of the studies provided solely a qualitative view of the digital therapeutic process.

Place of Recruitment

A total of 61 (38.1%) studies recruited participants from health care settings, 44 (27.5%) used online methods, 15 (9.4%) were from community settings, 12 (7.5%) from educational settings, 14 (8.8%) from other settings, and 14 (8.8%) from a combination of the above.

The "other" category included a telephone service, postal invitations, emailed letters, and places of employment. The most common combination of recruitment places was "online and other" with 3 (1.9%) studies using this combination (Table 2).

Table 2. Study characteristics (n=160).

Characteristic	Frequency, n (%)
Year of publication	
2018	21 (13.1)
2020	18 (11.2)
2023, 2013	14 (8.8)
2014	13 (8.1)
2021	12 (7.5)
2019, 2016	10 (6.2)
2017, 2015	9 (5.6)
2022	8 (5)
2012, 2010	7 (4.4)
2011	5 (3.1)
2024	3 (1.9)
Country	
The United States	49 (30.6)
The United Kingdom	18 (11.2)
Australia	15 (9.4)
Germany	13 (8.1)
The Netherlands	12 (7.5)
Canada	9 (5.6)
Austria, China, Iran, Ireland, multinational, and Switzerland	3 (1.9)
Australia and United States, Denmark, Finland, Japan, New Zealand, Norway, Japan, and Sweden	2 (1.2)
Azerbaijan, Brazil, Egypt, India, Mexico, Nigeria, Oman, the Republic of Korea, Romania, South Africa, Spain, and Turkey	1 (0.6)
Methodological approach	
Quantitative	
RCT[a]	88 (55)
Non-RCT	18 (11.2)
Descriptive	25 (15.6)
Mixed methods	15 (9.4)
Qualitative	13 (8.1)
Other	1 (0.6)
Place of recruitment	
Health care settings	61 (38.1)
Online	44 (27.5)
Community settings	15 (9.4)
Other and combination	14 (8.8)
Educational settings	12 (7.5)

[a]RCT: randomized controlled trial.

Population

Sample Size

A total of 160 studies (186 articles) encompassed a sample of 70,720 participants (Table 3).

Table 3. Participant characteristics (n=70,720).

Characteristics	Value
Age (y), mean (SD; range)	38.52 (10.87; 16-89)
Gender , n (%)	
Woman	51,677 (73.07)
Man	16,779 (23.73)
Not reported	2147 (3.04)
Other	72 (0.1)
Transgender	18 (0.03)
Nonconforming, queer	9 (0.01)
Variant	5 (0.007)
Gender Expansive	2 (0.002)
Gender Fluid	1 (0.001)
Nonbinary	1 (0.001)
Race and ethnicity[a], n (%)	
African, First Nations or Metis	4 (0.01)
African American	262 (0.65)
African Canadian, Black and White, Indigenous, Latin American, Middle Eastern, and New Zealand Indian	1 (0.002)
American	2 (0.005)
American Indian or Alaskan Native	23 (0.06)
Asian	450 (1.11)
Asian American	3 (0.01)
Black	338 (0.83)
Black African American	136 (0.34)
Caucasian	1241 (3.06)
Chinese	90 (0.22)
Dutch	667 (1.65)
European	105 (0.26)
European New Zealander	41 (0.1)
Han	73 (0.18)
Hawaiian or Pacific Islander	17 (0.04)
Hispanic or Latino	236 (0.58)
Indigenous Australian	795 (1.96)
Iranian	95 (0.23)
Maori	5 (0.01)
Mexican	16,447 (40.59)
Native American	6 (0.01)
Non-Aboriginal and Torres Strait Islander	100 (0.25)
Non-Dutch	12 (0.03)
Non-Hispanic or Latino	239 (0.59)
White	5235 (12.92)
Minority group	1731 (4.27)
Missing	10,511 (25.94)
Mixed or >1	171 (0.42)

Characteristics	Value
Other	1478 (3.65)
Clinical characteristics in 160 studies , n (%)	
Depression	87 (54.4)
Depression and anxiety	35 (21.9)
Major depressive disorder	21 (13.1)
Perinatal depression	11 (6.9)
Depression and stress	7 (4.4)
Depression in older adults and depression with suicidality	3 (1.9)
Not applicable	1 (0.6)
Severity, in 160 studies n (%)	
Moderate and above	42 (26.2)
Mild and above	35 (21.9)
Mild to moderate	31 (19.4)
Severe	20 (12.5)
Moderate	14 (8.8)
Not reported	9 (5.6)
Mild	5 (3.1)
Not applicable	4 (2.5)

[a]Race and ethnicity in the included studies: not reported, n=84 (52.5%); reported, n=76 (47.5%).

Age

Age data were analyzed in relation to means and age ranges (where stated). Most studies (138/160, 86.2%) provided mean age data, with an average of 38.52 (SD 10.87) years. Moreover, 45.6% (73/160) of the studies provided a minimum age, and 41.2% (66/160) of the studies provided a maximum age, with an average age range of 38.52 years. In terms of specific age groupings (18-24, 25-39, 40-64, and >65 years), 15.6% (25/160) studies spanned all 4 categories, and 41.2% (66/160) studies included >1 category.

Gender

Of the 160 studies (participants, n=70,720) reviewed, 51,677 (73.07%) of the participants were women , 16,779 (23.73%) were men , and 2147 (3.04%) participants did not report their gender. The diversity of gender generated 7 different categories, including transgender (n=18, 0.03%), queer (n=9, 0.01%), nonconforming (n=9, 0.01%), variant (n=5, 0.007%), gender expansive (n=2, 0.002%), gender fluid (n=1, 0.001%), and nonbinary (n=1, 0.001%). A total of 72 (0.1%) participants did not specify their gender. Most of the studies (145/160, 90.1%) had a mixture of men and women in their sample. However, 14 (8.8%) studies recruited women only while, 1 (0.6%) study men only; 8 (5%) studies recruited more men than women. While 4 (2.5%) studies had an equal number of men and women participants in the sample, 2 (1.2%) studies did not report gender characteristics.

Race and Ethnicity

Of the 160 studies (participants, n=70,720), 84 (52.5%) did not report race and ethnic data, leaving 30,197 (42.7%) participants not accounted for. In the remaining studies, there was no information about the race and ethnicity of an additional 10,511 (14.86%) participants, which brings the overall number of participants who did not report their race and ethnicity to 40,708 (57.56%). A total of 76 (47.5%) studies considered race and ethnicity in their analysis involving 40,523 (57.3%) participants and groups such as Asian (n=713, 1.76%; Asian American, Asian, Chinese, New Zealand Indian, Iranian, Han, and Middle Eastern); Black (n=741, 1.83%; Black African American, African American, African, African Canadian, and Black); Hispanic (n=16,684, 41.17%; Mexican, Hispanic Latino, and Latin American); Indigenous (n=851, 2.1%; First Nations or Metis, American Indian or Alaska Native, Maori, Indigenous Australians, Native Americans, Hawaiian Pacific Islander, and Indigenous); White (n=7291, 17.99%; White, Caucasian, European, European New Zealander, Dutch, and American); other (n=3560, 8.79%; non-Hispanic or Latino, non-Dutch, Other, non-Aboriginal Torres Strait Islander, and minority groups); and mixed (n=172, 0.42%; Black and White, mixed or >1).

Clinical Characteristics and Severity

There was a spread in terms of the severity of depression among the 160 studies, with 36 (22.5%) studies focusing on mild and above, 31 (19.4%) studies focusing on mild to moderate, and 27 (16.9%) on moderate to severe depression. A smaller number of studies (n=20, 12.5%) focused on severe depression only. Some studies (n=21, 13.1%) included major depressive disorder, perinatal depression (n=12, 7.5%), and (n=3, 1.9%) depression in older adults. A group of studies (n=35, 21.9%) focused on both depression and anxiety, and 3 (1.9%) studies included multiple categories. There were also studies (n=9, 5.6%) that discussed other concerns such as suicidal ideation or alcohol misuse; diverse populations such as lesbian, gay, bisexual, and transgender communities; and different contexts such as workplace or rural communities.

Research Question 1: Mapping the Types of Digital Interventions

MoodGYM was the most popular named digital intervention (15/160, 9.4%), which used an interactive workbook (Table 4). Beating the Blues was also common (8/160, 5%) and involved an 8-session course supported by a counselor. Other interventions included Good Days Ahead (4/160, 2.5%), a 9-lesson psychoeducational program, and Mindful Mood Balance (3/160, 1.9%), an 8-week course integrating mindfulness meditation with CBT. Furthermore, 5% (8/160) studies used a combination of digital interventions. Some interventions involved apps; for example, Project EVO (3/160, 1.9%) was a video game training app.

Table 4. Frequency of named interventions (n=70).

Named intervention	Frequency, n (%)
MoodGYM[a]	15 (21)
Beating the Blues[b]	8 (11)
Good Days Ahead	4 (6)
GET.ON Mood Enhancer, Health tips[c], iPST[c] Meru health program, Mindful Mood Balance, Project EVO[c]	3 (4)
BluePages[d], Cognitive Bias Modification Imagery, Deprexis, Intellicare, Living to the Full, Mom Mood Booster, My Compass[e], My Strength Inc, The Journal, This Way Up	2 (3)
Ascent, BAML, BetterHelp, Bluewatch, Colour Your Life, Cooperation After Divorce, Dario Behavioural Health, e-SMART, Empower@Home, Feel Stress Free, Feeling Better, Get Happy, Good Life Compass, Gratitude Visit[f], Happy@Work, HappyMom, Hdep, iFightDepression, Just In Time Adaptive Intervention, LifeApp'tite, Life Flex, Making the Golden Years Golden Again, MamaKits, Man Central, MARIGOLD, Master your Mood, MasterMind, Mindbalance, Mindfulness Virtual Community, MindWise 2.0, Moodbuster, MoodHacker, Mood Manager, MoodGarden[g], Motherly, NexJ Health Inc, OctaVis, Online Life Review, Online Writing, Op Koers Online, OPTT, Overcoming Thoughts, PaarBalance, Pacifica, Peak, PRIME-D, Signature Strength[f], Sinasprite, Sokoon, SPARX, Talkspace, TAPI, Three Good Things, Three Funny Things[f], Thrive, Todac, VIDA, Wellness Workshop	1 (1)

[a]Used in combination with other named interventions (n=4).

[b]Used in combination with MoodGYM (n=1).

[c]Project EVO, iPST, and Health tips used in combination (n=3).

[d]Used in combination with MoodGYM (n=2).

[e]Used in combination with MoodGYM (n=1).

[f]Gratitude visit, Three Good Things, Three Funny Things, and signature strength used in combination (n=1).

[g]Used in combination with MoodGYM (n=1).

Of the 160 studies reviewed, most interventions (n=132, 82.5%) used CBT as their framework, 6 (3.8%) referred to positive psychology interventions, and only 1 (0.6%) adopted a psychoanalytic and psychodynamic approach. In 8 (5%) studies, the interventions were based on an integrative framework such as combining CBT with positive psychology, while 13 (8.1%) studies described interventions in which the theoretical models were either not clearly defined or did not fall within the 4 main approaches to psychotherapy (ie, cognitive-behavioral, humanistic, psychoanalytic and psychodynamic, and integrative). Table 5 shows the definitions of specific types of interventions.

Out of the 160 studies reviewed, 93 (58.1%) were web based, 25 (15.6%) used mobile apps, 13 (8.1%) used computer programs, 1 (0.6%) involved virtual reality, and 1 (0.6%) was avatar based. Moreover, 26 (16.2%) studies used a combination of the above methods.

The interventions included messaging, emails, and calls (17/160, 10.6%), online peer support (11/160, 6.9%), online face-to-face contact (7/160, 4.4%), and virtual or augmented reality (1/160, 0.6%). Interventions were also delivered via participants watching videos (19/160, 11.9%) and listening to audio (11/160, 11.9%).

In terms of psychotherapeutic approaches, the digital interventions were mainly based on CBT (132/160, 82.5%). A small number of interventions used integrative approach (8/160, 5%), positive psychology (5/160, 3.1%) and psychodynamic and psychoanalytic psychotherapy (1/160, 0.6%).

Table 5. Definitions of the types of interventions (n=160).

Type of intervention	Definition	Studies, n (%)
No human contact and stand-alone	No human involvement in the therapeutic process, for example, an app that was fully automated and sent reminders to users via notifications or emails, but the participants did not have any contact with a human therapist	58 (36.2)
Human contact	Participants had web-based or offline sessions with the therapist. The contact could be synchronous (eg, a Zoom video call) or asynchronous (eg, an email), but the therapist was involved in the clients' journey. For example, the contact could include live Zoom video calls only or Zoom video calls and then text follow-up with the therapist. However, the contact had to be with a therapist, not a researcher	11 (6.9)
Blended	Both "stand-alone" and "human contact" interventions were used. For example, the participant worked through modules independently on a website and then met with a therapist via a Zoom video call or in person. The degree of contact with the therapist varied	91 (56.9)

Most of the therapeutic activities (86/160, 53.8%) involved web-based psychoeducation drawn primarily from CBT, including cognitive restructuring (34/160, 21.3%), behavioral activation or activity planning (23/160, 14.8%), mood rating (14/160, 8.8%), problem-solving (16/160, 10%), goal setting (5/160, 3.1%), and graded exposure or behavioral experiments (4/160, 2.5%). Some interventions (22/160, 13.8%) were delivered using creative means such as games or quizzes, music, journal writing, comic books, stories, animations, singing, bibliotherapy, and graphics. Homework with therapeutic tasks also featured (19/160, 11.9%). Different forms of relaxation (17/160, 10.6%) including visualization, guided relaxation, body scan, yoga, self-hypnosis, deep breathing, and progressive muscular relaxation were mentioned. Mindfulness and meditation (16/160, 10%) and acceptance and commitment therapy or values–based interventions (7/160, 4.4%) were also common. Less-common approaches included working with emotions or emotion regulation strategies (5/160, 3.1%), social skills learning (5/160, 3.1%, relapse prevention (2/160, 1.2%), coping skills training (2/160, 1.2%), cognitive training for memory improvement (1/160, 0.6%), neuroplasticity principles (1/160, 0.6%), and an emotional faces memory task (1/160, 0.6%).

Where reported, the most common dosage for interventions was 6 to 12 weeks in length (73/160, 45.6%), 6 to 12 sessions (44/160, 27.5%), and 6 to 12 modules (30/160, 18.6%). A further 23 (14.4%) studies lasted for <6 weeks, while 6 (3.6%) took >12 weeks. Most commonly, the interventions included 8 (mean 9.66) sessions of 60 (mean 43.35) minutes and were spread over 8 (mean 8.29) weeks, taking place once a week. However, due to variation, the average frequency per week was 2.18 sessions. At least 17 (10.6%) studies stated access as "ad libitum" or "self-paced."

Research Question 2: Clients' and Practitioners' Perspectives on Helpful and Unhelpful Factors Identified in Qualitative and Mixed Methods Studies

Overview

Qualitative (13/160, 8.1%) and mixed methods (15/160, 9.4%) studies were used to identify clients' and practitioners' perspectives on helpful and unhelpful factors in digital psychotherapeutic interventions for depression. In line with the categorization above, these interventions were divided into 3 groups: with no human contact or stand-alone, with human contact (via digital means), and a combination of human contact with stand-alone approach (blended).

In order to identify the specific helpful and unhelpful factors within each of the above groups, 4 authors used a modified behavior change model [33] as applied by Liverpool et al [34], outlining the barriers and facilitators to engagement in digital mental health interventions including person-specific and intervention-specific influencing factors. In terms of person-specific factors, motivation, opportunity, and capability were listed as the influencing factors. In terms of intervention-specific influencing factors, suitability, usability, and acceptability were identified.

Interventions That Did Not Involve Human Contact or Stand-Alone Interventions

The interventions included in this category involved 11fully automated programs. Of these 11 interventions, 10 (91%) were delivered through dedicated websites [40-49] and 1 (9%) was delivered through a fully automated app [50].

A total of 9 (82%) studies discussed helpful and unhelpful factors from the clients' perspectives. In addition, 1 (9%) study [46] was interested in consulting the public, as potential users, to better understand how interventions for depression could be improved for lesbian and gay users. Another study [50] involved collecting data from potential users and 2 groups of experts formed by researchers and health care professionals.

For the interventions delivered via websites, the most frequently reported (6/10, 60% studies) helpful factors related to "motivation." Examples of what "motivated" people to engage included gaining improvement in mental health, learning coping skills in difficult situations (eg, when anxious), achieving behavioral change, gaining awareness and insight, learning self-reflection skills, and having a sense of achievement and self-efficacy [44,49]. According to participants in the reviewed studies, digital psychotherapies also provided an "opportunity" to engage in online approaches, which was seen as equally valid as seeing a professional [47]. Reference was also made to the "suitability" of the interventions, as it allowed clients to undertake therapy at their own pace, time, and place [41], and "acceptability," when participants liked working in private due to discomfort related to working with personal issues [49].

Unhelpful factors related to the interventions delivered via website included issues with "acceptability," involving comments that the exercises were overwhelming, disconnected from experiences, "typical" advice [48], and repetitive. Participants also commented that they would have preferred a more engaging interactive format [49]. Unhelpful factors also related to the limitations of "opportunity," as the intervention was too flexible and consequently easy to avoid or too difficult to sustain as it required personal initiative. Users also reported the tasks were too demanding and felt like "work," which decreased their inclination to engage [42].

The study that focused on a fully automated app [50] discussed "acceptability," including an appealing visual layout and organization of content of the app, as well as the offer of a wide range of psychological strategies. End users also reported satisfaction with the increased self-awareness promoted by the app, which kept them "motivated" to continue engaging with the intervention.

Some unhelpful factors included the lack of contact with the therapist, which made the therapy "unacceptable" and ineffective [42], and a lack of guidance, making it difficult for participants to know whether they had used the content appropriately [41] or wasted their "opportunity."

Interventions That Involved Human Contact

There were 4 qualitative or mixed methods studies that focused on interventions involving human contact; 2 (50%) studies used a website [40,51], 1 (25%) study used a combination of digital formats (eg, videoconferencing and self-help materials) [52], and 1 (25%) study used videoconferencing as a primary mode of delivery [53]. Helpful and unhelpful factors were reported from the clients' perspectives.

In terms of websites, helpful factors focused mainly on "usability," such as explanations that helped with engagement and managing expectations [51]. Improved symptoms of anxiety and social support increased the "motivation" to continue using the intervention [40]. However, intervention "usability" problems were also mentioned as an unhelpful factor. For example, some clients indicated that when there were no attempts to manage the expectations of tasks, this led to difficulties and uncertainty how to respond [51]. No significant change in depression or dysfunctional thinking led to decreased "motivation" for engagement in therapy; this was also seen as an unhelpful factor [40].

When videoconferencing was used, participants reported feeling satisfied, as they were able to engage in therapy from their own environment and adapt it to their own needs (acceptability) [53]. However, they appreciated having an in-person assessment before the digital intervention. Initially, participants were worried that they would feel separated from the therapist, but they did not experience this during the intervention. In terms of unhelpful factors, participants reported "usability" issues especially when there were no technical explanations, and their privacy was compromised. In 1 study that implemented a combination of virtual reality and email [52], the intervention seemed "acceptable," as the self-help materials provided a calming and structured way to reflect on difficulties without a therapist's input [52]. Unhelpful aspects related to "usability" issues and uncertainty about the function of the group (where the intervention included online peer support) [52].

In addition, 1 study [54] highlighted the need for collaborative platforms in old age as older participants (aged > 65 years) used media less often than younger participants.

Interventions That Combined Human Contact With Automated Interventions (Blended)

Qualitative and mixed methods studies within the "blended" category (20/28, 71%) used mainly human contact alongside noncontact automated interventions.

Human contact methods included face-to-face appointments, telephone or video calls, emails, or asynchronous messaging (eg, via a forum). The most common approach was to use a combination of ≥2 different contact methods (8/28, 29%). In terms of the automated intervention component, 39% (11/28) studies used a website.

In these blended interventions, the use of daily practices, reminders, and likable content were viewed as "helpful" since they "motivated" clients to engage with the intervention [222]. Similarly, learning skills that focused on new ways of relating to negative thoughts, emotions, and sensations [55] and using likable content and helpful activities [56] "motivated" clients to engage. Other "motivating" factors included resources and tools to help manage stress and anxiety [57]; tools for soothing and helping to improve mood [58]; insights and resources offered by a coach [57]; and interventions interrupting the downward spiral of negative thinking [59], increasing the awareness of personal warning signs of impending relapse [222]; and confrontations by therapists in relation to completing more online sessions [60].

Other examples of helpful factors included the "opportunity" to access additional sources of social support, which made some clients feel more connected and less lonely [56,58]. Examples of "suitability," from the therapists' perspective, included the freedom to work anywhere at any time [61]. Clients, too, favored flexible scheduling [62]. Blended approaches were also viewed as addressing the issue of inadequate access to specialist care [63].

Examples of unhelpful factors included comments that the modules were not engaging and too lengthy and that certain program features were complicated and hard to follow; these issues had an impact on "usability" [62,64,65]. Other examples of unhelpful factors included an interface that felt less "acceptable"; some clients commented that the programs were not advanced enough [66], too structured [60], and not user-friendly [56]. In some cases, clients felt that they were asked to write "essays," which was not useful [59]. Clients also commented on the costs related to data charges [58], time and scheduling issues [222], and technical difficulties with access [65], which challenged the "suitability" of the interventions. From practitioners' perspectives, the constraints of hectic practice, inadequate knowledge, and competing tasks made it more difficult to use [63].

A lack of human contact, real-time interaction, dialogue, and guidance left users feeling a responsibility that required too much from them and sometimes left them feeling lonely [58,61]. The absence of synchronous group interaction led to feeling a loss of interpersonal learning [222], which again seemed like a lost "opportunity" of meaningful therapeutic interaction.

Research Question 3: Effectiveness of the Interventions

Overall Efficacy of Digital Psychotherapies on Depression

A total of 80 studies (intervention group, n=8444; control group, n=7628; total n=16,072) were included in the meta-analysis to examine the effects of digital psychotherapy interventions plus standard care compared to control groups (standard care alone, waiting list, or active controls) for depression in adult participants. The overall effect of digital psychotherapies on depression was moderate to large and statistically significant in favor of digital psychotherapies (Hedges *g*=–0.61, 95% CI –0.75 to –0.47; *Z*=–8.58; *P*<.001) immediately after the intervention (Figure 2 [40,52,59,65,68-80,82,87,89,91,93,94,101,106,109,113,114,116-136,138-146,149-152,156-158,161-166,168,170-174,177,178,181,189,191,195,198,201,206,212,213,216,218,223,224]). The resulting funnel plot from the overall depression outcomes (Figure 3) did not appear to have significant asymmetry (Egger intercept=0.07; *P*=.94).

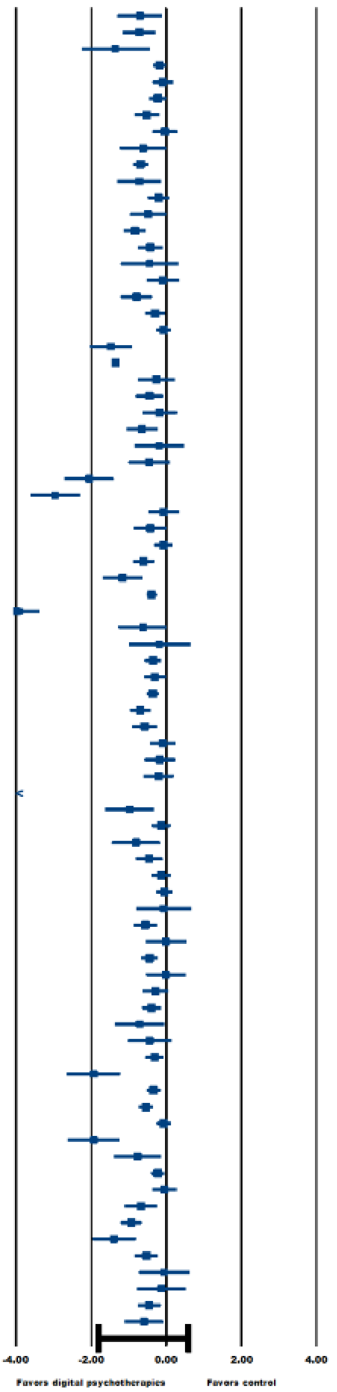
The *Q* test revealed a value of 1348.221 with 79 degrees of freedom and *P*<.001. Using a criterion α of .100, the null hypothesis that the true effect size is the same in all these studies was rejected. The I^2 statistic indicates that 94% of the variance in observed effects reflects the variance in true effects rather than a sampling error. As shown in Figure 4, assuming that the true effects are normally distributed (in *g* units), it was estimated that the prediction interval falls within –1.811 to 0.587, indicating that the true effect size in 95% of all comparable populations falls in this interval.

To explore the potential sources of heterogeneity, subgroup analyses (Figure 5; delivery format: blended [40,59,69,71,75-78,80,94,106,109,113,116,120,123,130,132,136,138,163,178,182,186,187,195,198,201,212,213,216,218], contact with human [52,65,68,82,118,121,127,144,157,161,167,170,174,175,188,223], no human contact [67,70,72-74,79,87,89,91,93,101,122,125,126,128,134,139,141-143,145,149,151,164,165,171,177,181,191,206,224], session frequency: more than one per week [40,52,59,69,71,74-77,79,80,87,89,91,93,101,106,113,120,121,123,125,128,130,132,134,138,139,144,145,151,157,164,165,168,170,172,174,181,182,186,188,195,201,213, 216], once per week [65,67,68,70,72,73,78,82,94,116,118,122,136,141-143,161,163,171,175,177,178, 187,191,212,223,224]) were conducted based on the intervention characteristics of the studies using a sufficient number of trials (80 for deliver format; 74 for session frequency) and participants. Therefore, the covariate distribution was not concerning in the subgroups. The type of delivery of the intervention (eg, blended, contact, or no contact with humans) did not have statistically significant modifying effects on the results of digital psychotherapies in comparison with the control group (*P*=.48). The intervention effects however are consistently in favor of digital psychotherapies for trials delivered in the 3 different formats studied, although the intervention effect is slightly greater for the trials delivered through a blended format (Hedges *g*=–0.73) than for the trials using contact with human (Hedges *g*=–0.42) or no contact with humans (Hedges *g*=–0.40). Similarly, the frequency of the intervention (once per week or more than once per week) did not have statistically significant modifying effects on the results of digital psychotherapies in comparison with the control group (*P*=.97). However, the intervention effect is slightly greater for trials that delivered the intervention more than once per week (Hedges *g*=–0.60) than for trials offering once per week intervention (Hedges *g*=–0.40). As the residual unexplained heterogeneity between the trials within all these subgroups is still persistent, the validity of the intervention effect is uncertain and requires further exploration to discuss the potential confounding variables.

Figure 2. Overall efficacy of digital psychotherapies on depression outcomes in adults.

Study name	Hedges g	Lower limit	Upper limit	Sample size: Intervention	Sample size: Control
Al-Alawi, 2021	-0.70	-1.29	-0.11	22	24
Alavi, 2016	-0.72	-1.14	-0.29	43	46
Anguera, 2017	-1.35	-2.25	-0.45	12	10
Arean, 2016	-0.17	-0.34	-0.01	420	206
Baikie, 2012	-0.09	-0.36	0.18	139	86
Birney, 2016	-0.23	-0.45	-0.00	150	150
Bisby, 2023	-0.52	-0.85	-0.19	75	70
Blackwell, 2015	-0.03	-0.35	0.28	76	74
Bowler, 2012	-0.61	-1.22	-0.00	21	21
Buntrock, 2015	-0.69	-0.89	-0.49	202	204
Calkins, 2015	-0.72	-1.29	-0.14	24	24
de Graaf, 2010	-0.20	-0.48	0.08	97	97
Douma, 2020	-0.48	-0.96	-0.00	34	33
Ebert, 2018	-0.84	-1.12	-0.55	102	102
El Morr, 2020	-0.43	-0.75	-0.10	68	80
Ellis, 2011	-0.44	-1.19	0.31	13	13
Fatori, 2023	-0.09	-0.52	0.34	37	44
Forman-Hoffman, 2024	-0.79	-1.20	-0.39	54	46
Geraedts, 2015	-0.30	-0.55	-0.04	116	115
Gilbody, 2015	-0.07	-0.25	0.12	210	239
Goma, 2023	-1.47	-2.03	-0.92	37	27
Hald, 2020	-1.35	-1.45	-1.25	1031	825
Hatcher, 2018	-0.26	-0.75	0.23	35	28
Heller, 2020	-0.45	-0.81	-0.08	54	65
Hirsch, 2017	-0.17	-0.62	0.28	36	38
Høifødt, 2013	-0.64	-1.05	-0.24	52	44
Hur, 2018	-0.19	-0.85	0.47	17	17
Iacoviello, 2018	-0.45	-1.00	0.10	28	23
Ime, 2023	-2.06	-2.70	-1.41	28	27
Jannati, 2020	-2.96	-3.61	-2.31	38	37
Jelinek, 2020	-0.07	-0.47	0.34	69	35
Keller, 2021	-0.43	-0.86	0.01	32	54
Kenter, 2016	-0.07	-0.31	0.17	136	133
Kingston, 2020	-0.61	-0.89	-0.33	102	103
Kivi, 2014	-1.16	-1.68	-0.64	30	35
Klein, 2016	-0.39	-0.51	-0.27	509	504
Kramer, 2021	-3.97	-4.55	-3.39	67	69
Lappalainen, 2015	-0.62	-1.26	0.01	18	20
Lemma, 2013	-0.18	-1.00	0.64	16	8
Levesque, 2011	-0.36	-0.57	-0.14	174	176
Levin, 2011	-0.30	-0.59	-0.02	99	91
Löbner, 2019	-0.36	-0.51	-0.20	320	327
Lockman, 2017	-0.70	-0.96	-0.43	97	140
Lu, 2023	-0.58	-0.91	-0.25	72	73
Ludtke, 2018	-0.09	-0.43	0.24	65	67
MacLean, 2020	-0.17	-0.57	0.23	47	46
McCloud, 2020	-0.20	-0.60	0.20	41	57
Moberg, 2019	-4.51	-4.84	-4.18	253	247
Nakao, 2018	-0.98	-1.62	-0.34	20	20
Oehler, 2020	-0.13	-0.38	0.11	133	129
Otared, 2021	-0.81	-1.44	-0.17	20	20
O'Toole, 2019	-0.46	-0.81	-0.11	60	69
Pfeiffer, 2020	-0.13	-0.39	0.12	108	128
Phillips, 2014	-0.05	-0.26	0.16	164	176
Pinto, 2015	-0.07	-0.80	0.66	12	16
Pots, 2016	-0.56	-0.86	-0.25	82	87
Preschl, 2011	-0.01	-0.54	0.52	25	28
Proudfoot, 2013	-0.45	-0.67	-0.22	126	198
Proyer, 2014	-0.01	-0.53	0.50	85	17
Reins, 2019	-0.28	-0.63	0.06	65	66
Richards, 2020	-0.39	-0.64	-0.14	198	91
Richter, 2022	-0.71	-1.37	-0.06	20	17
Ritvo, 2021	-0.44	-1.02	0.14	22	23
Rollman, 2018	-0.31	-0.55	-0.07	253	92
Sandoval, 2015	-1.93	-2.63	-1.23	25	20
Schneider, 2017	-0.34	-0.52	-0.15	235	220
Segal, 2020	-0.55	-0.73	-0.36	230	230
Sergeant, 2020	-0.07	-0.25	0.12	253	213
Sethi, 2020	-1.93	-2.62	-1.24	23	23
Shah, 2018	-0.76	-1.39	-0.14	29	16
Silverstone, 2017	-0.23	-0.40	-0.06	275	255
Thase, 2018	-0.05	-0.36	0.26	77	77
Tulbure, 2018	-0.68	-1.10	-0.25	64	33
van der Zanden, 2012	-0.94	-1.20	-0.67	121	123
Vernmark, 2010	-1.39	-1.96	-0.82	29	29
Warmerdam, 2013	-0.53	-0.84	-0.23	102	71
Westerhof, 2017	-0.05	-0.72	0.61	27	12
Williams, 2022	-0.13	-0.77	0.51	19	17
Wright, 2022	-0.46	-0.76	-0.16	95	80
Yeung, 2018	-0.59	-1.10	-0.09	29	32
Prediction Interval	-0.61	-1.81	0.59		

Hedges *g* and 95%CI



Q=3.054 df=6, p=0.802; Isq=0.00%

Figure 3. Funnel plot for overall effects without removal of outliers.

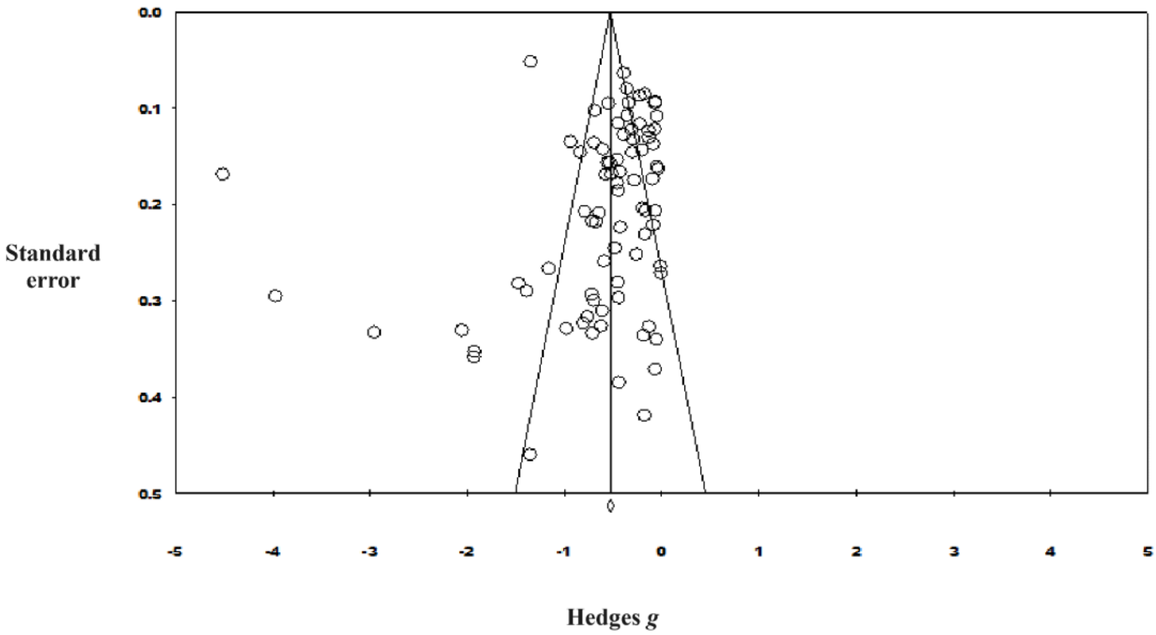


Figure 4. Distribution of true effects of the overall efficacy of digital psychotherapies for depression in adults. The mean effect size is –0.61 (95% CI–0.75 to –0.47). The true effect size is 95% of all comparable populations falls in the interval -1.81 to 0.59.

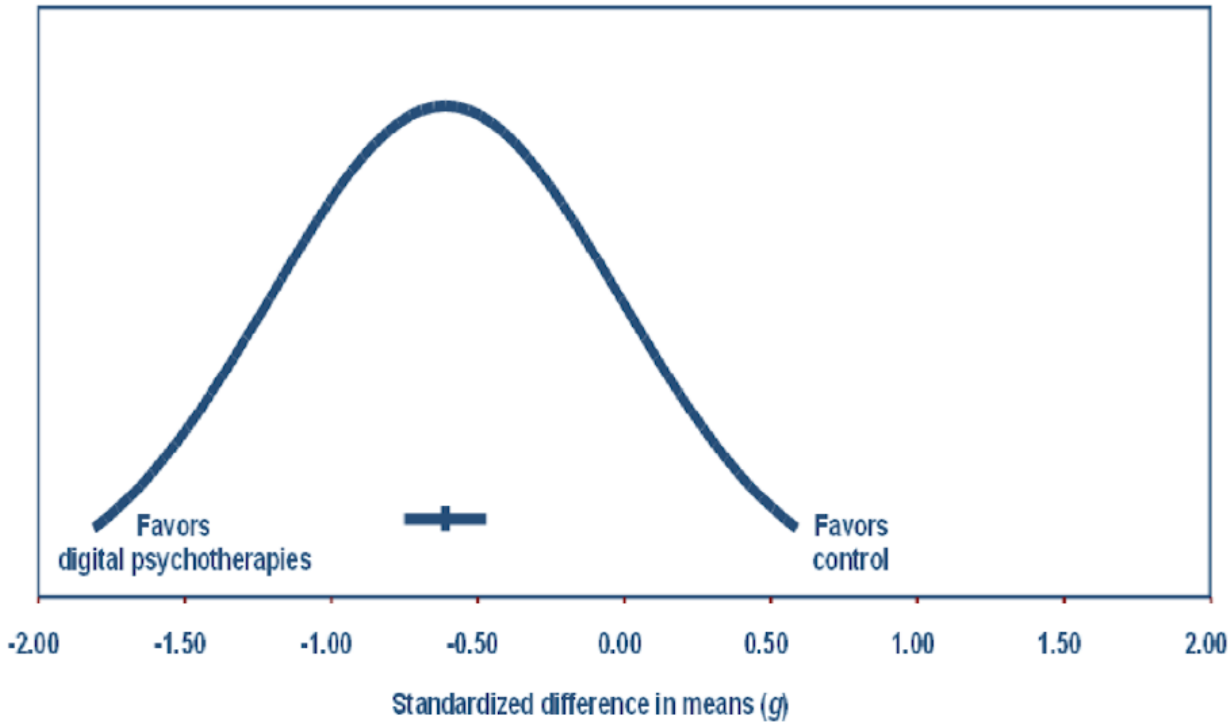
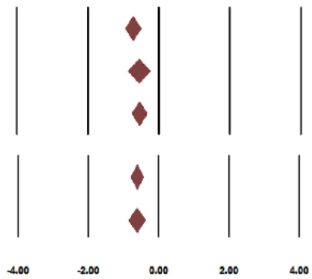


Figure 5. Subgroup analyses. *Q-test for between-group heterogeneity, mixed-effects model. **Six studies that did not offer session frequency data or did not fit into these two categories were excluded from this analysis. .

Outcome Subgroups	No.of Studies	I squared %	Hedges *g* (95% CI)	Hedges *g* (95% CI)	*P* value for between group heterogeniety *
Delivery Format					
Blended	32	93.28		-0.73 (-0.78 to -0.67)	
Contact with human	16	74.29		-0.42 (-0.51 to -0.32)	0.48
No contact with human	32	95.83		-0.40 (-0.45 to -0.35)	
Session Frequency **					
> than once per week	47	95.65		-0.60 (-0.63 to -0.55)	0.97
Once per week	27	89.25		-0.40 (-0.46 to -0.34)	



Quality Appraisal

Methodological quality varied across the included studies (Multimedia Appendix 3).

Most of the qualitative studies (12/13, 92%) were assessed as meeting all 5 MMAT criteria [31] (marked as "yes"), 2 (15%) studies were judged as not using appropriate qualitative approaches to answer the research question, and the findings did not seem adequately derived from the data (judged as "no" and "can't tell/unclear"). In 2 (15%) studies, the interpretation of] the results was judged as not being sufficiently substantiated by data.

Out of 15 nonrandomized quantitative trials (eg, quasi-experimental studies), 1 (7%) study was assessed as meeting all 5 MMAT criteria and 14 (93%) studies were assessed as "no" or "unclear" in terms of meeting at least 1 MMAT criteria, including presenting complete outcome data or appropriately accounting for confounder (n=8, 53%), administration of the intervention (n=7, 47%), having representative target samples (n=5, 33%), and using appropriate measurement tools (n=1, 7%).

Out of 25 quantitative descriptive studies, only 4 (16%) were judged as meeting all 5 criteria on MMAT; 13 (52%) studies were judged as "no" or "unclear" in terms of having representative samples, 5 (20%) for appropriate sampling strategies, and 4 (16%) in relation to measurement tools. A total of 18 (72%) studies were judged as "no" or "unclear" in terms of low risk of nonresponse bias, and 2 (8%) studies were judged as not having appropriate statistical analysis.

Most of the mixed methods studies without an RCT component were judged as not meeting at least 1 MMAT criteria (12/15, 80%). Only 2 (13%) studies met all 5 MMAT criteria, while 2 (13%) studies were judged as "no" or "unclear" for having adequate rationale for using mixed methods design. All studies (15/15, 100%) were judged as having qualitative and quantitative components of the study effectively integrated, but 4 (27%) studies were judged as "no" or "unclear" for providing adequate interpretation of the integration. A total of 10 (67%) studies were judged as "no" or "unclear" for adequately addressing divergences and inconsistencies between the quantitative and qualitative results, and 8 (53%) studies were judged as "no" or "unclear" for adherence to quality criteria for each type of methods.

Regarding the RCTs and mixed methods studies with an RCT component (83 studies) following Higgins and Green's [30] risk-of-bias assessment tool, random sequence allocation was judged as having a low risk of bias for most of the studies (n=74, 89%). Studies used methods such as software-generated block random sequence [223], algorithm-generated sequence [67], web-based sequence generator [68], computer random number generator [69], manual random sequence generation written in sealed envelopes [70,71], and via independent researchers [72]. A total of 9 (11%) studies, while identified as randomized trials, did not offer sufficient information on how the randomization took place [52,73], and therefore, they were marked as having an "unclear" risk of bias. Many authors did not adequately report their method of concealment, and therefore the risk of bias for this criterion was evaluated as "unclear" for 47% (n=39) of the studies and "high" for 18% (n=15) of the studies. Only 34% (n=29) of the studies clearly reported their method of concealment. When done successfully, authors used blocked randomization delivered in sealed envelopes from a centralized point instead of dividing participants across multiple recruitment centers [23]. Blinding was assessed with respect to participants, personnel, and outcome assessors. We judged only 9% (n=8) of the studies as having implemented blinding successfully with a "low" risk of bias [74,75]. Many studies were not explicit in their reporting of this procedure ("unclear"; n=32, 39%). In some contexts, authors reported that it was difficult or impossible to blind the participants in relation to the type of intervention. In other studies [76,77], blinding of personnel involved in the study was not possible ("high"; n=41, 49%). The risk of bias for blinding of outcome assessment was marked as "unclear" for 38 (46%) studies [78,224] and "high" for 22% (n=18) of the studies. Only 32% (n=27) of the studies were judged as having "low" risk of bias in terms of blinding of outcomes [79,80].

There were some important differences between the outcome assessment measurements across the studies. Blinding of assessors was often not feasible when depression was assessed as a self-reported measure rather than a clinician-rated measure. In many instances, authors reported the reasons for dropout, offered transparent reporting of attrition, and used intention-to-treat analysis (n=74, 89%). This was “unclear” in 10% (n=8) of the studies, and the risk of bias was “high” in only 1% (n=1) of the studies. The risk of selective reporting of the results (including depression outcomes) was judged as “low” for 80% (n=66) of the studies, as the differences within and between intervention groups were given regardless of the direction, magnitude, or statistical significance. The risk of selective reporting was judged as “high” and “unclear” only in 10% (n=8) and 11% (n=9) of the studies, respectively. Overall, selection bias and performance bias were identified as very likely to influence the quality of the results in the included studies.

Discussion

The systematic review identified the current literature on the topic of digital psychotherapies for adults experiencing depressive symptoms, including the most common types of interventions, the clients’ and practitioners’ views on helpful and unhelpful factors, and the effectiveness of the digital interventions for depression.

Characteristics of Reviewed Studies

A total of 186 eligible papers (160 primary studies and 26 sibling studies) met the inclusion criteria for this review. The studies accounted for 70,720 participants, including 51,677 (female: 73.07%) participants, which constituted a 3:1 female-to-male participants ratio. When age was reported, a relatively young mean age was present (mean 38.52 years). It therefore appears that women and younger people were overrepresented in the sample. This is reflected in the mental health literature reporting that women are twice as likely to report depression than men [225-227], including postnatal depression. The impacts of social inequalities and gender-based oppression on mental health and depression are well documented by the World Health Organization’s Women’s Mental Health report [228] as well as more recently by other researchers [229-231], highlighting that younger women are particularly susceptible to depression, especially in parts of the world where women struggle with additional burdens linked with unequal division of labor based on gender roles. Other predisposing factors to depression include social roles, cultural norms, and higher vulnerability to violence. Furthermore, women may be attracted to using digital devices, as they can give easier and quicker access as well as more privacy and anonymity. It is possible that digital interventions are particularly attractive to younger women especially in postnatal periods [13], and it might be useful to tailor the developments of these interventions to this group in particular.

The fact that fewer men were included in the review may be closely linked with fixed gender roles. For example, it is possible that, for men, admitting that they have depression is seen as a sign of weakness [232], which may prevent them from accessing psychological help [233]. Studies also show differences between men and women in terms of externalizing and internalizing mental health issues [234]. It seems that men are likely to show depression in an external way, for example, by outbursts of anger [235], smoking [236], physical inactivity [237], or alcohol abuse [238]. For women, depression may be more internalized and linked with social isolation [239] and loneliness [236], and they are more likely to seek psychological help. It is possible that different ways of working with men, and young men in particular, may be needed in psychological therapies for depression by involving more physical or creative approaches to psychological support.

More than a half of the studies (105/160, 65.6%) recruited participants from health care settings or used online methods. In terms of the severity of depressive symptoms, participants had depression as their primary diagnosis with a wide range of severity from mild to severe. It was noticeable, however, that fewer studies focused on severe depression and that these studies tended to have smaller sample sizes. It is possible that because digital interventions are in early stages of development, participants with mild to moderate depression might be more suitable for these studies. It is also possible that digital psychotherapies may be less useful for people with severe depression where face-to-face or more individualized therapy may be more appropriate due to the need for careful monitoring of factors such as suicidal ideation or withdrawal. Köhnen et al [8] reported large heterogeneity of studies with acute depression (eg, due to treatment duration) as well as a higher likelihood of negative events during treatment (eg, deterioration, withdrawal, and dropout), which makes it more difficult to research.

Most studies (133/160, 83%) took place in countries from the global north (ie, the United States, Europe, Australia, and Canada), locations where digital technologies and attention to people’s mental health are prominent. However, the impact of race and ethnicity on mental health was often overlooked. As a result, >50% (84/160) of the studies did not report ethnic data about their participants. These omissions are salient in a context where inequalities in health care have been observable among ethnically diverse groups [240], although efforts have been made to adapt treatments for depression for different cultural groups [241,242]. The lack of collection and reporting of ethnic data in the sampled studies demonstrate that researchers need to engage with current debates on the usefulness, desirability, and feasibility of cross-sectional analysis of mental health issues concerning ethnic data.

Nevertheless, ethnicity data were gathered from 57.3% (40,523/70,720) of the participants included in this systematic review. Hispanic, Latino, or Mexican were the most common group of participants (n=16,684, 41.17%), but most of these participants came from 1 study that was conducted in Mexico [56]. Other studies reported fairly diverse race and ethnicity, including 1.76% (n=713) Asian, 1.83% (n=741) Black, 2.1% (n=851) Indigenous, 17.99% (n=7291) White, 8.79% (n=3560) other, and 0.42% (n=172) mixed. It is possible that the diversity in race and ethnicity in the studies where these data were collected suggests a high need for psychological support among non-White groups [240,243-246]. It is important to note, however, that the prevalence of depression and comorbid psychiatric disorders is not uniform across racial and ethnic groups [247,248]. Ongoing debates question the extent to which existing treatments are effective for diverse ethnic populations [249].

The multiple interpretations of the terms further complicate the report of race and ethnicity data. *Race* and *ethnicity* were sometimes used interchangeably, and *nationality* was used at times as a substitute. The 64 studies incorporating race and ethnicity data in their analysis showed a complex spectrum of classification, as evidenced by >30 reported ethnic groups within what could be considered as homogenous groups. Beyond observable physical characteristics (eg, blackness, brownness, or whiteness), studies referred to finer distinctions such as being African, African American, or African Canadian, which indicate how context can impact the experience of depression in individuals of a seemingly homogenous racial group. Similar findings were observed in the category of indigeneity where studies provided further descriptors such as Native Hawaiian, Pacific Islander, American Indian, Maori, and Aboriginal. The use of subgroups suggests the potential for a detailed examination of how ancestry, culture, and geographical context shape the understandings of depression.

Summaries of the key findings are presented in the subsequent sections in relation to the research questions of the study.

Research Question 1: What are the Most Common Types of Digital Psychotherapeutic Interventions for Adults With Depression?

The most common types of digital psychotherapeutic interventions were MoodGYM and Beating the Blues, followed by Good Days Ahead, Health Tips, Mindful Mood Balance, and Project EVO. There were several programs with no human contact (58/160, 36.3%), but more than half of the studies (91/160, 56.9%) referred to blended versions. Over half of the programs (n=93, 58.1%) were web based with only a few involving virtual reality and avatars. Some included relaxation and mindful techniques and others included creative means such as games, music, and comic books. Therapeutically, the vast majority (132/160, 82.5%) used CBT as their framing approach delivered in 6 to 12 weeks in a self-paced manner.

Research Question 2: What Are the Clients' and Practitioners' Perceptions on Helpful and Unhelpful Aspects of Digital Psychotherapeutic Interventions for Adults With Depression?

Clients' and practitioners' perspectives on helpful and unhelpful factors were examined for interventions that did not involve human contact (automated interventions), those that involved human contact (via digital media), and those that combined human contact with automated interventions (blended).

Non–human contact and automated interventions facilitated motivation and offered an opportunity, which otherwise would not have been available. Engaging in one's own time, place, and overall pace was also named as a helpful factor. It was, however, easy for participants to feel overwhelmed and disconnected and to believe that they had simply received generic, common-sense advice. The interventions were experienced as too flexible in some cases, making engagement difficult to sustain over time. In other cases, the tasks were experienced as tedious work.

For interventions that involved human contact, for example, via a website, the most helpful factors from a client's perspective related to usability, which included explanations about tasks. However, "usability" was at times experienced by participants as unhelpful, especially when explanations were not available and there was confusion about how a client should respond. Experience of social support and the perception that their symptoms were improving helped motivate clients to continue using the intervention. However, when they felt no significant changes in their symptoms, clients often became less motivated and even disengaged from the intervention.

For blended approaches, the use of daily practices, reminders, and resources to manage stress and anxiety were seen as motivating. Learning skills to deal with negative thoughts and feelings was also helpful. However, participants commented that they appreciated and felt more attracted to contents that were likable. Blended approaches also provided opportunities to engage with peers (online) and with the therapist; both sets of interactions were seen as helpful. Participants appreciated the flexibility of blended approaches, especially given the difficulty in accessing face-to-face services. The adaptability of blended approaches to patients' needs and preferences has been noted in literature as different interventions can be easily tailored [250]. However, many unhelpful factors were also identified. These included difficulties with the usability of the interventions (eg, lengthy modules and limited attractiveness) and limited acceptability (not all the digital material made immediate sense; some were too structured or too challenging). In terms of opportunity, participants would have liked face-to-face interactions and synchronous groups, which were not available.

Research Question 3: What is the Effectiveness of Digital Psychotherapeutic Interventions for Adults With Depression?

The meta-analysis that included 80 studies with 16,072 participants suggested that there is some certainty in the evidence showing moderate to large effects supporting that digital psychotherapies are likely to reduce depression in comparison with control conditions (Hedges g=–0.61). Furthermore, subgroup analysis revealed a positive effect of digital psychotherapies to reduce scores of depression across all delivery formats. The blended approach appeared to have the greatest positive effects (Hedges g=–0.73 over –0.42 for human contact or –0.40 for no human contact). These findings highlight the benefits of digital interventions. The high impact of the blended approach may be the result of combining the benefits of both human and automated engagement. The value of this type of intervention is also reflected in the range of helpful and unhelpful factors identified in the qualitative and mixed methods studies reviewed and discussed in the *Research Question 2: Clients' and Practitioners' Perspectives on Helpful and Unhelpful Factors Identified in Qualitative and Mixed Methods Studies* section. However, we do not know if there is a statistically significant difference between the different digital approaches since they all indicate moderate and high impact. Further investigation on the differences between these approaches will be needed.

It is also worth considering the group of participants targeted by the different interventions. For example, a recent systematic review and meta-analysis of RCTs of smartphone app–based psychological interventions [16] found greater reduction of symptoms in moderate to severe depression than in mild to moderate depression. However, they commented that the findings might be related to the groups of clients that normally do not access mental health services (eg, health care staff during the COVID-19 pandemic). Our systematic review revealed participants with a range of mild to severe symptoms of depression (with less focus on severe depression), and, as highlighted above, it is possible that blended approaches are more relevant for this group of participants. These interesting findings highlight the need to combine the growth of digital interventions with user preferences and personalized care.

Our subgroup analysis also revealed that all forms of delivery, regardless of the frequency and duration of the sessions, were likely to support the reduction of depression in comparison with control groups. However, there was a greater effect size for interventions delivered more often than once per week (Hedges g=–0.60 vs –0.40 for interventions delivered once per week), a format of delivery that is more easily provided digitally, minimizing cost and human effort. This challenges the common practice of in-person psychotherapy that tends to be once a week. Although there is evidence that higher session frequency could lead to faster recovery, this is not common practice in in-person psychotherapy. These findings [251] provide useful information that could have a direct impact on future digital developments.

Although we were unable to perform a subgroup analysis for digital psychotherapies with different types of delivery and content due to the diversity of interventions, it is well established that creative tools such as creative writing or use of music as well as yoga or relaxation can impact the levels of engagement [252]. Further exploration is needed in this area to develop digital interventions that do not rely so heavily on CBT and focus on meeting the needs of clients from diverse cultural, socioeconomic, or educational backgrounds [253,254]. An example of this could be Arts for the Blues, a creative therapy for depression that has been offered in a digital format [255].

Comparison With Other Reviews

The review examines all digital forms of psychotherapy for depression; this is a new contribution to the existing evidence base, creating a cohesive picture. Unlike other reviews that have focused on specific client populations such as those with severe depression [8], chronic health conditions [9-11], and perinatal depression [13] or children and young people [14-17], this review examines digital forms of psychotherapy for all types of depression affecting all populations. Unlike systematic reviews conducted on specific psychotherapeutic approaches such as CBT only [11,18-20] or specific types of delivery (eg, smartphone app [12]), this review included all forms of psychotherapy and gathered information about different types of digital interventions. Given the increase in mental health concerns and depression worldwide, and the growth of digital psychotherapies in recent years, by bringing all populations, approaches, and digital products together in one review, we have been able to acquire a comprehensive picture of the field. This enables more insight and breadth to identify useful interventions as suggested by the recent NICE [3] guidelines.

As per other reviews [256], this review found that the evidence base would benefit from more diverse study participants, higher quality quantitative studies, and more detailed studies concerning which mechanisms within specific interventions lead to a change of outcomes.

Limitations

The comprehensive nature of the review is certainly one of its major strengths. However, at the same time, depth may have been compromised to accommodate for this “broad stroke” approach. It is possible, for example, that this could be achieved in future studies if qualitative and quantitative studies were reviewed separately.

By contrast, if the aim is to acquire a greater bird’s eye view of this rich and fast-growing field, it is possible that future studies may also consider umbrella reviews (ie, overviews of systematic reviews) including multivariate analysis to obtain a better understanding of how the different reviews relate to each other.

Another limitation relates to the quality of the reviewed studies. Although most qualitative studies (12/13, 92%) were regarded as of “good enough” quality, meeting all MMAT criteria, over a third of the quantitative studies (55/146, 37.6%) had no randomization. In terms of quantitative studies with RCT design (83/146, 56.8%) most studies (80/83, 96%) did not meet all 5 of the MMAT criteria; they were of poor quality and therefore were given less significance in this review.

Despite the risk of bias identified in the RCTs, findings from the meta-analysis and the subgroup analyses offer an invaluable overview of the field, which has not been presented before. The inclusion of 80 studies with 16,072 participants suggests a degree of precision since it is likely that the total sample reaches powered levels (eg, in Grading of Recommendations, Assessment, Development, and Evaluations terms it reached an optimal information size).

Potential Biases in the Review Process

All aspects of the review were checked by >1 reviewer to minimize possible biases. The initial general data extraction was completed by 4 people. This was followed by data extraction focused on specific characteristics of the studies, which was completed by most of the reviewers. Any issues and discrepancies were discussed in weekly meetings. The reviewers came from different professional backgrounds and research skills. The team comprised psychotherapists and arts therapists, for example, with specialized knowledge in the field and qualitative research skills essential for investigating clients’ and practitioners’ perspectives of helpful and unhelpful factors. It also comprised psychologists and allied health professionals with quantitative research skills, essential for calculating effect sizes.

Implications for Practice, Policy, and Research

In the current context of a shortage of and increasing demand for mental health interventions, digital approaches hold great promise. However, this review demonstrates that while digital interventions offer flexibility and autonomy for both providers and participants, blended interventions seem important for a positive experience in the treatment of depression and related conditions. Some of the reviewed studies documented limitations with asynchronous interventions and interventions that have no human contact [41,42,48]; these should not be ignored, especially for susceptible and at-risk populations.

Blended approaches delivered more than once a week seem to be particularly useful for the participants, giving the opportunity for a contact with a therapist as well as a platform that can be used between sessions supporting engagement. Digital interventions seem to be helpful for people from diverse ethnic groups and young women in particular. It will be important to tailor and target the digital interventions specifically for this group, and more research is needed in this area.

Most of the reviewed studies focused on CBT, and it seems important to use alternative theoretical approaches for digital psychotherapies, including creative interventions that can accommodate a range of service users' preferences especially minority populations. In addition, we agree with Fordham et al [257] that new research in the treatment of depression and other disorders ought to shift emphasis from investigating the general effectiveness of interventions to understanding how to achieve a greater effect size for specific populations.

Given the importance of clients' perceptions of what is helpful in digital psychotherapies, it is paramount that interventions that have high levels of acceptability and usability and prioritize a positive user experience are investigated further.

The fact that this review found that motivation to engage in a treatment was one of the most helpful factors from the clients' perspective suggests that digital interventions can have a useful auxiliary role to encourage clients to engage in treatments for depression and mitigate against the risk of dropout.

In the context of the COVID-19 pandemic, it is widely accepted that digital approaches show great promise as treatments for mental health problems [258]. However, this review found more research from the perspectives of clients rather than practitioners and therefore, further research is needed to assist practitioners to be aware of the evidence and efficacy of digital interventions they may use in clinical practice [258].

Nondigital interventions could consider the inclusion of a digital option to encourage participants to remain in treatment for depression. Further research could be done to understand the role of digital interventions for enhancing motivation and the mechanisms by which patients can be encouraged to stay in treatment, thus reducing dropout rates.

Conclusions

The review examines the digital forms of psychotherapy for depression, which is a new contribution to the existing evidence base. Unlike other reviews that have focused on specific client populations and specific psychotherapeutic approaches or modes of delivery, this review included all forms of delivery and gathered information about different types of digital interventions for depression. Given the increase in mental health concerns and depression worldwide and the growth of digital psychotherapies in recent years, by bringing all populations, approaches, and digital products together in 1 review, we have been able to acquire a comprehensive picture of the field, which enables more insights and breadth to identify useful interventions as suggested by the recent NICE guidelines. The review aimed to answer three research questions in relation to digital psychotherapies for depression: (1) the most common types of interventions, (2) the clients' and practitioners' perspectives of helpful and unhelpful aspects, and (3) the effectiveness of the interventions. Digital interventions fell into 3 categories including interventions with no human contact and stand-alone interventions, interventions with human contact, and blended including both stand-alone and "human contact" interventions. Blended interventions formed the biggest group of studies. Most of the digital interventions were web based and involved online psychoeducation drawn primarily from CBT delivered once a week. Blended approaches seem to be particularly useful for participants, giving the opportunity for contact with a therapist as well as a platform that can be used between sessions supporting engagement. In terms of the effectiveness of the digital interventions for depression, meta-analysis revealed a moderate to large effect on depression. Analysis of studies with blended approaches revealed the largest effect size in comparison to interventions involving human contact only or no human contact. In addition, the review found that digital interventions seem to be particularly helpful for people from diverse ethnic groups and young women and therefore new research in the treatment of depression ought to shift emphasis from investigating the general effectiveness of interventions to understanding how to achieve a greater effect size for specific populations.

Acknowledgments

This systematic review was funded by the UK Research and Innovation Strategic Priorities Fund (2020-2021) administered by the University of Salford, Greater Manchester, the United Kingdom.

Data Availability

The data sets analyzed during this study are available in Multimedia Appendix 2.

No new data were generated during this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Studies in systematic review.
[DOCX File , 59 KB - mental_v11i1e55500_app1.docx]

Multimedia Appendix 2
Characteristics of reviewed studies.
[XLSX File (Microsoft Excel File), 143 KB - mental_v11i1e55500_app2.xlsx]

Multimedia Appendix 3
Risk of bias grading and quality appraisal tables.
[DOCX File , 43 KB - mental_v11i1e55500_app3.docx]

Multimedia Appendix 4
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist.
[DOCX File , 31 KB - mental_v11i1e55500_app4.docx]

References

1. Depressive disorder (depression). World Health Organization. 2023 Mar 31. URL: https://www.who.int/news-room/fact-sheets/detail/depression [accessed 2023-06-22]
2. Global burden of disease study 2019 (GBD 2019). Global Burden of Disease Collaborative Network. 2019 Oct. URL: https://vizhub.healthdata.org/gbd-results/?params=gbd-api-2019-permalink/d780dffbe8a381b25e1416884959e88b [accessed 2023-06-22]
3. Depression in adults: treatment and management. NG 222. National Institute for Health and Care Excellence. 2022 Jun 22. URL: https://www.nice.org.uk/guidance/ng222/chapter/Recommendations [accessed 2023-06-22]
4. Liverpool S, Moinuddin M, Aithal S, Owen M, Bracegirdle K, Caravotta M, et al. Mental health and wellbeing of further and higher education students returning to face-to-face learning after COVID-19 restrictions. PLoS One 2023 Jan 23;18(1):e0280689 [FREE Full text] [doi: 10.1371/journal.pone.0280689] [Medline: 36689440]
5. IAPT guide for delivering treatment remotely during the coronavirus pandemic. National Health Service England and National Health Service Improvements. 2020 Mar 26. URL: https://www.yhscn.nhs.uk/media/PDFs/mhdn/Mental%20Health/COVID-19/C0073_MHLDA_Covid19_Guidance_IAPT_FINAL.pdf [accessed 2023-06-22]
6. The five year forward view for mental health. The Independent Mental Health Taskforce to the NHS in England. National Health Service England. 2016 Feb. URL: https://www.england.nhs.uk/wp-content/uploads/2016/02/Mental-Health-Taskforce-FYFV-final.pdf [accessed 2023-06-22]
7. NHS long term plan. National Health Service England. 2019 Jan. URL: https://www.longtermplan.nhs.uk/wp-content/uploads/2019/08/nhs-long-term-plan-version-1.2.pdf [accessed 2023-06-22]
8. Köhnen M, Dreier M, Seeralan T, Kriston L, Härter M, Baumeister H, et al. Evidence on technology-based psychological interventions in diagnosed depression: systematic review. JMIR Ment Health 2021 Feb 10;8(2):e21700 [FREE Full text] [doi: 10.2196/21700] [Medline: 33565981]
9. White KM, Williamson C, Bergou N, Oetzmann C, de Angel V, Matcham F, et al. A systematic review of engagement reporting in remote measurement studies for health symptom tracking. NPJ Digit Med 2022 Jun 29;5(1):82 [FREE Full text] [doi: 10.1038/s41746-022-00624-7] [Medline: 35768544]
10. Paalimäki-Paakki K, Virtanen M, Henner A, Nieminen MT, Kääriäinen M. Effectiveness of digital counseling environments on anxiety, depression, and adherence to treatment among patients who are chronically ill: systematic review. J Med Internet Res 2022 Jan 06;24(1):e30077 [FREE Full text] [doi: 10.2196/30077] [Medline: 34989681]
11. Lee S, Oh JW, Park KM, Lee S, Lee E. Digital cognitive behavioral therapy for insomnia on depression and anxiety: a systematic review and meta-analysis. NPJ Digit Med 2023 Mar 25;6(1):52 [FREE Full text] [doi: 10.1038/s41746-023-00800-3] [Medline: 36966184]
12. Serrano-Ripoll MJ, Zamanillo-Campos R, Fiol-DeRoque MA, Castro A, Ricci-Cabello I. Impact of smartphone app-based psychological interventions for reducing depressive symptoms in people with depression: systematic literature review and meta-analysis of randomized controlled trials. JMIR Mhealth Uhealth 2022 Jan 27;10(1):e29621 [FREE Full text] [doi: 10.2196/29621] [Medline: 35084346]
13. Motrico E, Conejo-Cerón S, Martín-Gómez C, Gómez I, Fonseca A, Moreno-Peral P. Effectiveness of web-based and mobile-based psychological interventions to prevent perinatal depression: study protocol for a systematic review and meta-analysis of randomized controlled trials. Internet Interv 2021 Dec;26:100471 [FREE Full text] [doi: 10.1016/j.invent.2021.100471] [Medline: 34754756]
14. Ivlev I, Beil TL, Haynes JS, Patnode CD. Rapid evidence review of digital cognitive-behavioral therapy for adolescents with depression. J Adolesc Health 2022 Jul;71(1):14-29. [doi: 10.1016/j.jadohealth.2022.01.220] [Medline: 35256238]

15. Wickersham A, Barack T, Cross L, Downs J. Computerized cognitive behavioral therapy for treatment of depression and anxiety in adolescents: systematic review and meta-analysis. J Med Internet Res 2022 Apr 11;24(4):e29842 [FREE Full text] [doi: 10.2196/29842] [Medline: 35404263]
16. Li SH, Achilles MR, Werner-Seidler A, Beames JR, Subotic-Kerry M, O'Dea B. Appropriate use and operationalization of adherence to digital cognitive behavioral therapy for depression and anxiety in youth: systematic review. JMIR Ment Health 2022 Aug 17;9(8):e37640 [FREE Full text] [doi: 10.2196/37640] [Medline: 35976180]
17. Martinez K, Menéndez-Menéndez MI, Bustillo A. Awareness, prevention, detection, and therapy applications for depression and anxiety in serious games for children and adolescents: systematic review. JMIR Serious Games 2021 Dec 16;9(4):e30482 [FREE Full text] [doi: 10.2196/30482] [Medline: 34927589]
18. Chow DY, Jiang X, You JH. Information technology-based versus face-to-face cognitive-behavioural therapy for anxiety and depression: a systematic review and meta-analysis. J Affect Disord 2022 Aug 01;310:429-440. [doi: 10.1016/j.jad.2022.05.048] [Medline: 35577156]
19. Kambeitz-Ilankovic L, Rzayeva U, Völkel L, Wenzel J, Weiske J, Jessen F, et al. A systematic review of digital and face-to-face cognitive behavioral therapy for depression. NPJ Digit Med 2022 Sep 15;5(1):144 [FREE Full text] [doi: 10.1038/s41746-022-00677-8] [Medline: 36109583]
20. Fadipe MF, Aggarwal S, Johnson C, Beauchamp JE. Effectiveness of online cognitive behavioural therapy on quality of life in adults with depression: a systematic review. J Psychiatr Ment Health Nurs 2023 Oct 03;30(5):885-898. [doi: 10.1111/jpm.12924] [Medline: 37010913]
21. Andrews G, Basu A, Cuijpers P, Craske M, McEvoy P, English C, et al. Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: an updated meta-analysis. J Anxiety Disord 2018 Apr;55:70-78 [FREE Full text] [doi: 10.1016/j.janxdis.2018.01.001] [Medline: 29422409]
22. Renn BN, Hoeft TJ, Lee HS, Bauer AM, Areán PA. Preference for in-person psychotherapy versus digital psychotherapy options for depression: survey of adults in the U.S. NPJ Digit Med 2019 Feb 11;2(1):6 [FREE Full text] [doi: 10.1038/s41746-019-0077-1] [Medline: 31304356]
23. Berger T. The therapeutic alliance in internet interventions: a narrative review and suggestions for future research. Psychother Res 2017 Sep;27(5):511-524. [doi: 10.1080/10503307.2015.1119908] [Medline: 26732852]
24. Russell GI. Screen Relations: The Limits of Computer-Mediated Psychoanalysis and Psychotherapy. New York City, NY: Routledge; 2015.
25. Martinez-Martin N, Kreitmair K. Ethical issues for direct-to-consumer digital psychotherapy apps: addressing accountability, data protection, and consent. JMIR Ment Health 2018 Apr 23;5(2):e32 [FREE Full text] [doi: 10.2196/mental.9423] [Medline: 29685865]
26. Smith K, Moller N, Cooper M, Gabriel L, Roddy J, Sheehy R. Video counselling and psychotherapy: a critical commentary on the evidence base. Couns Psychother Res 2021 Jun 25;22(1):92-97. [doi: 10.1002/capr.12436]
27. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
28. Bowling A, Ebrahim S. Handbook of Health Research Methods. New York, NY: Open University Press; 2005.
29. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ 2014 Mar 07;348:g1687 [FREE Full text] [doi: 10.1136/bmj.g1687] [Medline: 24609605]
30. Higgins JP, Green S. Cochrane Handbook for Systematic Reviews of Interventions. Hoboken, NJ: John Wiley & Sons; 2011.
31. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018 Dec 18;34(4):285-291. [doi: 10.3233/EFI-180221]
32. Popay J, Roberts H, Sowden A. Guidance on the conduct of narrative synthesis in systematic reviews: a product from the ESRC methods programme. Lancaster University. 2006. URL: https://www.lancaster.ac.uk/media/lancaster-university/content-assets/documents/fhm/dhr/chir/NSsynthesisguidanceVersion1-April2006.pdf [accessed 2023-06-22]
33. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42 [FREE Full text] [doi: 10.1186/1748-5908-6-42] [Medline: 21513547]
34. Liverpool S, Mota C, Sales CM, Čuš A, Carletto S, Hancheva C, et al. Engaging children and young people in digital mental health interventions: systematic review of modes of delivery, facilitators, and barriers. J Med Internet Res 2020 Jun 23;22(6):e16317 [FREE Full text] [doi: 10.2196/16317] [Medline: 32442160]
35. Andrade C. Mean difference, standardized mean difference (SMD), and their use in meta-analysis: as simple as it gets. J Clin Psychiatry 2020 Sep 22;81(5):20f13681 [FREE Full text] [doi: 10.4088/JCP.20f13681] [Medline: 32965803]
36. Higgins JP, Eldridge S, Li T. Including variants on randomized trials. In: Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, editors. Cochrane Handbook for Systematic Reviews of Interventions. Version 6.4. Hoboken, NJ: John Wiley & Sons; 2023.
37. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634 [FREE Full text] [doi: 10.1136/bmj.315.7109.629] [Medline: 9310563]

38. Sterne JA, Sutton AJ, Ioannidis JP, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ 2011 Jul 22;343(jul22 1):d4002. [doi: 10.1136/bmj.d4002] [Medline: 21784880]
39. Borenstein M, Higgins JP. Meta-analysis and subgroups. Prev Sci 2013 Apr 13;14(2):134-143. [doi: 10.1007/s11121-013-0377-7] [Medline: 23479191]
40. Ellis LA, Campbell AJ, Sethi S, O'Dea BM. Comparative randomized trial of an online cognitive-behavioral therapy program and an online support group for depression and anxiety. J Cyber Ther Rehabil 2011;4(4):467 [FREE Full text]
41. Gerhards SA, Abma TA, Arntz A, de Graaf LE, Evers SM, Huibers MJ, et al. Improving adherence and effectiveness of computerised cognitive behavioural therapy without support for depression: a qualitative study on patient experiences. J Affect Disord 2011 Mar;129(1-3):117-125. [doi: 10.1016/j.jad.2010.09.012] [Medline: 20889214]
42. Knowles SE, Lovell K, Bower P, Gilbody S, Littlewood E, Lester H. Patient experience of computerised therapy for depression in primary care. BMJ Open 2015 Nov 30;5(11):e008581 [FREE Full text] [doi: 10.1136/bmjopen-2015-008581] [Medline: 26621513]
43. Lucassen MF, Hatcher S, Stasiak K, Fleming T, Shepherd M, Merry SN. The views of lesbian, gay and bisexual youth regarding computerised self-help for depression: an exploratory study. Adv Ment Health 2014 Dec 17;12(1):22-33. [doi: 10.5172/jamh.2013.12.1.22]
44. Richards D, Timulak L. Client-identified helpful and hindering events in therapist-delivered vs. self-administered online cognitive-behavioural treatments for depression in college students. Couns Psychol Q 2012 Sep;25(3):251-262. [doi: 10.1080/09515070.2012.703129]
45. Richards D, Timulak L. Satisfaction with therapist-delivered vs. self-administered online cognitive behavioural treatments for depression symptoms in college students. Br J Guid Counc 2013 Apr;41(2):193-207. [doi: 10.1080/03069885.2012.726347]
46. Rozbroj T, Lyons A, Pitts M, Mitchell A, Christensen H. Improving self-help e-therapy for depression and anxiety among sexual minorities: an analysis of focus groups with lesbians and gay men. J Med Internet Res 2015 Mar 11;17(3):e66 [FREE Full text] [doi: 10.2196/jmir.4013] [Medline: 25761775]
47. Schneider J, Sarrami Foroushani P, Grime P, Thornicroft G. Acceptability of online self-help to people with depression: users' views of MoodGYM versus informational websites. J Med Internet Res 2014 Mar 28;16(3):e90 [FREE Full text] [doi: 10.2196/jmir.2871] [Medline: 24681717]
48. Walsh S, Szymczynska P, Taylor SJ, Priebe S. The acceptability of an online intervention using positive psychology for depression: a qualitative study. Internet Interv 2018 Sep;13:60-66 [FREE Full text] [doi: 10.1016/j.invent.2018.07.003] [Medline: 30206520]
49. Shkel J, Green G, Le S, Kaveladze B, Marcotte V, Rushton K, et al. Understanding users' experiences of a novel web-based cognitive behavioral therapy platform for depression and anxiety: qualitative interviews from pilot trial participants. JMIR Form Res 2023 Jun 20;7:e46062 [FREE Full text] [doi: 10.2196/46062] [Medline: 37338967]
50. Fuller-Tyszkiewicz M, Richardson B, Klein B, Skouteris H, Christensen H, Austin D, et al. A mobile app-based intervention for depression: end-user and expert usability testing study. JMIR Ment Health 2018 Aug 23;5(3):e54 [FREE Full text] [doi: 10.2196/mental.9445] [Medline: 30139722]
51. Ekberg S, Barnes RK, Kessler DS, Malpass A, Shaw AR. Managing clients' expectations at the outset of online Cognitive Behavioural Therapy (CBT) for depression. Health Expect 2016 Jun 02;19(3):557-569 [FREE Full text] [doi: 10.1111/hex.12227] [Medline: 25088009]
52. Lemma A, Fonagy P. Feasibility study of a psychodynamic online group intervention for depression. Psychoanal Psychol 2013 Jul;30(3):367-380. [doi: 10.1037/a0033239]
53. Nogami W, Nakagawa A, Kato N, Sasaki Y, Kishimoto T, Horikoshi M, et al. Efficacy and acceptability of remote cognitive behavioral therapy for patients with major depressive disorder in Japanese clinical settings: a case series. Cogn Behav Pract 2022 May 05;30(4):635-645 [FREE Full text] [doi: 10.1016/j.cbpra.2022.04.002] [Medline: 36059862]
54. Eichenberg C, Schott M, Sawyer A, Aumayr G, Plößnig M. Feasibility and conceptualization of an e-Mental health treatment for depression in older adults: mixed-methods study. JMIR Aging 2018 Oct 23;1(2):e10973 [FREE Full text] [doi: 10.2196/10973] [Medline: 31518235]
55. Felder J, Dimidjian S, Beck A, Boggs JM, Segal Z. Mindful mood balance: a case report of Web-based treatment of residual depressive symptoms. Perm J 2014 Nov 3;18(4):58-62 [FREE Full text] [doi: 10.7812/TPP/14-019] [Medline: 25141988]
56. Lara MA, Tiburcio M, Aguilar Abrego A, Sánchez-Solís A. A four-year experience with a web-based self-help intervention for depressive symptoms in Mexico. Rev Panam Salud Publica 2014;35(5-6):399-406. [Medline: 25211568]
57. Schlosser DA, Campellone TR, Truong B, Anguera JA, Vergani S, Vinogradov S, et al. The feasibility, acceptability, and outcomes of PRIME-D: a novel mobile intervention treatment for depression. Depress Anxiety 2017 Jun 18;34(6):546-554 [FREE Full text] [doi: 10.1002/da.22624] [Medline: 28419621]
58. Seshu U, Khan HA, Bhardwaj M, Sangeetha C, Aarthi G, John S, et al. A qualitative study on the use of mobile-based intervention for perinatal depression among perinatal mothers in rural Bihar, India. Int J Soc Psychiatry 2021 Aug 16;67(5):467-471. [doi: 10.1177/0020764020966003] [Medline: 33059490]

59. Westerhof G, Lamers SM, Postel MG, Bohlmeijer ET. Online therapy for depressive symptoms: an evaluation of counselor-led and peer-supported life review therapy. Gerontologist 2019 Jan 09;59(1):135-146. [doi: 10.1093/geront/gnx140] [Medline: 28961958]
60. Mol M, Dozeman E, Provoost S, van Schaik A, Riper H, Smit JH. Behind the scenes of online therapeutic feedback in blended therapy for depression: mixed-methods observational study. J Med Internet Res 2018 May 03;20(5):e174 [FREE Full text] [doi: 10.2196/jmir.9890] [Medline: 29724708]
61. Holst A, Nejati S, Björkelund C, Eriksson MC, Hange D, Kivi M, et al. Patients' experiences of a computerised self-help program for treating depression - a qualitative study of internet mediated cognitive behavioural therapy in primary care. Scand J Prim Health Care 2017 Mar;35(1):46-53 [FREE Full text] [doi: 10.1080/02813432.2017.1288813] [Medline: 28277055]
62. Danaher BG, Milgrom J, Seeley JR, Stuart S, Schembri C, Tyler MS, et al. Web-based intervention for postpartum depression: formative research and design of the MomMoodBooster program. JMIR Res Protoc 2012 Nov 22;1(2):e18 [FREE Full text] [doi: 10.2196/resprot.2329] [Medline: 23612274]
63. Wilhelmsen M, Høifødt RS, Kolstrup N, Waterloo K, Eisemann M, Chenhall R, et al. Norwegian general practitioners' perspectives on implementation of a guided web-based cognitive behavioral therapy for depression: a qualitative study. J Med Internet Res 2014 Sep 10;16(9):e208 [FREE Full text] [doi: 10.2196/jmir.3556] [Medline: 25208886]
64. Currie SL, McGrath PJ, Day V. Development and usability of an online CBT program for symptoms of moderate depression, anxiety, and stress in post-secondary students. Comput Human Behav 2010 Nov;26(6):1419-1426. [doi: 10.1016/j.chb.2010.04.020]
65. Hatcher S, Whittaker R, Patton M, Miles WS, Ralph N, Kercher K, et al. Web-based therapy plus support by a coach in depressed patients referred to secondary mental health care: randomized controlled trial. JMIR Ment Health 2018 Jan 23;5(1):e5 [FREE Full text] [doi: 10.2196/mental.8510] [Medline: 29362207]
66. Gega L, Smith J, Reynolds S. Cognitive behaviour therapy (CBT) for depression by computer vs. therapist: patient experiences and therapeutic processes. Psychother Res 2013 Mar;23(2):218-231. [doi: 10.1080/10503307.2013.766941] [Medline: 23390994]
67. Sergeant S, Mongrain M. An online optimism intervention reduces depression in pessimistic individuals. J Consult Clin Psychol 2014 Apr;82(2):263-274. [doi: 10.1037/a0035536] [Medline: 24417602]
68. Geraedts AS, van Dongen JM, Kleiboer AM, Wiezer NM, van Mechelen W, Cuijpers P, et al. Economic evaluation of a web-based guided self-help intervention for employees with depressive symptoms: results of a randomized controlled trial. J Occup Environ Med 2015 Jun;57(6):666-675. [doi: 10.1097/JOM.0000000000000423] [Medline: 25741795]
69. Anguera JA, Gunning FM, Areán PA. Improving late life depression and cognitive control through the use of therapeutic video game technology: a proof-of-concept randomized trial. Depress Anxiety 2017 Jun 03;34(6):508-517. [doi: 10.1002/da.22588] [Medline: 28052513]
70. Jelinek L, Arlt S, Moritz S, Schröder J, Westermann S, Cludius B. Brief web-based intervention for depression: randomized controlled trial on behavioral activation. J Med Internet Res 2020 Mar 26;22(3):e15312 [FREE Full text] [doi: 10.2196/15312] [Medline: 32213470]
71. O'Toole MS, Arendt MB, Pedersen CM. Testing an app-assisted treatment for suicide prevention in a randomized controlled trial: effects on suicide risk and depression. Behav Ther 2019 Mar;50(2):421-429. [doi: 10.1016/j.beth.2018.07.007] [Medline: 30824256]
72. Keller A, Babl A, Berger T, Schindler L. Efficacy of the web-based PaarBalance program on relationship satisfaction, depression and anxiety - a randomized controlled trial. Internet Interv 2021 Mar;23:100360 [FREE Full text] [doi: 10.1016/j.invent.2020.100360] [Medline: 33520669]
73. Kingston J, Becker L, Woeginger J, Ellett L. A randomised trial comparing a brief online delivery of mindfulness-plus-values versus values only for symptoms of depression: does baseline severity matter? J Affect Disord 2020 Nov 01;276:936-944. [doi: 10.1016/j.jad.2020.07.087] [Medline: 32745830]
74. Blackwell SE, Browning M, Mathews A, Pictet A, Welch J, Davies J, et al. Positive imagery-based cognitive bias modification as a web-based treatment tool for depressed adults: a randomized controlled trial. Clin Psychol Sci 2015 Jan 06;3(1):91-111 [FREE Full text] [doi: 10.1177/2167702614560746] [Medline: 25984421]
75. El Morr C, Ritvo P, Ahmad F, Moineddin R, MVC Team. Effectiveness of an 8-week web-based mindfulness virtual community intervention for university students on symptoms of stress, anxiety, and depression: randomized controlled trial. JMIR Ment Health 2020 Jul 17;7(7):e18595 [FREE Full text] [doi: 10.2196/18595] [Medline: 32554380]
76. Douma M, Maurice-Stam H, Gorter B, Krol Y, Verkleij M, Wiltink L, et al. Online psychosocial group intervention for parents: positive effects on anxiety and depression. J Pediatr Psychol 2021 Feb 19;46(2):123-134 [FREE Full text] [doi: 10.1093/jpepsy/jsaa102] [Medline: 33230541]
77. Nakao S, Nakagawa A, Oguchi Y, Mitsuda D, Kato N, Nakagawa Y, et al. Web-based cognitive behavioral therapy blended with face-to-face sessions for major depression: randomized controlled trial. J Med Internet Res 2018 Sep 21;20(9):e10743 [FREE Full text] [doi: 10.2196/10743] [Medline: 30249583]
78. Calkins AW, McMorran KE, Siegle GJ, Otto MW. The effects of computerized cognitive control training on community adults with depressed mood. Behav Cogn Psychother 2014 Mar 03;43(5):578-589. [doi: 10.1017/s1352465814000046]

79. Rollman BL, Herbeck Belnap B, Abebe KZ, Spring MB, Rotondi AJ, Rothenberger SD, et al. Effectiveness of online collaborative care for treating mood and anxiety disorders in primary care: a randomized clinical trial. JAMA Psychiatry 2018 Jan 01;75(1):56-64 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.3379] [Medline: 29117275]
80. Vernmark K, Lenndin J, Bjärehed J, Carlsson M, Karlsson J, Oberg J, et al. Internet administered guided self-help versus individualized e-mail therapy: a randomized trial of two versions of CBT for major depression. Behav Res Ther 2010 May;48(5):368-376. [doi: 10.1016/j.brat.2010.01.005] [Medline: 20152960]
81. Ahmedani BK, Belville-Robertson T, Hirsch A, Jurayj A. An online mental health and wellness intervention supplementing standard care of depression and anxiety. Arch Psychiatr Nurs 2016 Dec;30(6):666-670. [doi: 10.1016/j.apnu.2016.03.003] [Medline: 27888957]
82. Alavi N, Hirji A, Sutton C, Naeem F. Online CBT is effective in overcoming cultural and language barriers in patients with depression. J Psychiatr Pract 2016 Jan;22(1):2-8. [doi: 10.1097/PRA.0000000000000119] [Medline: 26813483]
83. Alavi N, Moghimi E, Stephenson C, Gutierrez G, Jagayat J, Kumar A, et al. Comparison of online and in-person cognitive behavioral therapy in individuals diagnosed with major depressive disorder: a non-randomized controlled trial. Front Psychiatry 2023;14:1113956 [FREE Full text] [doi: 10.3389/fpsyt.2023.1113956] [Medline: 37187863]
84. Andrews B, Klein B, Corboy D, McLaren S, Watson S. Video chat therapist assistance in an adaptive digital intervention for anxiety and depression: reflections from participants and therapists. Prof Psychol Res Pr 2023 Dec;54(6):418-429. [doi: 10.1037/pro0000527]
85. Andrews B, Klein B, van Nguyen H, Corboy D, McLaren S, Watson S. Efficacy of a digital mental health biopsychosocial transdiagnostic intervention with or without therapist assistance for adults with anxiety and depression: adaptive randomized controlled trial. J Med Internet Res 2023 Jun 12;25:e45135 [FREE Full text] [doi: 10.2196/45135] [Medline: 37307046]
86. Anguera JA, Jordan JT, Castaneda D, Gazzaley A, Areán PA. Conducting a fully mobile and randomised clinical trial for depression: access, engagement and expense. BMJ Innov 2016 Jan;2(1):14-21 [FREE Full text] [doi: 10.1136/bmjinnov-2015-000098] [Medline: 27019745]
87. Arean PA, Hallgren KA, Jordan JT, Gazzaley A, Atkins DC, Heagerty PJ, et al. The use and effectiveness of mobile apps for depression: results from a fully remote clinical trial. J Med Internet Res 2016 Dec 20;18(12):e330 [FREE Full text] [doi: 10.2196/jmir.6482] [Medline: 27998876]
88. Araghi NM, Zarei MA, Saei S, Yousefi Nodeh HR, Mahmoudi E. The effect of online cognitive behavioral therapy on depressive symptoms in recovered patients with COVID-19. J Educ Health Promot 2022;11:70 [FREE Full text] [doi: 10.4103/jehp.jehp_727_21] [Medline: 35372626]
89. Baikie KA, Geerligs L, Wilhelm K. Expressive writing and positive writing for participants with mood disorders: an online randomized controlled trial. J Affect Disord 2012 Feb;136(3):310-319. [doi: 10.1016/j.jad.2011.11.032] [Medline: 22209127]
90. Bantjes J, Kazdin AE, Cuijpers P, Breet E, Dunn-Coetzee M, Davids C, et al. A web-based group cognitive behavioral therapy intervention for symptoms of anxiety and depression among university students: open-label, pragmatic trial. JMIR Ment Health 2021 May 27;8(5):e27400 [FREE Full text] [doi: 10.2196/27400] [Medline: 34042598]
91. Birney AJ, Gunn R, Russell JK, Ary DV. MoodHacker mobile web app with email for adults to self-manage mild-to-moderate depression: randomized controlled trial. JMIR Mhealth Uhealth 2016 Jan 26;4(1):e8 [FREE Full text] [doi: 10.2196/mhealth.4231] [Medline: 26813737]
92. Bowie CR, Gupta M, Holshausen K, Jokic R, Best M, Milev R. Cognitive remediation for treatment-resistant depression: effects on cognition and functioning and the role of online homework. J Nerv Ment Dis 2013 Aug;201(8):680-685. [doi: 10.1097/NMD.0b013e31829c5030] [Medline: 23896849]
93. Bowler JO, Mackintosh B, Dunn BD, Mathews A, Dalgleish T, Hoppitt L. A comparison of cognitive bias modification for interpretation and computerized cognitive behavior therapy: effects on anxiety, depression, attentional control, and interpretive bias. J Consult Clin Psychol 2012 Dec;80(6):1021-1033 [FREE Full text] [doi: 10.1037/a0029932] [Medline: 22963595]
94. Buntrock C, Ebert D, Lehr D, Riper H, Smit F, Cuijpers P, et al. Effectiveness of a web-based cognitive behavioural intervention for subthreshold depression: pragmatic randomised controlled trial. Psychother Psychosom 2015;84(6):348-358. [doi: 10.1159/000438673] [Medline: 26398885]
95. Christensen H, Farrer L, Batterham PJ, Mackinnon A, Griffiths KM, Donker T. The effect of a web-based depression intervention on suicide ideation: secondary outcome from a randomised controlled trial in a helpline. BMJ Open 2013 Jun 28;3(6):e002886 [FREE Full text] [doi: 10.1136/bmjopen-2013-002886] [Medline: 23811172]
96. Clarke J, Proudfoot J, Birch MR, Whitton AE, Parker G, Manicavasagar V, et al. Effects of mental health self-efficacy on outcomes of a mobile phone and web intervention for mild-to-moderate depression, anxiety and stress: secondary analysis of a randomised controlled trial. BMC Psychiatry 2014 Sep 26;14:272 [FREE Full text] [doi: 10.1186/s12888-014-0272-1] [Medline: 25252853]
97. Cluxton-Keller F, Buteau J, Williams M, Stolte P, Monroe-Cassel M, Bruce ML. Engaging rural young mothers in a technology-based intervention for depression. Child Youth Serv 2019;40(2):158-183 [FREE Full text] [doi: 10.1080/0145935X.2018.1561264] [Medline: 31274940]

98. Collins S, Byrne M, Hawe J, O'Reilly G. Evaluation of a computerized cognitive behavioural therapy programme, MindWise (2.0), for adults with mild-to-moderate depression and anxiety. Br J Clin Psychol 2018 Jun;57(2):255-269. [doi: 10.1111/bjc.12165] [Medline: 29197102]
99. Danaher BG, Milgrom J, Seeley JR, Stuart S, Schembri C, Tyler MS, et al. MomMoodBooster web-based intervention for postpartum depression: feasibility trial results. J Med Internet Res 2013 Nov 04;15(11):e242 [FREE Full text] [doi: 10.2196/jmir.2876] [Medline: 24191345]
100. Darnell D, Pullmann MD, Hull TD, Chen S, Areán P. Predictors of disengagement and symptom improvement among adults with depression enrolled in talkspace, a technology-mediated psychotherapy platform: naturalistic observational study. JMIR Form Res 2022 Jun 22;6(6):e36521 [FREE Full text] [doi: 10.2196/36521] [Medline: 35731563]
101. de Graaf LE, Hollon SD, Huibers MJ. Predicting outcome in computerized cognitive behavioral therapy for depression in primary care: a randomized trial. J Consult Clin Psychol 2010 Apr;78(2):184-189. [doi: 10.1037/a0018324] [Medline: 20350029]
102. de Graaf LE, Gerhards SA, Arntz A, Riper H, Metsemakers JF, Evers SM, et al. One-year follow-up results of unsupported online computerized cognitive behavioural therapy for depression in primary care: a randomized trial. J Behav Ther Exp Psychiatry 2011 Mar;42(1):89-95. [doi: 10.1016/j.jbtep.2010.07.003] [Medline: 20723885]
103. Dehn L, Kater L, Piefke M, Botsch M, Driessen M, Beblo T. Training in a comprehensive everyday-like virtual reality environment compared to computerized cognitive training for patients with depression. Comput Hum Behav 2018 Feb;79:40-52. [doi: 10.1016/j.chb.2017.10.019]
104. Dimidjian S, Beck A, Felder JN, Boggs JM, Gallop R, Segal ZV. Web-based mindfulness-based cognitive therapy for reducing residual depressive symptoms: an open trial and quasi-experimental comparison to propensity score matched controls. Behav Res Ther 2014 Dec;63:83-89 [FREE Full text] [doi: 10.1016/j.brat.2014.09.004] [Medline: 25461782]
105. Duarte A, Walker S, Littlewood E, Brabyn S, Hewitt C, Gilbody S, et al. Cost-effectiveness of computerized cognitive-behavioural therapy for the treatment of depression in primary care: findings from the Randomised Evaluation of the Effectiveness and Acceptability of Computerised Therapy (REEACT) trial. Psychol Med 2017 Jul;47(10):1825-1835 [FREE Full text] [doi: 10.1017/S0033291717000289] [Medline: 28228182]
106. Ebert DD, Buntrock C, Lehr D, Smit F, Riper H, Baumeister H, et al. Effectiveness of web- and mobile-based treatment of subthreshold depression with adherence-focused guidance: a single-blind randomized controlled trial. Behav Ther 2018 Jan;49(1):71-83. [doi: 10.1016/j.beth.2017.05.004] [Medline: 29405923]
107. Eriksson MC, Kivi M, Hange D, Petersson EL, Ariai N, Häggblad P, et al. Long-term effects of Internet-delivered cognitive behavioral therapy for depression in primary care - the PRIM-NET controlled trial. Scand J Prim Health Care 2017 Jun;35(2):126-136 [FREE Full text] [doi: 10.1080/02813432.2017.1333299] [Medline: 28585868]
108. Farrer L, Christensen H, Griffiths KM, Mackinnon A. Web-based cognitive behavior therapy for depression with and without telephone tracking in a national helpline: secondary outcomes from a randomized controlled trial. J Med Internet Res 2012 Jun 27;14(3):e68 [FREE Full text] [doi: 10.2196/jmir.1859] [Medline: 22738715]
109. Fatori D, Zuccolo P, Xavier MO, Matijasevich A, Polanczyk GV. Smartphone-assisted online brief cognitive behavioral therapy to treat maternal depression: findings of a randomized controlled trial. Braz J Psychiatry 2023 Mar 13;45(1):50-53 [FREE Full text] [doi: 10.47626/1516-4446-2022-2679] [Medline: 36049141]
110. Figueroa CA, DeMasi O, Hernandez-Ramos R, Aguilera A. Who benefits most from adding technology to depression treatment and how? An analysis of engagement with a texting adjunct for psychotherapy. Telemed J E Health 2021 Jan;27(1):39-46 [FREE Full text] [doi: 10.1089/tmj.2019.0248] [Medline: 32213012]
111. Fogarty AS, Proudfoot J, Whittle EL, Clarke J, Player MJ, Christensen H, et al. Preliminary evaluation of a brief web and mobile phone intervention for men with depression: men's positive coping strategies and associated depression, resilience, and work and social functioning. JMIR Ment Health 2017 Aug 10;4(3):e33 [FREE Full text] [doi: 10.2196/mental.7769] [Medline: 28798009]
112. Forman-Hoffman VL, Nelson BW, Ranta K, Nazander A, Hilgert O, de Quevedo J. Significant reduction in depressive symptoms among patients with moderately-severe to severe depressive symptoms after participation in a therapist-supported, evidence-based mobile health program delivered via a smartphone app. Internet Interv 2021 Sep;25:100408 [FREE Full text] [doi: 10.1016/j.invent.2021.100408] [Medline: 34401367]
113. Forman-Hoffman VL, Sihvonen S, Wielgosz J, Kuhn E, Nelson BW, Peiper NC, et al. Therapist-supported digital mental health intervention for depressive symptoms: a randomized clinical trial. J Affect Disord 2024 Mar 15;349:494-501 [FREE Full text] [doi: 10.1016/j.jad.2024.01.057] [Medline: 38211747]
114. Fundoiano-Hershcovitz Y, Breuer Asher I, Ritholz MD, Feniger E, Manejwala O, Goldstein P. Specifying the efficacy of digital therapeutic tools for depression and anxiety: retrospective, 2-cohort, real-world analysis. J Med Internet Res 2023 Sep 22;25:e47350 [FREE Full text] [doi: 10.2196/47350] [Medline: 37738076]
115. Gerhards SA, de Graaf LE, Jacobs LE, Severens JL, Huibers MJ, Arntz A, et al. Economic evaluation of online computerised cognitive-behavioural therapy without support for depression in primary care: randomised trial. Br J Psychiatry 2010 Apr;196(4):310-318. [doi: 10.1192/bjp.bp.109.065748] [Medline: 20357309]

116. Gilbody S, Littlewood E, Hewitt C, Brierley G, Tharmanathan P, Araya R, et al. Computerised cognitive behaviour therapy (cCBT) as treatment for depression in primary care (REEACT trial): large scale pragmatic randomised controlled trial. BMJ 2015 Nov 11;351:h5627 [FREE Full text] [doi: 10.1136/bmj.h5627] [Medline: 26559241]
117. Goldin PR, Lindholm R, Ranta K, Hilgert O, Helteenvuori T, Raevuori A. Feasibility of a therapist-supported, mobile phone-delivered online intervention for depression: longitudinal observational study. JMIR Form Res 2019 Jan 22;3(1):e11509 [FREE Full text] [doi: 10.2196/11509] [Medline: 30682726]
118. Gomà M, Arias-Pujol E, Prims E, Ferrer J, Lara S, Glover V, et al. Internet-based interdisciplinary therapeutic group (Grupo Interdisciplinar Online, GIO) for perinatal anxiety and depression-a randomized pilot study during COVID-19. Arch Womens Ment Health 2024 Jun;27(3):405-415 [FREE Full text] [doi: 10.1007/s00737-023-01412-2] [Medline: 38150150]
119. Gräfe V, Berger T, Hautzinger M, Hohagen F, Lutz W, Meyer B, et al. Health economic evaluation of a web-based intervention for depression: the EVIDENT-trial, a randomized controlled study. Health Econ Rev 2019 Jun 07;9(1):16 [FREE Full text] [doi: 10.1186/s13561-019-0233-y] [Medline: 31175475]
120. Hald GM, Ciprić A, Øverup CS, Štulhofer A, Lange T, Sander S, et al. Randomized controlled trial study of the effects of an online divorce platform on anxiety, depression, and somatization. J Fam Psychol 2020 Sep;34(6):740-751. [doi: 10.1037/fam0000635] [Medline: 32077737]
121. Heller HM, Hoogendoorn AW, Honig A, Broekman BF, van Straten A. The effectiveness of a guided internet-based tool for the treatment of depression and anxiety in pregnancy (MamaKits online): randomized controlled trial. J Med Internet Res 2020 Mar 23;22(3):e15172 [FREE Full text] [doi: 10.2196/15172] [Medline: 32202505]
122. Hirsch A, Luellen J, Holder JM, Steinberg G, Dubiel T, Blazejowskyj A, et al. Managing depressive symptoms in the workplace using a web-based self-care tool: a pilot randomized controlled trial. JMIR Res Protoc 2017 Apr 04;6(4):e51 [FREE Full text] [doi: 10.2196/resprot.7203] [Medline: 28377368]
123. Høifødt RS, Lillevoll KR, Griffiths KM, Wilsgaard T, Eisemann M, Waterloo K, et al. The clinical effectiveness of web-based cognitive behavioral therapy with face-to-face therapist support for depressed primary care patients: randomized controlled trial. J Med Internet Res 2013 Aug 05;15(8):e153 [FREE Full text] [doi: 10.2196/jmir.2714] [Medline: 23916965]
124. Hollinghurst S, Peters TJ, Kaur S, Wiles N, Lewis G, Kessler D. Cost-effectiveness of therapist-delivered online cognitive-behavioural therapy for depression: randomised controlled trial. Br J Psychiatry 2010 Oct;197(4):297-304. [doi: 10.1192/bjp.bp.109.073080] [Medline: 20884953]
125. Hur J, Kim B, Park D, Choi SW. A scenario-based cognitive behavioral therapy mobile app to reduce dysfunctional beliefs in individuals with depression: a randomized controlled trial. Telemed J E Health 2018 Sep;24(9):710-716. [doi: 10.1089/tmj.2017.0214] [Medline: 29323626]
126. Iacoviello BM, Murrough JW, Hoch MM, Huryk KM, Collins KA, Cutter GR, et al. A randomized, controlled pilot trial of the emotional faces memory task: a digital therapeutic for depression. NPJ Digit Med 2018;1:23 [FREE Full text] [doi: 10.1038/s41746-018-0025-5] [Medline: 30854473]
127. İme Y. The effect of online cognitive behavioral group counseling on anxiety, depression, stress and resilience in maraş-centered earthquake survivors. J Ration Emot Cogn Behav Ther 2023 Sep 13;42(2):459-474. [doi: 10.1007/s10942-023-00526-x]
128. Jannati N, Mazhari S, Ahmadian L, Mirzaee M. Effectiveness of an app-based cognitive behavioral therapy program for postpartum depression in primary care: a randomized controlled trial. Int J Med Inform 2020 Sep;141:104145. [doi: 10.1016/j.ijmedinf.2020.104145] [Medline: 32480319]
129. Jonassaint CR, Gibbs P, Belnap BH, Karp JF, Abebe KK, Rollman BL. Engagement and outcomes for a computerised cognitive-behavioural therapy intervention for anxiety and depression in African Americans. BJPsych Open 2017 Jan;3(1):1-5 [FREE Full text] [doi: 10.1192/bjpo.bp.116.003657] [Medline: 28058109]
130. Kenter RM, Cuijpers P, Beekman A, van Straten A. Effectiveness of a web-based guided self-help intervention for outpatients with a depressive disorder: short-term results from a randomized controlled trial. J Med Internet Res 2016 Mar 31;18(3):e80 [FREE Full text] [doi: 10.2196/jmir.4861] [Medline: 27032449]
131. Kim DR, Hantsoo L, Thase ME, Sammel M, Epperson CN. Computer-assisted cognitive behavioral therapy for pregnant women with major depressive disorder. J Womens Health (Larchmt) 2014 Oct;23(10):842-848 [FREE Full text] [doi: 10.1089/jwh.2014.4867] [Medline: 25268672]
132. Kivi M, Eriksson MC, Hange D, Petersson EL, Vernmark K, Johansson B, et al. Internet-based therapy for mild to moderate depression in Swedish primary care: short term results from the PRIM-NET randomized controlled trial. Cogn Behav Ther 2014;43(4):289-298 [FREE Full text] [doi: 10.1080/16506073.2014.921834] [Medline: 24911260]
133. Kladnitski N, Smith J, Allen A, Andrews G, Newby JM. Online mindfulness-enhanced cognitive behavioural therapy for anxiety and depression: outcomes of a pilot trial. Internet Interv 2018 Sep;13:41-50 [FREE Full text] [doi: 10.1016/j.invent.2018.06.003] [Medline: 30206518]
134. Klein JP, Berger T, Schröder J, Späth C, Meyer B, Caspar F, et al. Effects of a psychological internet intervention in the treatment of mild to moderate depressive symptoms: results of the EVIDENT study, a randomized controlled trial. Psychother Psychosom 2016;85(4):218-228 [FREE Full text] [doi: 10.1159/000445355] [Medline: 27230863]

135. Lutz W, Arndt A, Rubel J, Berger T, Schröder J, Späth C, et al. Defining and predicting patterns of early response in a web-based intervention for depression. J Med Internet Res 2017 Jun 09;19(6):e206 [FREE Full text] [doi: 10.2196/jmir.7367] [Medline: 28600278]
136. Krämer LV, Grünzig SD, Baumeister H, Ebert DD, Bengel J. Effectiveness of a guided web-based intervention to reduce depressive symptoms before outpatient psychotherapy: a pragmatic randomized controlled trial. Psychother Psychosom 2021;90(4):233-242. [doi: 10.1159/000515625] [Medline: 33946072]
137. Krusche A, Cyhlarova E, Williams JM. Mindfulness online: an evaluation of the feasibility of a web-based mindfulness course for stress, anxiety and depression. BMJ Open 2013 Nov 29;3(11):e003498 [FREE Full text] [doi: 10.1136/bmjopen-2013-003498] [Medline: 24293203]
138. Lappalainen P, Langrial S, Oinas-Kukkonen H, Tolvanen A, Lappalainen R. Web-based acceptance and commitment therapy for depressive symptoms with minimal support: a randomized controlled trial. Behav Modif 2015 Nov;39(6):805-834. [doi: 10.1177/0145445515598142] [Medline: 26253644]
139. Levesque DA, van Marter DF, Schneider RJ, Bauer MR, Goldberg DN, Prochaska JO, et al. Randomized trial of a computer-tailored intervention for patients with depression. Am J Health Promot 2011;26(2):77-89. [doi: 10.4278/ajhp.090123-QUAN-27] [Medline: 22040388]
140. Liu J, Duan W, Xiao Z, Wu Y. The effectiveness of online group mindfulness-based cognitive therapy for outpatients with depression in China. J Affect Disord 2024 Apr 15;351:387-391 [FREE Full text] [doi: 10.1016/j.jad.2024.01.223] [Medline: 38281594]
141. Löbner M, Pabst A, Stein J, Dorow M, Matschinger H, Luppa M, et al. Computerized cognitive behavior therapy for patients with mild to moderately severe depression in primary care: A pragmatic cluster randomized controlled trial (@ktiv). J Affect Disord 2018 Oct 01;238:317-326. [doi: 10.1016/j.jad.2018.06.008] [Medline: 29902736]
142. Lokman S, Leone SS, Sommers-Spijkerman M, van der Poel A, Smit F, Boon B. Complaint-directed mini-interventions for depressive complaints: a randomized controlled trial of unguided web-based self-help interventions. J Med Internet Res 2017 Jan 04;19(1):e4 [FREE Full text] [doi: 10.2196/jmir.6581] [Medline: 28052840]
143. Lu Y, Li Y, Huang Y, Zhang X, Wang J, Wu L, et al. Effects and mechanisms of a web- and mobile-based acceptance and commitment therapy intervention for anxiety and depression symptoms in nurses: fully decentralized randomized controlled trial. J Med Internet Res 2023 Nov 27;25:e51549 [FREE Full text] [doi: 10.2196/51549] [Medline: 38010787]
144. Lüdtke T, Westermann S, Pult LK, Schneider BC, Pfuhl G, Moritz S. Evaluation of a brief unguided psychological online intervention for depression: a controlled trial including exploratory moderator analyses. Internet Interv 2018 Sep;13:73-81 [FREE Full text] [doi: 10.1016/j.invent.2018.06.004] [Medline: 30206522]
145. MacLean S, Corsi DJ, Litchfield S, Kucharski J, Genise K, Selaman Z, et al. Coach-facilitated web-based therapy compared with information about web-based resources in patients referred to secondary mental health care for depression: randomized controlled trial. J Med Internet Res 2020 Jun 09;22(6):e15001 [FREE Full text] [doi: 10.2196/15001] [Medline: 32515740]
146. Mahoney A, Li I, Haskelberg H, Millard M, Newby JM. The uptake and effectiveness of online cognitive behaviour therapy for symptoms of anxiety and depression during COVID-19. J Affect Disord 2021 Sep 01;292:197-203 [FREE Full text] [doi: 10.1016/j.jad.2021.05.116] [Medline: 34126311]
147. Mahoney A, Shiner CT, Grierson AB, Sharrock MJ, Loughnan SA, Harrison V, et al. Online cognitive behaviour therapy for maternal antenatal and postnatal anxiety and depression in routine care. J Affect Disord 2023 Oct 01;338:121-128. [doi: 10.1016/j.jad.2023.06.008] [Medline: 37295653]
148. Marcelle ET, Nolting L, Hinshaw SP, Aguilera A. Effectiveness of a multimodal digital psychotherapy platform for adult depression: a naturalistic feasibility study. JMIR Mhealth Uhealth 2019 Jan 23;7(1):e10948 [FREE Full text] [doi: 10.2196/10948] [Medline: 30674448]
149. McCloud T, Jones R, Lewis G, Bell V, Tsakanikos E. Effectiveness of a mobile app intervention for anxiety and depression symptoms in university students: randomized controlled trial. JMIR Mhealth Uhealth 2020 Jul 31;8(7):e15418 [FREE Full text] [doi: 10.2196/15418] [Medline: 32735221]
150. McMurchie W, Macleod F, Power K, Laidlaw K, Prentice N. Computerised cognitive behavioural therapy for depression and anxiety with older people: a pilot study to examine patient acceptability and treatment outcome. Int J Geriatr Psychiatry 2013 Nov;28(11):1147-1156. [doi: 10.1002/gps.3935] [Medline: 23495124]
151. Moberg C, Niles A, Beermann D. Guided self-help works: randomized waitlist controlled trial of Pacifica, a mobile app integrating cognitive behavioral therapy and mindfulness for stress, anxiety, and depression. J Med Internet Res 2019 Jun 08;21(6):e12556 [FREE Full text] [doi: 10.2196/12556] [Medline: 31199319]
152. Moghimi E, Stephenson C, Agarwal A, Nikjoo N, Malakouti N, Layzell G, et al. Efficacy of an electronic coganitive behavioral therapy program delivered via the online psychotherapy tool for depression and anxiety related to the COVID-19 pandemic: pre-post pilot study. JMIR Ment Health 2023 Dec 25;10:e51102 [FREE Full text] [doi: 10.2196/51102] [Medline: 37993984]
153. Morthland M, Shah A, Meadows JT, Scogin F. Development of an audio and computer cognitive behavioral therapy for depression in older adults. Aging Ment Health 2020 Aug;24(8):1207-1215. [doi: 10.1080/13607863.2019.1609901] [Medline: 31116017]

154. Moskowitz JT, Addington EL, Shiu E, Bassett SM, Schuette S, Kwok I, et al. Facilitator contact, discussion boards, and virtual badges as adherence enhancements to a web-based, self-guided, positive psychological intervention for depression: randomized controlled trial. J Med Internet Res 2021 Sep 22;23(9):e25922 [FREE Full text] [doi: 10.2196/25922] [Medline: 34550076]
155. Motter JN, Grinberg A, Lieberman DH, Iqnaibi WB, Sneed JR. Computerized cognitive training in young adults with depressive symptoms: effects on mood, cognition, and everyday functioning. J Affect Disord 2019 Feb 15;245:28-37. [doi: 10.1016/j.jad.2018.10.109] [Medline: 30366235]
156. Nelson C, Abraham K, Walters H, Pfeiffer P, Valenstein M. Integration of peer support and computer-based CBT for veterans with depression. Comput Human Behav 2014 Feb;31:57-64. [doi: 10.1016/j.chb.2013.10.012]
157. Oehler C, Görges F, Rogalla M, Rummel-Kluge C, Hegerl U. Efficacy of a guided web-based self-management intervention for depression or dysthymia: randomized controlled trial with a 12-month follow-up using an active control condition. J Med Internet Res 2020 Jul 14;22(7):e15361 [FREE Full text] [doi: 10.2196/15361] [Medline: 32673233]
158. Ofoegbu TO, Asogwa U, Otu MS, Ibenegbu C, Muhammed A, Eze B. Efficacy of guided internet-assisted intervention on depression reduction among educational technology students of Nigerian universities. Medicine (Baltimore) 2020 Feb;99(6):e18774 [FREE Full text] [doi: 10.1097/MD.0000000000018774] [Medline: 32028389]
159. Openshaw D, Pfister R, Silverbaltt H, Moen D. Providing mental health services to women diagnosed with depression in rural Utah communities: using technologically assisted psychotherapeutic intervention as the delivery medium. Rural Ment Health 2011;35(1):23-31. [doi: 10.1037/h0094782]
160. Orr LC, Graham AK, Mohr DC, Greene CJ. Engagement and clinical improvement among older adult primary care patients using a mobile intervention for depression and anxiety: case studies. JMIR Ment Health 2020 Jul 08;7(7):e16341. [doi: 10.2196/16341] [Medline: 32673236]
161. Otared N, Moharrampour NG, Vojoudi B, Najafabadi AJ. A group-based online acceptance and commitment therapy treatment for depression, anxiety symptomns and quality of life in healthcare workers during COVID-19. Int J Psicol Psicol Ther 2021;21(3):399-411 [FREE Full text]
162. Pettitt AK, Nelson BW, Forman-Hoffman VL, Goldin PR, Peiper NC. Longitudinal outcomes of a therapist-supported digital mental health intervention for depression and anxiety symptoms: a retrospective cohort study. Psychol Psychother 2024 Jun;97(2):288-300. [doi: 10.1111/papt.12517] [Medline: 38270220]
163. Pfeiffer PN, Pope B, Houck M, Benn-Burton W, Zivin K, Ganoczy D, et al. Effectiveness of peer-supported computer-based CBT for depression among veterans in primary care. Psychiatr Serv 2020 Mar 01;71(3):256-262. [doi: 10.1176/appi.ps.201900283] [Medline: 31931686]
164. Phillips R, Schneider J, Molosankwe I, Leese M, Foroushani PS, Grime P, et al. Randomized controlled trial of computerized cognitive behavioural therapy for depressive symptoms: effectiveness and costs of a workplace intervention. Psychol Med 2014 Mar;44(4):741-752 [FREE Full text] [doi: 10.1017/S0033291713001323] [Medline: 23795621]
165. Pinto MD, Greenblatt AM, Hickman RL, Rice HM, Thomas TM, Clochesy JM. Assessing the critical parameters of eSMART-MH: a promising avatar-based digital therapeutic intervention to reduce depressive symptoms. Perspect Psychiatr Care 2016 Jul;52(3):157-168. [doi: 10.1111/ppc.12112] [Medline: 25800698]
166. Pinto MD, Hickman RL, Clochesy J, Buchner M. Avatar-based depression self-management technology: promising approach to improve depressive symptoms among young adults. Appl Nurs Res 2013 Feb;26(1):45-48 [FREE Full text] [doi: 10.1016/j.apnr.2012.08.003] [Medline: 23265918]
167. Pots WT, Fledderus M, Meulenbeek PA, ten Klooster PM, Schreurs KM, Bohlmeijer ET. Acceptance and commitment therapy as a web-based intervention for depressive symptoms: randomised controlled trial. Br J Psychiatry 2016 Jan;208(1):69-77. [doi: 10.1192/bjp.bp.114.146068] [Medline: 26250745]
168. Pots WT, Trompetter HR, Schreurs KM, Bohlmeijer ET. How and for whom does web-based acceptance and commitment therapy work? Mediation and moderation analyses of web-based ACT for depressive symptoms. BMC Psychiatry 2016 May 23;16:158 [FREE Full text] [doi: 10.1186/s12888-016-0841-6] [Medline: 27215733]
169. Pratap A, Renn BN, Volponi J, Mooney SD, Gazzaley A, Arean PA, et al. Using mobile apps to assess and treat depression in Hispanic and Latino populations: fully remote randomized clinical trial. J Med Internet Res 2018 Aug 09;20(8):e10130 [FREE Full text] [doi: 10.2196/10130] [Medline: 30093372]
170. Preschl B, Maercker A, Wagner B. The working alliance in a randomized controlled trial comparing online with face-to-face cognitive-behavioral therapy for depression. BMC Psychiatry 2011 Dec 06;11:189 [FREE Full text] [doi: 10.1186/1471-244X-11-189] [Medline: 22145768]
171. Proudfoot J, Clarke J, Birch MR, Whitton AE, Parker G, Manicavasagar V, et al. Impact of a mobile phone and web program on symptom and functional outcomes for people with mild-to-moderate depression, anxiety and stress: a randomised controlled trial. BMC Psychiatry 2013 Nov 18;13:312 [FREE Full text] [doi: 10.1186/1471-244X-13-312] [Medline: 24237617]
172. Proyer RT, Gander F, Wellenzohn S, Ruch W. Positive psychology interventions in people aged 50-79 years: long-term effects of placebo-controlled online interventions on well-being and depression. Aging Ment Health 2014;18(8):997-1005 [FREE Full text] [doi: 10.1080/13607863.2014.899978] [Medline: 24712501]

173. Pugh NE, Hadjistavropoulos HD, Fuchs CM. Internet therapy for postpartum depression: a case illustration of emailed therapeutic assistance. Arch Womens Ment Health 2014 Aug;17(4):327-337. [doi: 10.1007/s00737-014-0439-2] [Medline: 24965602]
174. Reins JA, Boß L, Lehr D, Berking M, Ebert DD. The more I got, the less I need? Efficacy of internet-based guided self-help compared to online psychoeducation for major depressive disorder. J Affect Disord 2019 Mar 01;246:695-705. [doi: 10.1016/j.jad.2018.12.065] [Medline: 30611913]
175. Richards D, Enrique A, Eilert N, Franklin M, Palacios J, Duffy D, et al. A pragmatic randomized waitlist-controlled effectiveness and cost-effectiveness trial of digital interventions for depression and anxiety. NPJ Digit Med 2020;3:85 [FREE Full text] [doi: 10.1038/s41746-020-0293-8] [Medline: 32566763]
176. Richards D, Timulak L, Hevey D. A comparison of two online cognitive‐behavioural interventions for symptoms of depression in a student population: the role of therapist responsiveness. Couns Psychother Res 2013 Aug 16;13(3):184-193. [doi: 10.1080/14733145.2012.733715]
177. Richter LM, Machleit-Ebner A, Scherbaum N, Bonnet U. How effective is a web-based mental health intervention (Deprexis) in the treatment of moderate and major depressive disorders when started during routine psychiatric inpatient treatment as an adjunct therapy? A pragmatic parallel-group randomized controlled trial. Fortschr Neurol Psychiatr 2023 Jul;91(7-08):297-310. [doi: 10.1055/a-1826-2888] [Medline: 35882364]
178. Ritvo P, Knyahnytska Y, Pirbaglou M, Wang W, Tomlinson G, Zhao H, et al. Online mindfulness-based cognitive behavioral therapy intervention for youth with major depressive disorders: randomized controlled trial. J Med Internet Res 2021 Mar 10;23(3):e24380 [FREE Full text] [doi: 10.2196/24380] [Medline: 33688840]
179. Rotondi AJ, Belnap BH, Rothenberger S, Feldman R, Hanusa B, Rollman BL. Predictors of use and drop out from a web-based cognitive behavioral therapy program and health community for depression and anxiety in primary care patients: secondary analysis of a randomized controlled trial. JMIR Ment Health 2024 Jan 17;11:e52197 [FREE Full text] [doi: 10.2196/52197] [Medline: 38231552]
180. Sampson E. Implementing digital cognitive-behavioral therapy for major depressive disorder in routine psychiatric appointments: a pilot project in a rural population. J Psychosoc Nurs Ment Health Serv 2023 Oct;61(10):44-51. [doi: 10.3928/02793695-20230424-01] [Medline: 37134285]
181. Sandoval LR, Buckey JC, Ainslie R, Tombari M, Stone W, Hegel MT. Randomized controlled trial of a computerized interactive media-based problem solving treatment for depression. Behav Ther 2017 May;48(3):413-425 [FREE Full text] [doi: 10.1016/j.beth.2016.04.001] [Medline: 28390503]
182. Schneider BC, Schröder J, Berger T, Hohagen F, Meyer B, Späth C, et al. Corrigendum to ``bridging the "digital divide": a comparison of use and effectiveness of an online intervention for depression between baby boomers and millennials". J Affect Disord 2018 Dec 01;241:635. [doi: 10.1016/j.jad.2018.05.031] [Medline: 29866375]
183. Schueller SM, Mohr DC. Initial field trial of a coach-supported web-based depression treatment. Int Conf Pervasive Comput Technol Healthc 2015 Aug;2015:27 [FREE Full text] [doi: 10.4108/icst.pervasivehealth.2015.260115] [Medline: 26640741]
184. Schuster R, Kalthoff I, Walther A, Köhldorfer L, Partinger E, Berger T, et al. Effects, adherence, and therapists' perceptions of web- and mobile-supported group therapy for depression: mixed-methods study. J Med Internet Res 2019 Apr 28;21(5):e11860 [FREE Full text] [doi: 10.2196/11860] [Medline: 31066700]
185. Schuster R, Sigl S, Berger T, Laireiter AR. Patients' experiences of web- and mobile-assisted group therapy for depression and implications of the group setting: qualitative follow-up study. JMIR Ment Health 2018 Jul 11;5(3):e49 [FREE Full text] [doi: 10.2196/mental.9613] [Medline: 29997106]
186. Segal ZV, Dimidjian S, Beck A, Boggs JM, Vanderkruik R, Metcalf CA, et al. Outcomes of online mindfulness-based cognitive therapy for patients with residual depressive symptoms: a randomized clinical trial. JAMA Psychiatry 2020 Jun 01;77(6):563-573 [FREE Full text] [doi: 10.1001/jamapsychiatry.2019.4693] [Medline: 31995132]
187. Sethi S. Treating youth depression and anxiety: a randomised controlled trial examining the efficacy of computerised versus face‐to‐face cognitive behaviour therapy. Aust Psychol 2020 Nov 12;48(4):249-257. [doi: 10.1111/ap.12006]
188. Shah A, Morthland M, Scogin F, Presnell A, DiNapoli EA, DeCoster J, et al. Audio and computer cognitive behavioral therapy for depressive symptoms in older adults: a pilot randomized controlled trial. Behav Ther 2018 Nov;49(6):904-916. [doi: 10.1016/j.beth.2018.06.002] [Medline: 30316489]
189. Sharry J, Davidson R, McLoughlin O, Doherty G. A service-based evaluation of a therapist-supported online cognitive behavioral therapy program for depression. J Med Internet Res 2013 Jun 27;15(6):e121 [FREE Full text] [doi: 10.2196/jmir.2248] [Medline: 23807565]
190. Silva Almodovar A, Surve S, Axon DR, Cooper D, Nahata MC. Self-directed engagement with a mobile app (Sinasprite) and its effects on confidence in coping skills, depression, and anxiety: retrospective longitudinal study. JMIR Mhealth Uhealth 2018 Mar 16;6(3):e64 [FREE Full text] [doi: 10.2196/mhealth.9612] [Medline: 29549066]
191. Silverstone PH, Rittenbach K, Suen VY, Moretzsohn A, Cribben I, Bercov M, et al. Depression outcomes in adults attending family practice were not improved by screening, stepped-care, or online CBT during a 12-week study when compared to controls in a randomized trial. Front Psychiatry 2017;8:32 [FREE Full text] [doi: 10.3389/fpsyt.2017.00032] [Medline: 28373846]

192. Spates CR, Kalata AH, Ozeki S, Stanton CE, Peters S. Initial open trial of a computerized behavioral activation treatment for depression. Behav Modif 2013 May;37(3):259-297. [doi: 10.1177/0145445512455051] [Medline: 22987916]
193. Stearns-Yoder KA, Ryan AT, Smith AA, Forster JE, Barnes SM, Brenner LA. Computerized cognitive behavioral therapy intervention for depression among veterans: acceptability and feasibility study. JMIR Form Res 2022 Apr 25;6(4):e31835 [FREE Full text] [doi: 10.2196/31835] [Medline: 35468088]
194. Thase ME, McCrone P, Barrett MS, Eells TD, Wisniewski SR, Balasubramani GK, et al. Improving cost-effectiveness and access to cognitive behavior therapy for depression: providing remote-ready, computer-assisted psychotherapy in times of crisis and beyond. Psychother Psychosom 2020;89(5):307-313 [FREE Full text] [doi: 10.1159/000508143] [Medline: 32396917]
195. Thase ME, Wright JH, Eells TD, Barrett MS, Wisniewski SR, Balasubramani GK, et al. Improving the efficiency of psychotherapy for depression: computer-assisted versus standard CBT. Am J Psychiatry 2018 Mar 01;175(3):242-250 [FREE Full text] [doi: 10.1176/appi.ajp.2017.17010089] [Medline: 28969439]
196. Titov N, Dear BF, Staples LG, Bennett-Levy J, Klein B, Rapee RM, et al. MindSpot clinic: an accessible, efficient, and effective online treatment service for anxiety and depression. Psychiatr Serv 2015 Oct;66(10):1043-1050. [doi: 10.1176/appi.ps.201400477] [Medline: 26130001]
197. Titzler I, Egle V, Berking M, Gumbmann C, Ebert DD. Blended psychotherapy: treatment concept and case report for the integration of internet- and mobile-based interventions into brief psychotherapy of depressive disorders. Verhaltenstherapie 2022;32(1):230-244 [FREE Full text] [doi: 10.6084/m9.figshare.10505021]
198. Tulbure BT, Rusu A, Sava FA, Sălăgean N, Farchione TJ. A web-based transdiagnostic intervention for affective and mood disorders: randomized controlled trial. JMIR Ment Health 2018 May 24;5(2):e36 [FREE Full text] [doi: 10.2196/mental.8901] [Medline: 29798831]
199. van der Zanden R, Curie K, van Londen M, Kramer J, Steen G, Cuijpers P. Web-based depression treatment: associations of clients' word use with adherence and outcome. J Affect Disord 2014 May;160:10-13 [FREE Full text] [doi: 10.1016/j.jad.2014.01.005] [Medline: 24709016]
200. van der Zanden R, Galindo-Garre F, Curie K, Kramer J, Cuijpers P. Online cognitive-based intervention for depression: exploring possible circularity in mechanisms of change. Psychol Med 2014 Apr;44(6):1159-1170. [doi: 10.1017/S003329171300175X] [Medline: 23866103]
201. van der Zanden R, Kramer J, Gerrits R, Cuijpers P. Effectiveness of an online group course for depression in adolescents and young adults: a randomized trial. J Med Internet Res 2012 Jun 07;14(3):e86 [FREE Full text] [doi: 10.2196/jmir.2033] [Medline: 22677437]
202. Venkatesan A, Rahimi L, Kaur M, Mosunic C. Digital cognitive behavior therapy intervention for depression and anxiety: retrospective study. JMIR Ment Health 2020 Aug 26;7(8):e21304 [FREE Full text] [doi: 10.2196/21304] [Medline: 32845246]
203. Wagner B, Horn AB, Maercker A. Internet-based versus face-to-face cognitive-behavioral intervention for depression: a randomized controlled non-inferiority trial. J Affect Disord 2014 Jan;152-154:113-121. [doi: 10.1016/j.jad.2013.06.032] [Medline: 23886401]
204. Wahle F, Kowatsch T, Fleisch E, Rufer M, Weidt S. Mobile sensing and support for people with depression: a pilot trial in the wild. JMIR Mhealth Uhealth 2016 Sep 21;4(3):e111 [FREE Full text] [doi: 10.2196/mhealth.5960] [Medline: 27655245]
205. Wang L, Miller L. Assessment and disruption of ruminative episodes to enhance mobile cognitive behavioral therapy just-in-time adaptive interventions in clinical depression: pilot randomized controlled trial. JMIR Form Res 2023 Jan 05;7:e37270 [FREE Full text] [doi: 10.2196/37270] [Medline: 36602841]
206. Warmerdam L, van Straten A, Jongsma J, Twisk J, Cuijpers P. Online cognitive behavioral therapy and problem-solving therapy for depressive symptoms: exploring mechanisms of change. J Behav Ther Exp Psychiatry 2010 Mar;41(1):64-70. [doi: 10.1016/j.jbtep.2009.10.003] [Medline: 19913781]
207. Watts S, Mackenzie A, Thomas C, Griskaitis A, Mewton L, Williams A, et al. CBT for depression: a pilot RCT comparing mobile phone vs. computer. BMC Psychiatry 2013 Feb 07;13:49 [FREE Full text] [doi: 10.1186/1471-244X-13-49] [Medline: 23391304]
208. Welch ES, Weigand A, Hooker JE, Philip NS, Tyrka AR, Press DZ, et al. Feasibility of computerized cognitive-behavioral therapy combined with bifrontal transcranial direct current stimulation for treatment of major depression. Neuromodulation 2019 Dec;22(8):898-903. [doi: 10.1111/ner.12807] [Medline: 30153360]
209. Whiteside U, Richards J, Steinfeld B, Simon G, Caka S, Tachibana C, et al. Online cognitive behavioral therapy for depressed primary care patients: a pilot feasibility project. Perm J 2014;18(2):21-27 [FREE Full text] [doi: 10.7812/TPP/13-155] [Medline: 24867546]
210. Whitton AE, Proudfoot J, Clarke J, Birch MR, Parker G, Manicavasagar V, et al. Breaking open the black box: isolating the most potent features of a web and mobile phone-based intervention for depression, anxiety, and stress. JMIR Ment Health 2015;2(1):e3 [FREE Full text] [doi: 10.2196/mental.3573] [Medline: 26543909]
211. Wijnen BF, Lokman S, Leone S, Evers SM, Smit F. Complaint-directed mini-interventions for depressive symptoms: a health economic evaluation of unguided web-based self-help interventions based on a randomized controlled trial. J Med Internet Res 2018 Oct 01;20(10):e10455 [FREE Full text] [doi: 10.2196/10455] [Medline: 30274958]

212. Williams C, McClay CA, Martinez R, Morrison J, Haig C, Jones R, et al. Online cognitive behavioral therapy (CBT) life skills program for depression: pilot randomized controlled trial. JMIR Form Res 2022 Feb 17;6(2):e30489 [FREE Full text] [doi: 10.2196/30489] [Medline: 35175203]
213. Wright JH, Owen J, Eells TD, Antle B, Bishop LB, Girdler R, et al. Effect of computer-assisted cognitive behavior therapy vs usual care on depression among adults in primary care: a randomized clinical trial. JAMA Netw Open 2022 Feb 01;5(2):e2146716 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.46716] [Medline: 35142833]
214. Wu MS, Wickham RE, Chen SY, Chen C, Lungu A. Examining the impact of digital components across different phases of treatment in a blended care cognitive behavioral therapy intervention for depression and anxiety: pragmatic retrospective study. JMIR Form Res 2021 Dec 17;5(12):e33452 [FREE Full text] [doi: 10.2196/33452] [Medline: 34927591]
215. Xiang X, Kayser J, Ash S, Zheng C, Sun Y, Weaver A, et al. Web-based cognitive behavioral therapy for depression among homebound older adults: development and usability study. JMIR Aging 2023 Sep 19;6:e47691 [FREE Full text] [doi: 10.2196/47691] [Medline: 37725423]
216. Yeung A, Wang F, Feng F, Zhang J, Cooper A, Hong L, et al. Outcomes of an online computerized cognitive behavioral treatment program for treating chinese patients with depression: a pilot study. Asian J Psychiatr 2018 Dec;38:102-107. [doi: 10.1016/j.ajp.2017.11.007] [Medline: 29146042]
217. Zhang R, Nicholas J, Knapp AA, Graham AK, Gray E, Kwasny MJ, et al. Clinically meaningful use of mental health apps and its effects on depression: mixed methods study. J Med Internet Res 2019 Dec 20;21(12):e15644 [FREE Full text] [doi: 10.2196/15644] [Medline: 31859682]
218. Bisby MA, Balakumar T, Scott AJ, Titov N, Dear BF. An online therapist-guided ultra-brief treatment for depression and anxiety: a randomized controlled trial. Psychol Med 2024 Apr;54(5):902-913. [doi: 10.1017/S003329172300260X] [Medline: 37655527]
219. Amer NA, Shohieb SM, Eladrosy WM, Elbakry HM, Elrazek SM. Sokoon: a gamification-based cognitive behavioral therapy application – an application for depression, stress, and anxiety. Int J Gaming Comput Mediat Simul 2023;15(1):1-26 [FREE Full text] [doi: 10.4018/IJGCMS.324098] [Medline: 10.6084/m9.figshare.10505021]
220. Mohr DC, Schueller SM, Tomasino KN, Kaiser SM, Alam N, Karr C, et al. Comparison of the effects of coaching and receipt of app recommendations on depression, anxiety, and engagement in the IntelliCare platform: factorial randomized controlled trial. J Med Internet Res 2019 Aug 28;21(8):e13609 [FREE Full text] [doi: 10.2196/13609] [Medline: 31464192]
221. Broglia E, Millings A, Barkham M. Counseling with guided use of a mobile well-being app for students experiencing anxiety or depression: clinical outcomes of a feasibility trial embedded in a student counseling service. JMIR Mhealth Uhealth 2019 Aug 15;7(8):e14318 [FREE Full text] [doi: 10.2196/14318] [Medline: 31418424]
222. Boggs JM, Beck A, Felder JN, Dimidjian S, Metcalf CA, Segal ZV. Web-based intervention in mindfulness meditation for reducing residual depressive symptoms and relapse prophylaxis: a qualitative study. J Med Internet Res 2014 Mar 24;16(3):e87 [FREE Full text] [doi: 10.2196/jmir.3129] [Medline: 24662625]
223. Al-Alawi M, McCall RK, Sultan A, Al Balushi N, Al-Mahrouqi T, Al Ghailani A, et al. Efficacy of a six-week-long therapist-guided online therapy versus self-help internet-based therapy for COVID-19-induced anxiety and depression: open-label, pragmatic, randomized controlled trial. JMIR Ment Health 2021 Feb 12;8(2):e26683 [FREE Full text] [doi: 10.2196/26683] [Medline: 33512323]
224. Levin W, Campbell DR, McGovern KB, Gau JM, Kosty DB, Seeley JR, et al. A computer-assisted depression intervention in primary care. Psychol Med 2010 Oct 20;41(7):1373-1383. [doi: 10.1017/s0033291710001935]
225. Ferrari AJ, Charlson FJ, Norman RE, Patten SB, Freedman G, Murray CJ, et al. Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. PLoS Med 2013 Nov;10(11):e1001547 [FREE Full text] [doi: 10.1371/journal.pmed.1001547] [Medline: 24223526]
226. Kuehner C. Why is depression more common among women than among men? Lancet Psychiatry 2017 Feb;4(2):146-158 [FREE Full text] [doi: 10.1016/s2215-0366(16)30263-2]
227. Lim GY, Tam WW, Lu Y, Ho CS, Zhang MW, Ho RC. Prevalence of depression in the community from 30 countries between 1994 and 2014. Sci Rep 2018 Feb 12;8(1):2861 [FREE Full text] [doi: 10.1038/s41598-018-21243-x] [Medline: 29434331]
228. Women's mental health evidence based review. World Health Organization. 2000. URL: https://www.who.int/publications/i/item/WHO-MSD-MDP-00.1 [accessed 2023-06-22]
229. Yu S. Uncovering the hidden impacts of inequality on mental health: a global study. Transl Psychiatry 2018 May 18;8(1):98 [FREE Full text] [doi: 10.1038/s41398-018-0148-0] [Medline: 29777100]
230. Rungreangkulkij S, Rujiraprasert N. Gender inequality identified as an underlying cause of depression in Thai women. J Int Women Stud 2019;20(7):395 [FREE Full text]
231. Rafful C, Medina-Mora ME, Borges G, Benjet C, Orozco R. Depression, gender, and the treatment gap in Mexico. J Affect Disord 2012 Apr;138(1-2):165-169. [doi: 10.1016/j.jad.2011.12.040] [Medline: 22341482]
232. Krumm S, Checchia C, Koesters M, Kilian R, Becker T. Men's views on depression: a systematic review and metasynthesis of qualitative research. Psychopathology 2017 Mar 11;50(2):107-124. [doi: 10.1159/000455256] [Medline: 28285304]
233. Seidler ZE, Dawes AJ, Rice SM, Oliffe JL, Dhillon HM. The role of masculinity in men's help-seeking for depression: a systematic review. Clin Psychol Rev 2016 Nov;49:106-118. [doi: 10.1016/j.cpr.2016.09.002] [Medline: 27664823]

234. Otten D, Tibubos AN, Schomerus G, Brähler E, Binder H, Kruse J, et al. Similarities and differences of mental health in women and men: a systematic review of findings in three large German cohorts. Front Public Health 2021 Feb 5;9:553071 [FREE Full text] [doi: 10.3389/fpubh.2021.553071] [Medline: 33614574]
235. Winkler D, Pjrek E, Kasper S. Anger attacks in depression--evidence for a male depressive syndrome. Psychother Psychosom 2005 Aug 10;74(5):303-307. [doi: 10.1159/000086321] [Medline: 16088268]
236. Beutel ME, Brähler E, Wiltink J, Kerahrodi JG, Burghardt J, Michal M, et al. New onset of depression in aging women and men: contributions of social, psychological, behavioral, and somatic predictors in the community. Psychol Med 2018 Aug 22;49(07):1148-1155. [doi: 10.1017/s0033291718001848]
237. Perna L, Mielck A, Lacruz ME, Emeny RT, Holle R, Breitfelder A, et al. Socioeconomic position, resilience, and health behaviour among elderly people. Int J Public Health 2012 Apr 13;57(2):341-349. [doi: 10.1007/s00038-011-0294-0] [Medline: 21912944]
238. Angst J, Gamma A, Gastpar M, Lépine JP, Mendlewicz J, Tylee A, et al. Gender differences in depression. epidemiological findings from the European DEPRES I and II studies. Eur Arch Psychiatry Clin Neurosci 2002 Oct;252(5):201-209. [doi: 10.1007/s00406-002-0381-6] [Medline: 12451460]
239. Häfner S, Emeny R, Lacruz M, Baumert J, Herder C, Koenig W, et al. Association between social isolation and inflammatory markers in depressed and non-depressed individuals: results from the MONICA/KORA study. Brain Behav Immun 2011 Nov;25(8):1701-1707. [doi: 10.1016/j.bbi.2011.06.017] [Medline: 21756997]
240. Smedley A, Smedley BD. Race as biology is fiction, racism as a social problem is real: anthropological and historical perspectives on the social construction of race. Am Psychol 2005;60(1):16-26. [doi: 10.1037/0003-066x.60.1.16]
241. Husain A, Hodge DR. Islamically modified cognitive behavioral therapy: enhancing outcomes by increasing the cultural congruence of cognitive behavioral therapy self-statements. Int Soc Work 2016 Apr 30;59(3):393-405. [doi: 10.1177/0020872816629193]
242. Bernal G, Jiménez-Chafey MI, Domenech Rodríguez MM. Cultural adaptation of treatments: a resource for considering culture in evidence-based practice. Prof Psychol Res Pr 2009 Aug;40(4):361-368. [doi: 10.1037/a0016401]
243. Ratcliff KS. The Social Determinants of Health: Looking Upstream. New York, NY: Wiley-Blackwell; 2017.
244. Carmona-Huerta J, Durand-Arias S, Rodriguez A, Guarner-Catalá C, Cardona-Muller D, Madrigal-de-León E, et al. Community mental health care in Mexico: a regional perspective from a mid-income country. Int J Ment Health Syst 2021 Jan 11;15(1):7. [doi: 10.1186/s13033-020-00429-9]
245. Wainberg ML, Scorza P, Shultz JM, Helpman L, Mootz JJ, Johnson KA, et al. Challenges and opportunities in global mental health: a research-to-practice perspective. Curr Psychiatry Rep 2017 May;19(5):28 [FREE Full text] [doi: 10.1007/s11920-017-0780-z] [Medline: 28425023]
246. Kiekens WJ, la Roi C, Dijkstra JK. Sexual identity disparities in mental health among U.K. adults, U.S. adults, and U.S. adolescents: examining heterogeneity by race/ethnicity. Psychol Sex Orientat Gend Divers 2021 Dec;8(4):407-419. [doi: 10.1037/sgd0000432]
247. Lee H, Singh GK. Monthly trends in self-reported health status and depression by race/ethnicity and socioeconomic status during the COVID-19 pandemic, United States, April 2020 - May 2021. Ann Epidemiol 2021 Nov;63:52-62 [FREE Full text] [doi: 10.1016/j.annepidem.2021.07.014] [Medline: 34358622]
248. Weller BE, Blanford KL, Butler AM. Estimated prevalence of psychiatric comorbidities in U.S. adolescents with depression by race/ethnicity, 2011-2012. J Adolesc Health 2018 Jun;62(6):716-721. [doi: 10.1016/j.jadohealth.2017.12.020] [Medline: 29784115]
249. Le HN, Boyd RC, Lara MA. Treatment of depressive disorders and comorbidity in ethnic minority groups. In: Richards CS, O'Hara MW, editors. The Oxford Handbook of Depression and Comorbidity. Oxford, UK: Oxford University Press; 2014:528-545.
250. Wentzel J, van der Vaart R, Bohlmeijer ET, van Gemert-Pijnen JE. Mixing online and face-to-face therapy: how to benefit from blended care in mental health care. JMIR Ment Health 2016 Feb 09;3(1):e9 [FREE Full text] [doi: 10.2196/mental.4534] [Medline: 26860537]
251. Erekson DM, Lambert MJ, Eggett DL. The relationship between session frequency and psychotherapy outcome in a naturalistic setting. J Consult Clin Psychol 2015 Dec;83(6):1097-1107. [doi: 10.1037/a0039774] [Medline: 26436645]
252. Omylinska‐Thurston J, Karkou V, Parsons A, Nair K, Dubrow‐Marshall L, Starkey J, et al. Arts for the Blues: the development of a new evidence‐based creative group psychotherapy for depression. Couns Psychother Res 2020 Dec 14;21(3):597-607. [doi: 10.1002/capr.12373]
253. Martin C, Iqbal Z, Airey ND, Marks L. Improving access to psychological therapies (IAPT) has potential but is not sufficient: how can it better meet the range of primary care mental health needs? Br J Clin Psychol 2022 Mar 14;61(1):157-174. [doi: 10.1111/bjc.12314] [Medline: 34124792]
254. Omylinska‐Thurston J, McMeekin A, Walton P, Proctor G. Clients' perceptions of unhelpful factors in CBT in IAPT serving a deprived area of the UK. Couns Psychother Res 2019 Jul 22;19(4):455-464. [doi: 10.1002/capr.12249]
255. Farish-Edwards F, Parsons AS, Starkey J, Dubrow-Marshall L, Thurston SD, Omylinska-Thurston J, et al. Digitising creative psychological therapy: arts for the blues (a4b). In: Prescott J, editor. Digital Innovations for Mental Health Support. New York, NY: IGI Global; 2022.

256. Eisenstadt M, Liverpool S, Infanti E, Ciuvat RM, Carlsson C. Mobile apps that promote emotion regulation, positive mental health, and well-being in the general population: systematic review and meta-analysis. JMIR Ment Health 2021 Nov 08;8(11):e31170 [FREE Full text] [doi: 10.2196/31170] [Medline: 34747713]
257. Fordham B, Sugavanam T, Edwards K, Stallard P, Howard R, das Nair R, et al. The evidence for cognitive behavioural therapy in any condition, population or context: a meta-review of systematic reviews and panoramic meta-analysis. Psychol Med 2021 Jan 18;51(1):21-29. [doi: 10.1017/s0033291720005292]
258. Marshall JM, Dunstan DA, Bartik W. Smartphone psychological therapy during COVID-19: a study on the effectiveness of five popular mental health apps for anxiety and depression. Front Psychol 2021 Dec 13;12:775775 [FREE Full text] [doi: 10.3389/fpsyg.2021.775775] [Medline: 35069357]

Abbreviations

CBT: cognitive behavioral therapy
MMAT: Mixed Methods Appraisal Tool
NICE: National Institute for Health and Care Excellence
PICOS: Population, Intervention, Comparison, Outcomes and Study Design
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SMD: standardized mean difference
TIDieR: Template for Intervention Description and Replication

Edited by J Torous; submitted 14.12.23; peer-reviewed by H Shah, L Yu, D Singh; comments to author 31.12.23; revised version received 24.05.24; accepted 07.06.24; published 30.09.24.

Please cite as:
Omylinska-Thurston J, Aithal S, Liverpool S, Clark R, Moula Z, Wood J, Viliardos L, Rodríguez-Dorans E, Farish-Edwards F, Parsons A, Eisenstadt M, Bull M, Dubrow-Marshall L, Thurston S, Karkou V
Digital Psychotherapies for Adults Experiencing Depressive Symptoms: Systematic Review and Meta-Analysis
JMIR Ment Health 2024;11:e55500
URL: https://mental.jmir.org/2024/1/e55500
doi:10.2196/55500
PMID:39348177

Review

Leveraging Personal Technologies in the Treatment of Schizophrenia Spectrum Disorders: Scoping Review

Jessica D'Arcey[1,2], MA, MSc; John Torous[3], MBI, MD; Toni-Rose Asuncion[1,4], BSc, MA; Leah Tackaberry-Giddens[1], BSc; Aqsa Zahid[1,2], BSc; Mira Ishak[1]; George Foussias[1,2], MD, PhD; Sean Kidd[1,2], CPsych, PhD

[1]Schizophrenia Division, Centre for Addiction and Mental Health, Toronto, ON, Canada

[2]Clinical Psychological Sciences, Department of Psychology, University of Toronto Scarborough, Toronto, ON, Canada

[3]Department of Psychiatry, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, United States

[4]Department of Psychology, Research and Clinical Training, Concordia University, Montreal, QC, Canada

Corresponding Author:
Jessica D'Arcey, MA, MSc
Schizophrenia Division
Centre for Addiction and Mental Health
1025 Queen Street West
Toronto, ON, M6J 1H1
Canada
Phone: 1 647 385 2490
Email: jessica.darcey@mail.utoronto.ca

Abstract

Background: Digital mental health is a rapidly growing field with an increasing evidence base due to its potential scalability and impacts on access to mental health care. Further, within underfunded service systems, leveraging personal technologies to deliver or support specialized service delivery has garnered attention as a feasible and cost-effective means of improving access. Digital health relevance has also improved as technology ownership in individuals with schizophrenia has improved and is comparable to that of the general population. However, less digital health research has been conducted in groups with schizophrenia spectrum disorders compared to other mental health conditions, and overall feasibility, efficacy, and clinical integration remain largely unknown.

Objective: This review aims to describe the available literature investigating the use of personal technologies (ie, phone, computer, tablet, and wearables) to deliver or support specialized care for schizophrenia and examine opportunities and barriers to integrating this technology into care.

Methods: Given the size of this review, we used scoping review methods. We searched 3 major databases with search teams related to schizophrenia spectrum disorders, various personal technologies, and intervention outcomes related to recovery. We included studies from the full spectrum of methodologies, from development papers to implementation trials. Methods and reporting follow the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines.

Results: This search resulted in 999 studies, which, through review by at least 2 reviewers, included 92 publications. Included studies were published from 2010 to 2023. Most studies examined multitechnology interventions (40/92, 43%) or smartphone apps (25/92, 27%), followed by SMS text messaging (16/92, 17%) and internet-based interventions (11/92, 12%). No studies used wearable technology on its own to deliver an intervention. Regarding the stage of research in the field, the largest number of publications were pilot studies (32/92, 35%), followed by randomized control trials (RCTs; 20/92, 22%), secondary analyses (16/92, 17%), RCT protocols (16/92, 17%), development papers (5/92, 5%), and nonrandomized or quasi-experimental trials (3/92, 3%). Most studies did not report on safety indices (55/92, 60%) or privacy precautions (64/92, 70%). Included studies tend to report consistent positive user feedback regarding the usability, acceptability, and satisfaction with technology; however, engagement metrics are highly variable and report mixed outcomes. Furthermore, efficacy at both the pilot and RCT levels report mixed findings on primary outcomes.

Conclusions: Overall, the findings of this review highlight the discrepancy between the high levels of acceptability and usability of these digital interventions, mixed efficacy results, and difficulties with sustained engagement. The discussion highlights common patterns that may underscore this observation in the field; however, as this was a scoping review, a more in-depth systematic review or meta-analysis may be required to better understand the trends outlined in this review.

(*JMIR Ment Health 2024;11:e57150*) doi:10.2196/57150

KEYWORDS

schizophrenia; digital mental health; personal technology; access to specialized resources; mental health; scoping review; mental health care; feasibility; efficacy; clinical integration; support; specialized care; care; database; schizophrenia spectrum disorder; text messaging; text; user feedback; usability; acceptability; satisfaction; engagement; digital health; digital mental health; technology; health technology; mood disorder; mood disorders; neurodevelopment; eHealth; mobile phone

Introduction

Background

Research on digital mental health care in groups with serious mental illness (SMI), including schizophrenia spectrum disorders (SSDs), has grown slowly compared to other mental health conditions such as anxiety and depression [1,2]. Historically, this has been due to concerns about the ability to afford and use technology (ie, device ownership, access to the internet, and cellular data plans), understand the limitations of the technology (ie, privacy, crisis planning, and digital literacy), and make effective gains from digitally delivered content.

Recent research shows that cell phone and smartphone ownership and internet access are high in populations with SMI with and without psychosis [3,4] and have shown to be an adequate means of service delivery over the COVID-19 pandemic. Leveraging personal technologies (ie, laptops, tablets, smartphones, and wearable technologies) has also been framed as a critical measure to address access to specialized mental health care in remote areas and populations identified as marginalized, as access to smartphones and the internet is improving in these communities [5-7]. In addition, it has the potential to address barriers to stigma, time, and cost that limit the frequency and duration of available mental health care [8,9]. Given the potential to overcome such significant barriers, experts predict the use of personal technologies to deliver or assist with mental health care to continue to grow and hope to capitalize on the burst of enthusiasm brought about by the COVID-19 pandemic [10-12].

Although technologically delivered and assisted care is steadily gaining credibility and acceptability among mental health care professionals, there remain questions about the efficacy, safety and potential limitations of technologically delivered care in SSDs [13]. In fact, research has found that out of the publicly available technology advertised for mental health care, only 3% have scientific evidence supporting their use [14]. In schizophrenia and psychosis, a commercial review found that only 6 publicly available apps out of 700 were supported by scientific evidence, and many contained stigmatizing themes and misinformation [15].

To this end, it is critical to identify evidence-based digital mental health tools available for use on personal technologies by individuals with SSDs and evaluate the state of research in the field and readiness for implementation. Previous reviews in this area have tended to focus on 1 technical modality (eg, mobile apps [16] and SMS text messaging [17]) or 1 aspect of recovery (eg, relapse prevention [18]) but have yet to examine the field collectively across technological platforms, treatment targets, and therapeutic approaches to provide a high-level snapshot of the extant literature.

This Study

This study is a scoping review examining the use of personal technologies to deliver digital mental health interventions in the treatment of SSDs. The overarching aim was to examine the state of the existing evidence supporting digital health tools for individuals with SSDs. Specifically, we identified gaps in the literature and explored the next steps for future research as the field moves toward implementation. To this end, we examined studies across the research stage, from development papers to implementation studies, using various types of commonly owned personal technologies that may be used to support clinical outcomes and recovery in populations experiencing SSDs. Personal technologies were chosen as the target technology for this review, as it can be argued that this may be the most efficient way to increase access to specialized information and care for individuals who cannot access traditional care consistently. It is also plausible that personal technologies may be the most accessible and cost-effective technological aid to traditional care.

Methods

Overview

Given the complex and broad nature of our research question, our results are presented in alignment with the scoping review methodology, focusing on the state of research and identification of key concepts and gaps [19,20]. Our methods are structured according to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) principles endorsed by the Cochrane Collaboration [21] and the scoping review reporting framework [22].

Inclusion Criteria

The included studies were published in English, examined the use of digital health tools in populations with SSDs, and published between 2010 and 2023 (given the evolution of technology during this period). Population, Intervention, Comparison, and Outcome principles guided the formulation of the inclusion criteria using the categories population, intervention, and outcome of interest (control type was not applicable as we included a range of study types; refer to Textbox 1 for details). Studies were excluded if no full text was available (eg, editorials and conference presentations), if not published in English, if examined in a population with health challenges other than SSDs (eg, depression, substance use, and physical health conditions), if they did not use technology to deliver an intervention, or if they did not examine clinical targets related to SSDs. We also did not include studies that examined the pivot to virtual care (eg, phone, video, or other communication technology-based care) during the COVID-19 pandemic and virtually delivered care more broadly, as the only technological component was videoconferencing or telephone delivery of treatment as usual (TAU). The included studies sought to enhance the delivery of TAU (ie, technologically supported TAU) or deliver adjunct interventions (ie, interventions not included in TAU).

Textbox 1. Study inclusion criteria following the Population, Intervention, Comparison, and Outcome principles.

Population diagnosis

- Schizophrenia
- Schizophreniform
- Schizoaffective
- Psychotic disorder not otherwise specified
- Psychosis

Type of technology

- Mobile apps
- SMS text messaging
- Web based
- Videoconferencing
- Wearables
- Blended interventions

Outcome target

- Symptoms
- Functioning
- Service engagement
- Illness management

Study methodology

- Randomized controlled trials
- Pilot trials
- Qualitative studies
- Feasibility studies
- Protocol papers

Search Strategy

A total of 3 core databases were searched: PsycINFO, Embase, and MEDLINE. In addition, Cochrane and PROSPERO databases were searched for existing reviews and protocols. References of resulting included studies were hand searched.

Search strategies used for general databases were as follows:

- Population: Psychosis OR Schizophrenia OR Schizoaffective OR Schizophrenia Spectrum OR Psychotic Disorders OR First-Episode Psychosis OR Early-Episode Psychosis
- Intervention: SMS OR Short Message Service OR SMS-Survey OR Texting OR Text Message OR SMS Based System OR SMS Reminder OR Text Message Reminder OR Digital Health OR Telehealth OR Mobile Apps OR Mobile Applications OR Mobile Health OR eHealth OR mHealth OR Wearable Technology

Search strategies used for databases powered by OVID were as follows: schizophrenia spectrum.mp. OR psychotic disorder.mp. OR exp psychosis/AND (sms or short messag* service* or texting or text messag*).mp. OR (mobile apps* or smartphone app* or telehealth) OR (eHealth or mHealth or mobile health or internet intervention or web-based treatment or web-based intervention or wearabl*).mp.

Reviewer Protocol

Studies resulting from the initial search were exported into a reference manager (EndNote; Clarivate) [23], where the initial deduplication occurred, and then transferred into a web-based review management system, where the system again identified duplicates (Covidence; Veritas Health Innovation) [24]. This platform also allowed for the independent review of each study for inclusion. Authors (JD, MI, LT, AZ, and TA) conducted the review using the title, abstract, and full text. Review conflicts were reviewed by the first author (JD) and team and, if needed, by the last authors (SK and GF). At least 2 reviewers reviewed each publication.

Data Extraction

Both qualitative and quantitative data were extracted from the included studies. First, methodological information was collected, including research design and publication type, sample size, diagnostic group, length of study, and outcome measures. Second, intervention information was collected on the digital intervention target, the type of technology used, the evidence-based approach used, and the length of the intervention. Third, study outcome data were extracted, including qualitative or descriptive and quantitative data regarding primary outcomes. Both qualitative and quantitative data are reported descriptively.

Analysis

The reported results are descriptive, in line with scoping review methods, and provide a high-level overview of key findings. Extracted data are charted according to the type of technology used, followed by subcategories regarding the stage of research (eg, development, pilot feasibility and efficacy, and effectiveness and implementation trials), themes, and issues uncovered through data synthesis. Of note, we distinguish between blended care and multitechnological interventions. Blended care is defined as integrating technology into traditionally delivered care (ie, in-person care) [25], whereas we define multitechnology interventions as using >1 technological platform and may be integrated with traditional or virtual care. Furthermore, given the large number of included studies, findings report on primary outcomes. A more detailed review would be required to investigate secondary and tertiary findings. Primary outcomes pertaining to feasibility use highly varied measures; therefore, we provided a high-level description of outcomes.

Results

Overview

Our search yielded 999 studies, of which, through careful review, 92 (9.21%) studies were included (refer to Figure 1 for details of the search exclusion), investigating approximately 50 unique interventions. It is difficult to know the exact number of unique interventions as this is not always clearly reported. Included studies were published in a range of countries, predominately in North America (40/92, 43%), Europe (33/92, 36%), Australia (13/92, 14%), and East Asia (7/92, 7%). Studies were published from 2010 to 2023, with the majority published between 2018 and 2020 (50/92, 54%; Figure 2).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) breakdown of search results. Description of the review process that determined the included and excluded studies.

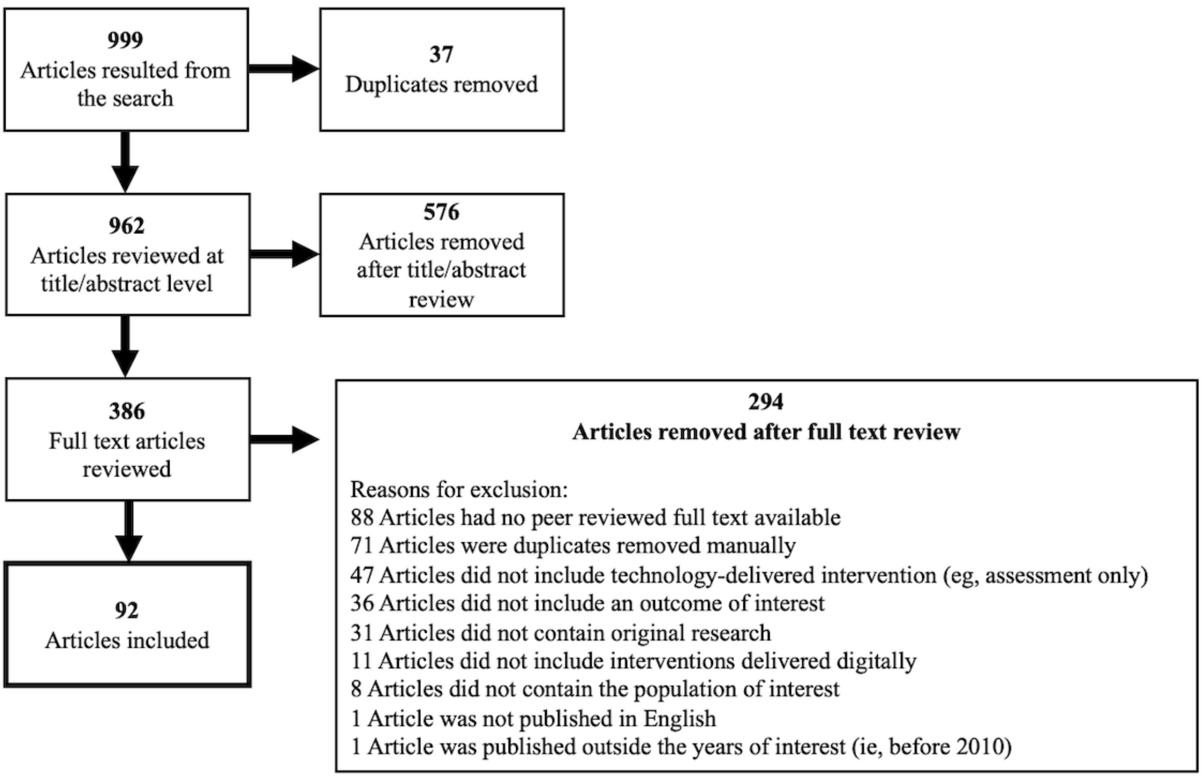
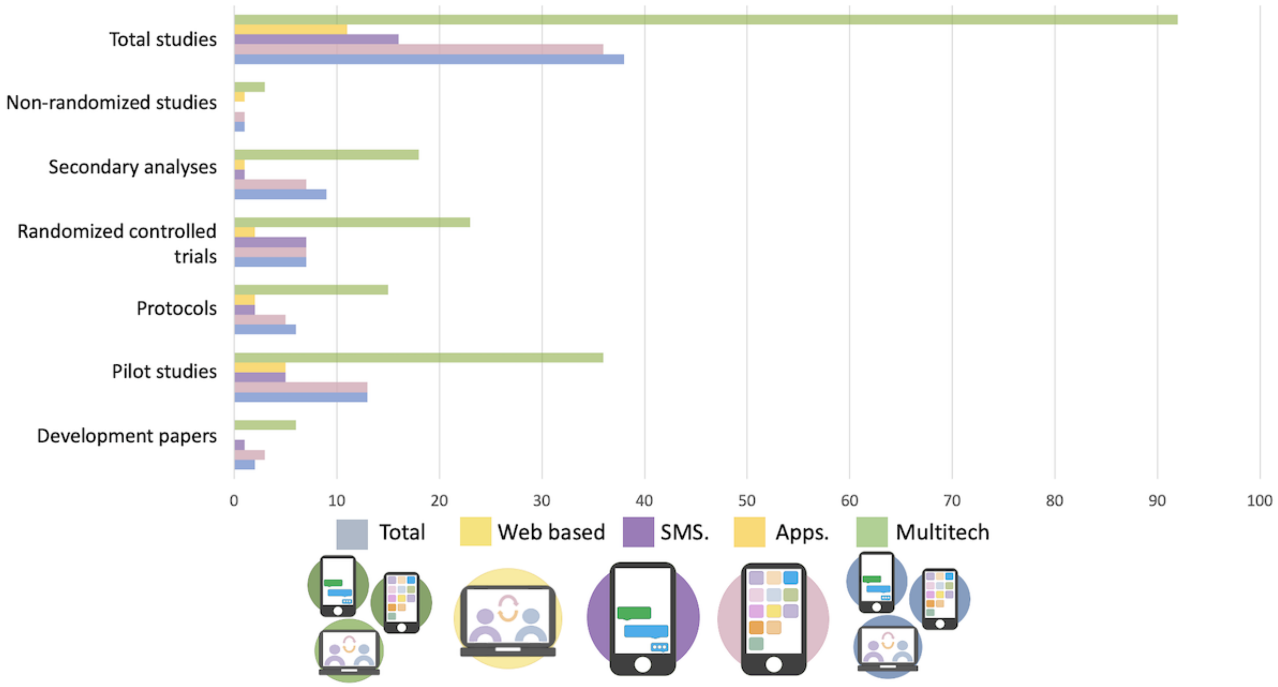


Figure 2. Number of publications over time by technological modality.



We primarily examined results through the intersecting lens of technology type and stage of research (ie, study design and aim). Among the 92 included studies, 11 (12%) focused on web or internet-based interventions, 25 (27%) examined smartphone apps, 16 (17%) investigated SMS text messaging, and 40 (43%) used multiple types of technology to deliver the interventions. No included studies used wearable technology to deliver an intervention. Studies included in this review that used videoconferencing or phone calls did so in combination with other types of technologies and are thus counted under multitechnology interventions.

Regarding the stage of the included research, of 92 studies, there are 5 (5%) qualitative studies investigating development and patient experiences, 32 (35%) pilot trials (26/32, 81% open trials and 6/32, 19% randomized control trials [RCTs]), 13 (14%) protocol papers, 23 (25%) conducted RCTs, 16 (17%) secondary analyses (both quantitative and qualitative), 3 (3%) nonrandomized trials, and no implementation trials. Given the number of secondary analyses, it is important to note the number of unique technologies being studied to reflect a potentially more accurate picture of the spread of digital mental health that is leveraging personal technologies in this population. There are 50 unique technologies studied within the 92 included studies. Several unique technologies (19/50, 38%) are investigated in >1 publication, with several published protocols suggesting ongoing investigations using these interventions. Contextually, it is important to note that all included trials recruited from mental health care clinics and institutions providing traditional care, and as such, all interventions included in this study can be considered blended or integrated to varying degrees with traditional care.

Overall Feasibility

Factors of feasibility in included studies vary, such as measures of acceptability, usability, technology engagement, and user feedback. The vast majority of included studies report high levels of acceptability and usability. Technology and user feedback are mixed and reveal several high-level themes. First, technology engagement tends to decrease over time regardless of technology type. Second, many studies report self-reflection and therapeutic rapport as positive features of using these technologies. Third, various types of prompts were used with varying success to enhance engagement.

Safety and Privacy Outcomes

Across all technologies and all study designs, 60% (55/92) of the studies did not report on any safety indices at all. The most common indicator (31/92, 34%) of safety used was reporting serious adverse events (SAE) and adverse events (AE). Most studies reported that no SAEs or AEs were related to the technological intervention, and if there were any present, they provided a short description of the SAE or AE. Furthermore, the vast majority of the studies (64/92, 70%) did not comment on the privacy measures taken by the study team or built into the technology used to deliver the intervention.

Overall Efficacy

There were 16 RCTs and 1 pilot RCT examining the efficacy of the technological intervention as its primary outcome; see Table 1). Regarding methodology, most RCTs use TAU (10/17, 58%) as their comparison group, with others using varying active comparisons, including phone calls, films, sham apps, and other therapies. RCTs were most commonly reported using single-blind designs (7/17, 41%); however, 2 (12%) were double blinded, and 8 (47%) did not use blinding procedures for their assessors or participants. Approximately 52% (9/17) of the trials showed a positive change (ie, improvement in target) during the course of the intervention in favor of the technology, with another showing positive change in 1 primary outcome but not both [26] and another showing no overall change but a change in the positive symptom subscale scores [27]. Compared to the control, 47% (8/17) of the RCTs show significant group differences in their primary outcome, with the same 2 studies reporting significant group differences in part of their primary outcomes [26,27]. Only 4 studies reported effect sizes, with 3 being interpretable as medium to high and 1 as low [28-31]. The type of effect sizes reported were inconsistent across studies and thus were interpreted as low, moderate, and high based on suggested interpretation guidelines for each type of effect size to aid in comparison. A more detailed examination of efficacy, which includes all study types with primary efficacy outcomes, can be found in the sections below divided by technology type.

Table 1. Included original randomized control trials with primary efficacy outcomes.

Author, year	Technological intervention information					Study design information				Study outcomes			
	Type	Name	Target	Approach	Length (months)	Population	Sample size	Control	Blinding	Primary outcome	Significant change	Significant group difference	Effect size
Schlosser et al [32], 2018	App[a] and SMS text message + hone	PRIME	Motivation	Mixed	3	SSDs[b]	32	TAU[c]	Single	Motivated behavior	No	No	NR[d]
Fisher et al [28], 2023	App and videoconference	PRIME	Motivation	Mixed and CR[e]	4	SSDs	100	Sham	Double	Cognitive functioning	Yes	Yes	Medium to high
Garety et al [33], 2021	App and web-assisted therapy (adjunct to 8 in-person sessions)	SlowMo	Paranoid ideation	CBT[f]	3	SSDs	361	TAU	Single	Paranoia	Yes	Yes	NR
Krzystanek et al [34], 2019	App, provider web portal, and telemedicine	MONEO	Symptom severity	Symptom monitoring and CR	12	Schizophrenia	290	Sham	None	Symptom severity	Yes	Yes	NR
Schulze et al [35], 2019	SMS text messaging and phone calls	Telca	Medication adherence	NR	6	Schizophrenia or BD[g]	120	TAU	None	Medication adherence	Yes	Yes	NR
Stentzel et al [36], 2021	SMS text messaging and phone calls	Telca	Medication adherence	NR	6	SMI[h]	118	TAU	None	Quality of Life	No	No	NR
Depp et al [29], 2019	App and single in-person therapy session	CBT2Go	Symptom severity	CBT	3	SSDs or BD1	229	TAU and sham	Single	Symptom severity	Yes	Yes	Low
Zhu et al [30], 2020	Embedded onto social media app	WeChat	Medication and illness management	NR	6	Schizophrenia	84	Phone calls	Single	Medication adherence	Yes	Yes	High
Beebe et al [26], 2014	SMS text messaging versus phone Calls	TIPS	Medication adherence and symptom severity	NR	3	SSDs	28	SMS text messaging vs phone vs combined	Single	Medication adherence and symptoms severity	Adherence: no; symptoms: yes	Adherence: no; symptoms: yes	NR
Cullen et al [27], 2020	SMS text messaging	T4RP	Relapse prevention	NR	6	Schizophrenia or SZA[i]	40	TAU	None	Symptom severity	Yes: positive symptoms only	Yes: positive symptoms only	NR
Montes et al [37], 2012	SMS text messaging	NR	Medication adherence	NR	3	Schizophrenia	254	TAU	None	Medication Adherence	Yes	Yes	NR

Author, year	Technological intervention information					Study design information				Study outcomes			
	Type	Name	Target	Approach	Length (months)	Population	Sample size	Control	Blinding	Primary outcome	Significant change	Significant group difference	Effect size
Pijnenborg et al [38], 2010	SMS text messaging	NR	Activity achievement	NR	1.75	SSDs	62	TAU	None	Goal attainment	No	NR	NR
Välimäki et al [39], 2017	SMS text messaging	NR	Relapse prevention	SDT[j]	12	SSDs	1139	TAU	None	Rehospitalization	NR	No	NR
Xu et al [31], 2019	SMS text messaging	LEAN	Medication adherence	NR	6	Schizophrenia	278	TAU	Single	Medication Adherence	Yes	Yes	Medium
Kaplan et al [40], 2011	Internet based	NR	Symptom severity	Peer support	12	Psychosis	300	Three conditions	None	Symptom severity	No	No	NR
Westermann et al [41], 2020	Internet based	iCBTp	Symptom severity	CBT	2	SSDs	101	TAU	Single	Symptom severity	No	No	NR
Dabit et al [42], 2021	App, telemedicine, and SMS text messaging	CLIMB	Social functioning	SCT[k]	2.25	Schizophrenia or SZA	24	Sham	Double	Social functioning	Yes	No	NR

[a]App: smartphone app.
[b]SSD: schizophrenia spectrum disorder.
[c]TAU: treatment as usual.
[d]NR: not reported.
[e]CR: cognitive remediation.
[f]CBT: cognitive behavioral therapy.
[g]BD: bipolar disorder.
[h]SMI: serious mental illness.
[i]SZA: schizoaffective disorder.
[j]SDT: self-determination theory.
[k]SCT: social cognition training.

Detailed results are outlined in the subsequent sections and focus on the stage of research and notable results. Key gaps and future directions are explored in the *Discussion* section. A summary of the included studies can be found in Multimedia Appendix 1 [26-120].

Multiple Technology Interventions

Overview

The largest proportion (40/92, 43%) of the included studies investigated interventions that leveraged multiple types of personal technologies. Multiple distinct personal technologies within 1 intervention were identified if they served (1) distinct purposes or (2) two distinct modes of communication. For example, an ecological momentary assessment (EMA) embedded with a mobile app would be considered 2 distinct technologies as an EMA collects information, whereas the mobile app shares information. In contrast, an addition of modules to an already existing web platform serves the same purpose as the original web platform and would count as 1 technology. Most commonly, interventions combined the use of a smartphone app and a web page that could be accessed by either a clinician or peer support [29,33,34,43-49], with some additionally combined with targeted in-person intervention sessions [33,47]. Other technology combinations included smartphone apps to deliver resources with added EMA and intervention approaches [50] and a web page [51] or wearable technology [52] to help users self-manage their mental health. Other combinations still included using a smartphone app or web page in combination with communication technology such as SMS text messaging [29,53,54], phone calls, videoconferencing [32], and email [55]. Finally, 2 interventions engage various communication strategies, such as SMS text messaging and phone calls [35,56] and a smartphone app, SMS text messaging, and telemedicine [42].

Of the 40 included studies investigating interventions leveraging multiple personal technologies 2 (5%) were development papers

[57,58], 14 (35%) were pilot studies [42,43,45,49,51,53,54,59-65], 6 (15%) were RCT protocols [47,48,52,56,66,67], 8 (20%) were RCTs [28,29,33-36,64], 9 (23%) were secondary analyses [44,55,68-72,74,75], and 1 (2%) was a nonrandomized trial [46]. There are 17 unique multitechnology interventions investigated using the 40 included studies. Targets of the multitechnology interventions included illness self-management (5/17, 29%) [43,52,55,58,62], relapse prevention (3/17, 18%) [46,47,49], medication adherence (2/17, 12%) [51,56], general symptom severity (2/17, 12%) [29,34,48], social functioning (1/17, 6%) [42], motivation (1/17, 6%) [53], coping with auditory hallucinations (1/17, 6%) [50], negative symptoms (1/17, 6%) [54], and cognitive function (1/17, 6%) [70]. Evidence-based approaches are reported by 77% (13/17) of these interventions, including cognitive behavioral therapy (CBT; n=3, 23%); peer support (n=1, 8%); cognitive remediation (n=1, 8%); information motivation behavioral skills (n=1, 8%); connectedness, hope and optimism, identity, meaning, and empowerment (n=1, 8%); social cognitive theory (n=1, 8%); and mixed approaches (n=6, 46%). For all studies (29/92, 32%) that investigated any of the 17 multitechnology intervention, the trial lengths ranged from 1 visit to 60 months, with a mode of 3 months.

Feasibility

Studies (23/40, 58%) investigating feasibility-related variables as the primary outcome reported varying measures of engagement, including study retention, intervention retention, and various means of measuring technology engagement (eg, number of log-ins per period, number of tasks completed, and number of responses). Regarding study retention, included studies reported good retention rates of 78% to 100% [28,43]. However, technology or intervention retention tended to decrease over time and was 42% to 52% at the end of the intervention periods [43,60]. Technology use over time also tended to decrease over time regardless of the measurement [28,43,60]. Despite this, studies reported high levels of response or completion throughout the study period (70%-84%) [49,51,53,59]. One study reported a relatively low response rate (22%); however, this represented a homework completion rate of at least once per week [54]. Another study reported on clinician engagement with the connected clinician portal (67% log-in rate), which was much lower than patient engagement (80% response rate) [49]. Some studies reported on factors impacting intervention engagement, which included low mood or depression [44,68], negative symptoms and low motivation [44], and fluctuations in wellness [71]. Technological factors were also reported to impact engagement. Specifically, ease of navigation, accessibility of resources, the fit of technology [44,71], and concerns regarding privacy [69] impacted engagement, whereas coaching facilitated engagement [53,55] as long as it did not feel too scripted [53]. Overall, patients tended to endorse the technologies as easy to use and helpful and were satisfied with the technology or interested in continued use [43,49,51,53,117].

Of note, 7 (30%) feasibility studies of the 23 included multitechnology studies focused on 1 intervention [58,61,63-65,72,75] in varying populations (ie, psychotic and mood disorders), in different contexts (ie, outpatient, assertive community treatment, and veteran association hospital), and with different comparison groups and augmentations (ie, the addition of video content and phone calls from interventionists). Throughout these studies, participants and clinicians rated the acceptability and usability of the intervention as high. Participants could access the app at any time, and the app sent out prompts and phone calls to support app engagement. Participants responded to most prompts [64,65] and phone calls from the mobile case manager or interventionist [65]. Despite this, patient app engagement over time declined in the studies that measured use over time [63,64] but reported better engagement over time compared to an in-person group [64]. One study reported on cost-effectiveness and found that the technologically supported intervention was significantly less costly, with no significant differences in efficacy compared to a fully clinician-delivered intervention [73].

Efficacy

Studies examining the efficacy of multitechnology interventions as a primary outcome (10/40, 25%) [29,32-36,42,46,70,74] reported mixed findings across a wide range of specific outcomes. One study examining self-guided CBT did not report an overall significant difference in symptom severity compared to TAU [29]. Another study investigating technology-assisted, clinician-delivered CBT for paranoid thinking found that the intervention significantly reduced paranoia across 3 measures at the end of a 3-month intervention compared to TAU [33]. Studies investigating social-cognitive-functional domains reported significant impacts on aspects of global cognition, attention [28], motivated behaviors (ie, anticipatory pleasure and improved effort expenditure for future social interaction) [32], improvements in affective symptoms [34], the identification of correct answers and cognitive fatigue [70], and social functioning [42]. A smartphone app–clinician portal alert system reported significant reductions in relapse, rates of hospitalization, and urgent care needs [46]. An intervention comprising biweekly clinician phone calls and weekly SMS text message check-ins showed improvements in medication adherence compared to controls [36]; however, it did not show improvements in other areas such as global functioning or quality of life [35].

Smartphone App Interventions

Overview

Of the 92 included studies, research investigating the use of smartphone apps included 25 (27%) publications, of which 2 (8%) were development papers [76,77], 8 (32%) were pilot studies [76,78-84], 1 (4%) was a nonrandomized study [85], 6 (24%) were protocol papers for RCTs [50,86-90], 3 (12%) were RCTs [30,91,92], and 5 (20%) were secondary analyses [93-97]. The average length of interventions was 4 months (range: 1 visit to 12 months and modes: 3 and 12 months). Pilot feasibility studies typically lasted 1 to 2 months; pilot efficacy trials lasted 2 to 6 months; and randomized trials lasted 3 or 12 months, including protocols. Sample size within trials is also highly variable (mean 41.65; median 31, range 8-229).

There were 15 unique smartphone app interventions investigated in the 25 included studies. Most of the included smartphone interventions were positioned as adjunctive to traditional mental health care or TAU and were aimed at illness self-management [76,80,88,91]; medication adherence [30]; and symptom severity [81,86,89,92], including specific domains such as auditory hallucinations [84,90], social skills [79,82], and metacognition [85]. Most apps reported evidence-based foundation therapeutic approaches, most commonly CBT (6/15, 40%) [76,78,81,83,84,89], followed by social skills training (1/15, 7%) [77], positive psychology (1/15, 7%) [82], acceptance and commitment therapy (1/15, 7%) [76], metacognitive intervention for schizophrenia (1/15, 7%) [85], or report mixed approaches (3/15, 20%) [88,91]. However, some (2/15, 13%) [30,80,90] interventions do not report following a specific evidence-based approach or omit this information. Two (2%) of the 15 included app interventions were specifically blended with in-person care [84,92].

Feasibility

Of the 25 studies investigating apps, there were 12 (48%) studies that reported primary outcomes related to the feasibility of apps [78,80-84,91-96]. All feasibility studies investigating apps reported high levels of acceptability, usability, and satisfaction. Studies using a prompt or EMA system reported medium to high response rates of 33% to 81% [33,92,93]. Studies reporting on task or module completion also reported medium to high completion rates of 42% to 95% [78,81,82,92]. Of 12 feasibility studies, 2 (17%) studies reported that engagement with apps decreased over time [91,94], 1 of which reported that 50% stopped using the app within 3 months [91]. Facilitators of app engagement were functions allowing synchronous communication with clinicians [81], integration of face-to-face sessions [96], and stronger therapeutic alliance [94], whereas barriers to engagement were lack of clinician support using the app, concerns regarding therapy [91], feeling the app was not personable, and there were too many sessions or prompts [81,83]. Some studies found participant factors associated with app engagement, such as age, employment status, race, and smartphone ownership [93,94].

Efficacy

Of 25 studies, 4 (16%) examined the efficacy of interventions delivered via smartphone apps [30,79,85,98], of which study examined an app delivering self-guided CBT [98]. Regarding overall symptom severity, a pilot study found significant changes across recovery and symptom outcomes [98]. Another pilot study examined the delivery of self-guided social skills training and found a moderate impact on social functioning at the end of the intervention; however, it found that gains were not maintained at follow-up [97]. A nonrandomized control comparison study found that a smartphone-delivered metacognitive therapy augmented by weekly mentoring sessions found positive impacts on metacognition, delusions, general pathology, and negative symptoms [85]. Finally, an RCT examined the effect of medication reminders sent through a smartphone app and reported that compared to TAU, those receiving reminders showed significantly better medication adherence, with large effect sizes reported [30].

SMS Text Messaging Interventions

Overview

In this review of 92 studies, 16 (17%) were included in which an intervention using SMS text messaging was investigated. Of the 16 papers, 1 (6%) was a development paper [99], 2 (12%) were protocol papers [100,118], 4 (25%) were pilot trials (ie, 2 (12%) were open trials [101,102] and 2 (12%) were pilot RCTs [103,104]), 6 (38%) were RCTs [26,27,31,37-39], and 3 (19%) were secondary analyses [105-107]. The sample sizes varied greatly between 14 and 1139 (mean 181, SD 274.7); however, most were <100 participants (10/16, 62%). The length of the intervention varied from 1 to 18 months, with a mean of 4 months and a mode of 6 months.

There were 11 unique SMS interventions investigated within the included 16 studies. Most (n=9, 82%) did not report following a specific evidence-based model, with one reporting the use of CBT [102], and another Self-Determination Theory [39]. Interventions using SMS text messaging targeted a variety of outcomes, mainly in the form of reminders, such as medication adherence (4/16, 25%) [26,31,37,102], appointment attendance (1/16, 6%) [100], and relapse and rehospitalization rates (1/16, 6%) [27]. Beyond service engagement, some studies aimed to support clinical outcomes, such as auditory hallucinations [102], social skills [102], goal attainment [38], and general support [39,101,103,105]. Most SMS-based interventions were framed as adjunctive care, and only 1 was reported as being intentionally blended with traditional care [103]. Most SMS-based interventions did not report to follow a specific evidence-based approach. One intervention was based on CBT [102], and another was based on self-determination theory [39].

Feasibility

Of the 16 studies investigating SMS text messaging, 5 (31%) reported feasibility as the primary outcome. Studies reported high text message response rates (69% to 87%) [101,103,104] and high levels of satisfaction, ease of use, and helpfulness of SMS text messages [103,104,107]. One of the 5 feasibility studies even found that therapeutic alliance was higher with the SMS text message interventionists compared to participants' community-based clinicians [101]. Another study explored areas that participants found useful to discuss with SMS text message interventionists, including mental health symptoms, coping strategies, lifestyle or well-being, social or leisure activity, motivation, and independent living skills [106].

Efficacy

Of the 16 included studies that investigated SMS text messaging, 13 (81%) reported on efficacy outcomes. The reported efficacy of SMS text messaging interventions seemed to vary depending on the outcomes measured. Most studies investigating SMS interventions aimed at medication adherence were associated with greater rates of adherence [31,37,102,105], especially among individuals with schizophrenia who were living independently [102]. Effects on medication adherence may be bolstered by the addition of phone check-ins by clinicians [26]. However, drop-off effects may occur once the reminders are no longer delivered [119]. In addition, there seemed to be potential positive effects on rehospitalization rates [31]; however, results were mixed as other researchers did not find evidence of a reduction in hospital admission rates, the time between hospitalizations, or time spent in a psychiatric hospital [39]. Regarding changes in recovery and symptom severity, some studies reported improvements in negative symptoms, positive symptoms [27], cognition, and global clinical symptoms [37]. These results were strengthened when the SMS text messaging intervention was coupled with a secondary intervention (eg, telephone intervention or assertive community treatment) [26,103]. Other researchers did not find significant improvements in symptom severity [39]. Mixed outcomes were reported for social and community functioning [38,104,105]. Studies reported mixed findings regarding service engagement. One study reported improved therapeutic alliance with their mobile clinician compared to in-person clinician visits, and participant feedback results reflected a sense of support among users when receiving daily SMS text messages [101,106].

Internet-Based Interventions

Overview

Of the 92 included studies, 11 (12%) described internet-based interventions that were not paired with other technology. Of these 11 papers, 1 (9%) was a cross-sectional study [108], 5 (45%) were pilot trials [109-112,120], 2 (18%) were RCT protocols [113,114], 2 (18%) were RCTs [40,41], and 1 (9%) was a secondary analysis [115]. The sample size of included studies examining internet-based interventions ranged from 10 to 300, with approximately 50% being >100 (mean 97). The length of the intervention ranged from 2 to 26 (mean 9; modes 2 and 18) months. There were 8 unique interventions investigated within the included 11 studies. The internet-based intervention treatment targets included recovery [110,113], cognition [112,120], working memory [111], symptom severity [40,41], and illness self-management [108]. Most internet-based interventions reported to be based on an evidence-based framework, including connectedness, hope and optimism, identity, meaning, and empowerment [110]; cognitive remediation [111,112,120]; CBT [41]; peer support [40]; and mixed approaches [113].

Feasibility

Of the 11 studies examining internet-based interventions, 6 studies examined feasibility using varied measures of feasibility and engagement. Study attrition was reported between 64% and 100% [109,112,114,120]. Given the range of interventions delivered, intervention engagement was measured differently in each study based on the intervention design. Of the 11 studies, 1 (9%) reported the number of log-ins (eg, mean 39.2 over the intervention) [109], 1 (9%) reported the mean time spent on the website per week (eg, 3 h/wk) [120], another (9%) reported percentage of participants completing at least 80% of modules according to a predetermined meaningful completion rate (eg, 70% of participants completed 80% of modules) [112], and 1 (9%) reported the percentage of participants that used the website (eg, 41% of participants used the website) [108]. Of the 11 studies, 2 (18%) reported that participant factors impacted engagement, such that younger age [108,120], higher education [108], being employed [108], and lower cognitive symptoms [120] led to better engagement.

One study specifically reported integrating a website into regular therapy sessions and found that the website was used in 95% of therapy sessions [114]. Qualitative feedback related to the use in sessions included that the website helped facilitate conversation and reflection in their sessions and was a key tool for engagement, leading to increased perceived recovery [114]. However, only 60% of participants reported use outside of sessions [114]. Another study presented findings related to a self-guided website [115] and found that the self-guided nature to some was motivating and cultivated a sense of autonomy in recovery, whereas others found the self-guided nature overwhelming, which interrupted engagement. Positive aspects reported related to the self-guided website were having at-demand resources more so than any clinician-driven service, having the means to distract from distress in a meaningful, positive way, and having the ability to interact with peers and psychoeducation was a normalizing experience [115].

Efficacy

Of the 11 studies examining internet-based interventions, there were 3 (27%) that reported primary outcomes pertaining to the efficacy of internet-based interventions [40,41,111]. Each of these interventions focused on different evidence-based approaches. One study examined the impact of online peer support (ie, comparing anonymous listserve group email communication or communication with peers via web-based bulletin board compared to TAU) on subjective recovery, quality of life, empowerment, and social support [40]. No significant group differences were found with regard to online peer support and the control group for either the listserve or bulletin board communication types [40]. Another study investigated remote cognitive remediation (ie, web-based computerized tasks, psychoeducation, and strategy development paired with weekly clinician calls) targeting working memory [111]. Compared to TAU, remote cognitive remediation was associated with improved working and episodic memory, with medium effect sizes reported [111]. Finally, another study looked at the effects of self-guided CBT for psychosis on various psychotic symptoms compared to a waitlist condition [41]. Individuals receiving self-guided CBT for psychosis demonstrated greater improvements in symptoms overall and self-reported hallucinations compared to the waitlist condition, with a small to medium effect size reported [41]. However, the study did not find significant differences in other positive psychotic symptoms [41].

Discussion

Principal Findings

The primary aim of this review was to provide a high-level overview of the use of personal technologies in research on digital mental health interventions for individuals with SSDs. Publication patterns show rapid growth in the area over the past 5 to 7 years, highlighting increasing interest in this area [10-12]. The lower number of publications reported for 2021 to 2023 is likely partly attributable to a combination of the large number of protocols published in 2020, as many of these trials may still be underway, and research challenges related to the COVID-19 pandemic. Personal technologies leveraged most often included the use of smartphone apps or an app in combination with other technology, such as a clinician web portal.

Our review confirms that personal technologies have already garnered acceptance among individuals with SSDs, as evidenced by consistently high ratings of acceptability, usability, and satisfaction. Some studies even report that augmenting care with technology improved engagement in traditional in-person or remote care [44,101,106]. In included studies which report participant feedback, participants typically report that these technologies facilitate self-reflection and understanding, better communication with their clinical teams, improved access to evidence-based resources, and support in times of need. Findings related to feasibility presented in this review are similar to findings reported elsewhere for groups with SMI [121].

Despite the high levels of acceptance and interest, sustained engagement in technological interventions is an ongoing concern [122] and is not well understood. Measures of engagement are highly inconsistent across studies, and include module completion, response rates, skills and homework completion, and number of posts, making it almost impossible to understand a cohesive story regarding engagement with digital tools in groups with SSDs. While simple metrics such as screen clicks or the number of days or hours used are commonly reported, the field lacks an understanding of clinically meaningful engagement for this metric [123]. Beyond access metrics, there is no consistent measure of technology use. The challenge of heterogeneity in engagement metrics is already well known [124], but a lack of progress in agreement around these common metrics hinders future advances in the field [125]. Furthermore, likely in part owing to a lack of definition in this area, means of addressing technology disengagement are also poorly understood. A review examining the use of persuasive system design (design features designed to address or enhance engagement) in technologies designed for use in depression and anxiety reported that persuasive system design efforts did not systematically lead to improved engagement but are associated with improved efficacy [126].

Moreover, some ongoing challenges may impact the feasibility of using such technologies in practice, including understanding the link between privacy protections embedded in these technologies and meaningful engagement. Privacy concerns are seldom outlined in the included publications, which may result from the dearth of available guidelines, policies, and reporting standards. Privacy concerns are often cited by health care providers as a key barrier to the use of technology in practice [127], and as such, to encourage technological adoption, more data on privacy features needs to be made available. There is also a dearth of information regarding patient safety and urgent or crisis resources and support. In this review, most of the included studies (55/92, 60%) did not report on safety at all, which is well above the proportions reported across the field of digital mental health (35%) [128]. The most common safety indicators reported in this review and across the field [128] were SAEs and other AEs, which are often mandatory reporting standards for ethics boards and do not provide nuanced information regarding technology use and patient well-being.

Furthermore, there are promising early results regarding efficacy, but overall findings are mixed. Pilot studies tend to show more optimistic outcomes across a range of mental health outcomes; however, once studies reach the RCT phase, findings become more mixed. Similar descriptions are found in other areas of SMI, such as bipolar disorder [129]. In areas such as depression and anxiety, where more research on efficacy has been conducted, clearer conclusions about effectiveness can be drawn [130]. In the included RCTs, it does not appear that study procedures, such as blinding, type of technology, or sample size, systematically impact outcomes, as no obvious patterns have emerged between study type or methodology. However, with few studies featuring an active control group, any assessment of efficacy is still preliminary. Future studies should consider the active intervention when planning a control group and the purpose of the technology (eg, if testing an evidence-based treatment that is typically delivered in person and not included in TAU, then TAU would not be an active control group, such as an in-person service delivering the evidence-based treatment would better represent an active control). However, there are some interventions that have garnered significant support and exhibit positive early results [33,53,60,84,86,113].

Of note, all included studies used samples of individuals currently in treatment and thus were all offered adjunctly to traditional clinician-delivered care. However, few studies described how the technology was blended or integrated into the care structure. As such, the field provides little direction on the role of technology and its integration into existing care models. Furthermore, there were few clinical integration and effectiveness studies, underscoring the next frontier for this research in SSDs. It is important to understand how external factors, including primary treatment methods, may impact engagement, efficacy, and effectiveness. For example, the included studies that described the role of technology within the clinical context described technology-forward (ie, self-guided and clinician-supported) and clinician-forward (ie, clinician-delivered and technology-supported) interventions. Clinician-forward interventions, in our small sample of studies that described clinical integration, seemed to demonstrate better engagement [96] and better outcomes [33,84,96] than technology-forward interventions, depending on the target and treatment approach [29].

Other reviews reported mixed findings in relation to blended (technology-supported) approaches. One review and meta-analysis of digital mental health interventions across disorders suggested that technology-supported interventions outperform (Hedges g=0.16; P<.001) traditional TAU (solely clinician-delivered) [131]. However, this review did not report on the more nuanced technological differences that we highlight in this review regarding whether the intervention was delivered by the technology (ie, technology-forward) or supported clinician-delivered care (ie, clinician-forward). Another review and meta-analysis suggested that technology engagement was associated with efficacy regardless of intervention type (ie, guided or unguided) [132], suggesting that engagement may be the underlying mechanism for effectiveness in these trials rather than the level of clinician involvement. Therefore, more research is needed to define the role of technology within clinical spaces to understand the implementation and best practices related to technology-assisted care.

Future research in this promising area should focus on the identified key gaps to move the field toward implementation readiness. First, studies should clarify the technology's intended role, intervention targets, and functionality. This is important for understanding implementation into existing clinical structures and interpreting study outcomes. Second, given the significant barriers of safety and privacy to clinical adoption, a more detailed exploration of safety and privacy indices is critical. Finally, a better understanding of barriers to participating in and disengagement from digital interventions is crucial for understanding its usefulness and scope in clinical settings.

Limitations

There are several limitations to this review. First, as a scoping review, we could not provide a detailed review of each study but rather a high-level examination of the findings in the field to date. Therefore, data have been synthesized and presented in accordance with outlined scoping review guidelines, and, as such, the methodological quality of each study was not rigorously reviewed. Second, we limited our search to using personal technologies to deliver or support specialized care for individuals with SSDs. Other technologies are being used in SSD treatment that this review did not include, such as virtual reality, avatar therapy, cognitive remediation, brain stimulation, and more. Third, we restricted this review to outcomes directly related to general recovery (ie, core symptoms and functioning) and treatment engagement and did not focus on specific domains such as cognitive deficits. Fourth, other interventions exist for comorbid conditions, such as substance use and smoking cessation [133], physical activity [134], and resources for family and social support [135]. Finally, we only included technologies that were used to deliver interventions; we did not include studies that were used to improve the understanding of SSDs, such as observational methods (eg, technology-assisted EMA or health services–based interventions, including those that are embedded in electronic medical records for medication algorithms, symptom monitoring, or other purposes). Future reviews may consider focusing on these additional areas of digital mental health to understand how other technologies may support the recovery of individuals with SSD diagnoses.

Conclusions

Overall, using personal technologies to deliver specialized care requires more careful consideration before advocating for broad implementation, highlighting the same challenges as other psychotherapeutic intervention research. Namely, the prolonged time between investigation and implementation is due to varied findings and a lack of cohesive targets and direction. Despite these challenges, as evidenced by this review, there is great promise for leveraging personal technology in mental health care to help provide pieces of the holistic care necessary for recovery in SSDs. Multifaceted mental health conditions such as SSDs are highly heterogeneous and may require multifaceted and flexible interventions. Although the exact nature of meaningful technological support is still being discovered, the studies in this review overwhelmingly demonstrate how flexibly personal technologies can support recovery from medication adherence to potentially delivering complex psychosocial interventions such as CBT. Furthermore, studies clearly demonstrated the feasibility of personal technologies to extend access to specialized information beyond traditional care and into everyday life. Therefore, these technologies are well suited to be integrated into existing specialty care structures for individuals with SSD diagnoses.

Acknowledgments

This study is funded by the Canadian Institute of Health Research. The first author (JD) is also supported by a Canadian Institute of Health Research Vanier Graduate Scholarship.

Conflicts of Interest

JT is the editor in chief of *JMIR Mental Health*. Research examining the App4Independence (A4i) is included in this review. A4i was developed and owned by the joint for-profit venture A4i. A4i is jointly owned by MEMOTEXT, the Centre for Addiction and Mental Health, and SK – hence having author engagement in this company.

Multimedia Appendix 1
Supplemental table of included studies divided by technology type.
[DOCX File , 1132 KB - mental_v11i1e57150_app1.docx]

Multimedia Appendix 2
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 498 KB - mental_v11i1e57150_app2.pdf]

References

1. Wozney L, Mcgrath P, Newton A, Hartling L, Curran J, Huguet A, et al. RE-AIMing e-mental health: a rapid review of current research. Mental Health Commission of Canada. 2017. URL: https://www.mentalhealthcommission.ca/wp-content/uploads/drupal/2017-08/eMH%20Literature%20Review_FINAL%20EN.pdf [accessed 2024-04-29]

2. Kasckow J, Felmet K, Appelt C, Thompson R, Rotondi A, Haas G. Telepsychiatry in the assessment and treatment of schizophrenia. Clin Schizophr Relat Psychoses 2014 Apr;8(1):21-7A [FREE Full text] [doi: 10.3371/CSRP.KAFE.021513] [Medline: 23428781]
3. Firth J, Cotter J, Torous J, Bucci S, Firth JA, Yung AR. Mobile phone ownership and endorsement of "mHealth" among people with psychosis: a meta-analysis of cross-sectional studies. Schizophr Bull 2016 Mar;42(2):448-455 [FREE Full text] [doi: 10.1093/schbul/sbv132] [Medline: 26400871]
4. Naslund JA, Aschbrenner KA, Bartels SJ. How people with serious mental illness use smartphones, mobile apps, and social media. Psychiatr Rehabil J 2016 Dec;39(4):364-367 [FREE Full text] [doi: 10.1037/prj0000207] [Medline: 27845533]
5. Ruiz-Martínez I, Esparcia J. Internet access in rural areas: brake or stimulus as post-COVID-19 opportunity? Sustainability 2020 Nov 18;12(22):9619. [doi: 10.3390/su12229619]
6. Graham AK, Weissman RS, Mohr DC. Resolving key barriers to advancing mental health equity in rural communities using digital mental health interventions. JAMA Health Forum 2021 Jun 04;2(6):e211149 [FREE Full text] [doi: 10.1001/jamahealthforum.2021.1149] [Medline: 36218752]
7. Schueller SM, Hunter JF, Figueroa C, Aguilera A. Use of digital mental health for marginalized and underserved populations. Curr Treat Options Psych 2019 Jul 5;6(3):243-255. [doi: 10.1007/s40501-019-00181-z]
8. Lal S, Malla A. Service engagement in first-episode psychosis: current issues and future directions. Can J Psychiatry 2015 Aug;60(8):341-345 [FREE Full text] [doi: 10.1177/070674371506000802] [Medline: 26454555]
9. Sreejith G, Menon V. Mobile phones as a medium of mental health care service delivery: perspectives and barriers among patients with severe mental illness. Indian J Psychol Med 2019 Oct 01;41(5):428-433. [doi: 10.4103/ijpsym.ijpsym_333_18]
10. Torous J, Jän Myrick K, Rauseo-Ricupero N, Firth J. Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow. JMIR Ment Health 2020 Mar 26;7(3):e18848 [FREE Full text] [doi: 10.2196/18848] [Medline: 32213476]
11. Gratzer D, Torous J, Lam RW, Patten SB, Kutcher S, Chan S, et al. Our digital moment: innovations and opportunities in digital mental health care. Can J Psychiatry 2021 Jan 30;66(1):5-8 [FREE Full text] [doi: 10.1177/0706743720937833] [Medline: 32603188]
12. Wind TR, Rijkeboer M, Andersson G, Riper H. The COVID-19 pandemic: the 'Black Swan' for mental health care and a turning point for e-health. Internet Interv 2020 Apr;20:100317 [FREE Full text] [doi: 10.1016/j.invent.2020.100317] [Medline: 32289019]
13. Kozloff N, Mulsant BH, Stergiopoulos V, Voineskos AN. The COVID-19 global pandemic: implications for people with schizophrenia and related disorders. Schizophr Bull 2020 Jul 08;46(4):752-757 [FREE Full text] [doi: 10.1093/schbul/sbaa051] [Medline: 32343342]
14. Marshall JM, Dunstan DA, Bartik W. The digital psychiatrist: in search of evidence-based apps for anxiety and depression. Front Psychiatry 2019 Nov 15;10:831 [FREE Full text] [doi: 10.3389/fpsyt.2019.00831] [Medline: 31803083]
15. Lagan S, Aquino P, Emerson MR, Fortuna K, Walker R, Torous J. Actionable health app evaluation: translating expert frameworks into objective metrics. NPJ Digit Med 2020;3:100 [FREE Full text] [doi: 10.1038/s41746-020-00312-4] [Medline: 32821855]
16. Firth J, Torous J. Smartphone apps for schizophrenia: a systematic review. JMIR Mhealth Uhealth 2015 Nov 06;3(4):e102 [FREE Full text] [doi: 10.2196/mhealth.4930] [Medline: 26546039]
17. D'Arcey J, Collaton J, Kozloff N, Voineskos AN, Kidd SA, Foussias G. The use of text messaging to improve clinical engagement for individuals with psychosis: systematic review. JMIR Ment Health 2020 Apr 02;7(4):e16993 [FREE Full text] [doi: 10.2196/16993] [Medline: 32238334]
18. Furukawa TA, Bighelli I. Digital aids for relapse prevention in schizophrenia. Lancet Psychiatry 2022 Jun;9(6):424-425. [doi: 10.1016/s2215-0366(22)00145-6]
19. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005 Feb;8(1):19-32. [doi: 10.1080/1364557032000119616]
20. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5(1):69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
21. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ 2015 Jan 02;350(jan02 1):g7647 [FREE Full text] [doi: 10.1136/bmj.g7647] [Medline: 25555855]
22. O'Brien KK, Colquhoun H, Levac D, Baxter L, Tricco AC, Straus S, et al. Advancing scoping study methodology: a web-based survey and consultation of perceptions on terminology, definition and methodological steps. BMC Health Serv Res 2016 Dec 26;16:305 [FREE Full text] [doi: 10.1186/s12913-016-1579-z] [Medline: 27461419]
23. EndNote. Version 20. EndNote. URL: https://endnote.com/ [accessed 2024-04-29]
24. The world's #1 systematic review tool. Covidence. URL: https://www.covidence.org/ [accessed 2024-04-29]
25. Wentzel J, van der Vaart R, Bohlmeijer ET, van Gemert-Pijnen JE. Mixing online and face-to-face therapy: how to benefit from blended care in mental health care. JMIR Ment Health 2016;3(1):e9 [FREE Full text] [doi: 10.2196/mental.4534] [Medline: 26860537]

26. Beebe L, Smith KD, Phillips C. A comparison of telephone and texting interventions for persons with schizophrenia spectrum disorders. Issues Ment Health Nurs 2014 May 25;35(5):323-329. [doi: 10.3109/01612840.2013.863412] [Medline: 24766166]
27. Cullen BA, Rodriguez K, Eaton WW, Mojtabai R, Von Mach T, Ybarra ML. Clinical outcomes from the texting for relapse prevention (T4RP) in schizophrenia and schizoaffective disorder study. Psychiatry Res 2020 Oct;292:113346 [FREE Full text] [doi: 10.1016/j.psychres.2020.113346] [Medline: 32750572]
28. Fisher M, Etter K, Murray A, Ghiasi N, LaCross K, Ramsay I, et al. The effects of remote cognitive training combined with a mobile app intervention on psychosis: double-blind randomized controlled trial. J Med Internet Res 2023 Nov 13;25:e48634 [FREE Full text] [doi: 10.2196/48634] [Medline: 37955951]
29. Depp CA, Perivoliotis D, Holden J, Dorr J, Granholm EL. Single-session mobile-augmented intervention in serious mental illness: a three-arm randomized controlled trial. Schizophr Bull 2019 Jun 18;45(4):752-762 [FREE Full text] [doi: 10.1093/schbul/sby135] [Medline: 30281086]
30. Zhu X, Li M, Liu P, Chang R, Wang Q, Liu J. A mobile health application-based strategy for enhancing adherence to antipsychotic medication in schizophrenia. Arch Psychiatr Nurs 2020 Dec;34(6):472-480. [doi: 10.1016/j.apnu.2020.08.001] [Medline: 33280669]
31. Xu D, Xiao S, He H, Caine ED, Gloyd S, Simoni J, et al. Lay health supporters aided by mobile text messaging to improve adherence, symptoms, and functioning among people with schizophrenia in a resource-poor community in rural China (LEAN): a randomized controlled trial. PLoS Med 2019 Apr 23;16(4):e1002785 [FREE Full text] [doi: 10.1371/journal.pmed.1002785] [Medline: 31013275]
32. Schlosser DA, Campellone TR, Truong B, Etter K, Vergani S, Komaiko K, et al. Efficacy of PRIME, a mobile app intervention designed to improve motivation in young people with schizophrenia. Schizophr Bull 2018 Aug 20;44(5):1010-1020 [FREE Full text] [doi: 10.1093/schbul/sby078] [Medline: 29939367]
33. Garety P, Ward T, Emsley R, Greenwood K, Freeman D, Fowler D, et al. Digitally supported CBT to reduce paranoia and improve reasoning for people with schizophrenia-spectrum psychosis: the SlowMo RCT. Efficacy Mech Eval 2021 Aug;8(11):1-90. [doi: 10.3310/eme08110] [Medline: 34398537]
34. Krzystanek M, Borkowski M, Skałacka K, Krysta K. A telemedicine platform to improve clinical parameters in paranoid schizophrenia patients: results of a one-year randomized study. Schizophr Res 2019 Mar;204:389-396. [doi: 10.1016/j.schres.2018.08.016] [Medline: 30154027]
35. Schulze LN, Stentzel U, Leipert J, Schulte J, Langosch J, Freyberger HJ, et al. Improving medication adherence with telemedicine for adults with severe mental illness. Psychiatr Serv 2019 Mar 01;70(3):225-228. [doi: 10.1176/appi.ps.201800286] [Medline: 30651059]
36. Stentzel U, van den Berg N, Moon K, Schulze LN, Schulte J, Langosch JM, et al. Telemedical care and quality of life in patients with schizophrenia and bipolar disorder: results of a randomized controlled trial. BMC Psychiatry 2021 Jun 29;21(1):318 [FREE Full text] [doi: 10.1186/s12888-021-03318-8] [Medline: 34187420]
37. Montes JM, Medina E, Gomez-Beneyto M, Maurino J. A short message service (SMS)-based strategy for enhancing adherence to antipsychotic medication in schizophrenia. Psychiatry Res 2012 Dec 30;200(2-3):89-95. [doi: 10.1016/j.psychres.2012.07.034] [Medline: 22901437]
38. Pijnenborg GH, Withaar FK, Brouwer WH, Timmerman ME, van den Bosch RJ, Evans JJ. The efficacy of SMS text messages to compensate for the effects of cognitive impairments in schizophrenia. British J Clinic Psychol 2010 Dec 24;49(2):259-274. [doi: 10.1348/014466509x467828]
39. Välimäki M, Kannisto KA, Vahlberg T, Hätönen H, Adams CE. Short text messages to encourage adherence to medication and follow-up for people with psychosis (Mobile.Net): randomized controlled trial in Finland. J Med Internet Res 2017 Jul 12;19(7):e245 [FREE Full text] [doi: 10.2196/jmir.7028] [Medline: 28701292]
40. Kaplan K, Salzer MS, Solomon P, Brusilovskiy E, Cousounis P. Internet peer support for individuals with psychiatric disabilities: a randomized controlled trial. Soc Sci Med 2011 Jan;72(1):54-62. [doi: 10.1016/j.socscimed.2010.09.037] [Medline: 21112682]
41. Westermann S, Rüegg N, Lüdtke T, Moritz S, Berger T. Internet-based self-help for psychosis: findings from a randomized controlled trial. J Consult Clin Psychol 2020 Oct;88(10):937-950. [doi: 10.1037/ccp0000602] [Medline: 32790453]
42. Dabit S, Quraishi S, Jordan J, Biagianti B. Improving social functioning in people with schizophrenia-spectrum disorders via mobile experimental interventions: results from the CLIMB pilot trial. Schizophr Res Cogn 2021 Dec;26:100211 [FREE Full text] [doi: 10.1016/j.scog.2021.100211] [Medline: 34381699]
43. Kidd SA, Feldcamp L, Adler A, Kaleis L, Wang W, Vichnevetski K, et al. Feasibility and outcomes of a multi-function mobile health approach for the schizophrenia spectrum: App4Independence (A4i). PLoS One 2019 Jul 15;14(7):e0219491 [FREE Full text] [doi: 10.1371/journal.pone.0219491] [Medline: 31306439]
44. Austin SF, Frøsig A, Buus N, Lincoln T, von Malachowski A, Schlier B, et al. Service user experiences of integrating a mobile solution (IMPACHS) into clinical treatment for psychosis. Qual Health Res 2021 Apr 23;31(5):942-954. [doi: 10.1177/1049732320986556] [Medline: 33491577]
45. Fortuna KL, DiMilia PR, Lohman MC, Bruce ML, Zubritsky CD, Halaby MR, et al. Feasibility, acceptability, and preliminary effectiveness of a peer-delivered and technology supported self-management intervention for older adults with serious

mental illness. Psychiatr Q 2018 Jun 26;89(2):293-305 [FREE Full text] [doi: 10.1007/s11126-017-9534-7] [Medline: 28948424]
46. Bonet L, Torous J, Arce D, Blanquer I, Sanjuán J. ReMindCare, an app for daily clinical practice in patients with first episode psychosis: a pragmatic real-world study protocol. Early Interv Psychiatry 2021 Feb;15(1):183-192 [FREE Full text] [doi: 10.1111/eip.12960] [Medline: 32253830]
47. Gumley A, Bradstreet S, Ainsworth J, Allan S, Alvarez-Jimenez M, Beattie L, et al. Early signs monitoring to prevent relapse in psychosis and promote well-being, engagement, and recovery: protocol for a feasibility cluster randomized controlled trial harnessing mobile phone technology blended with peer support. JMIR Res Protoc 2020 Jan 09;9(1):e15058 [FREE Full text] [doi: 10.2196/15058] [Medline: 31917372]
48. Rüegg N, Moritz S, Berger T, Lüdtke T, Westermann S. An internet-based intervention for people with psychosis (EviBaS): study protocol for a randomized controlled trial. BMC Psychiatry 2018 Apr 13;18(1):102 [FREE Full text] [doi: 10.1186/s12888-018-1644-8] [Medline: 29653532]
49. Niendam TA, Tully LM, Iosif AM, Kumar D, Nye KE, Denton JC, et al. Enhancing early psychosis treatment using smartphone technology: a longitudinal feasibility and validity study. J Psychiatr Res 2018 Dec;96:239-246. [doi: 10.1016/j.jpsychires.2017.10.017] [Medline: 29126059]
50. Bell IH, Fielding-Smith SF, Hayward M, Rossell SL, Lim MH, Farhall J, et al. Smartphone-based ecological momentary assessment and intervention in a coping-focused intervention for hearing voices (SAVVy): study protocol for a pilot randomised controlled trial. Trials 2018 May 02;19(1):262 [FREE Full text] [doi: 10.1186/s13063-018-2607-6] [Medline: 29720208]
51. Kreyenbuhl J, Record EJ, Himelhoch S, Charlotte M, Palmer-Bacon J, Dixon LB, et al. Development and feasibility testing of a smartphone intervention to improve adherence to antipsychotic medications. Clin Schizophr Relat Psychoses 2019 Dec;12(4):152-167. [doi: 10.3371/csrp.krre.070816]
52. Byrne S, Kotze B, Ramos F, Casties A, Starling J, Harris A. Integrating a mobile health device into a community youth mental health team to manage severe mental illness: protocol for a randomized controlled trial. JMIR Res Protoc 2020 Nov 02;9(11):e19510 [FREE Full text] [doi: 10.2196/19510] [Medline: 33136053]
53. Schlosser D, Campellone T, Kim D, Truong B, Vergani S, Ward C, et al. Feasibility of PRIME: a cognitive neuroscience-informed mobile app intervention to enhance motivated behavior and improve quality of life in recent onset schizophrenia. JMIR Res Protoc 2016 Apr 28;5(2):e77 [FREE Full text] [doi: 10.2196/resprot.5450] [Medline: 27125771]
54. Granholm E, Holden J, Dwyer K, Mikhael T, Link P, Depp C. Mobile-assisted cognitive behavioral therapy for negative symptoms: open single-arm trial with schizophrenia patients. JMIR Ment Health 2020 Dec 01;7(12):e24406 [FREE Full text] [doi: 10.2196/24406] [Medline: 33258792]
55. Arnold C, Villagonzalo KA, Meyer D, Farhall J, Foley F, Kyrios M, et al. Predicting engagement with an online psychosocial intervention for psychosis: exploring individual- and intervention-level predictors. Internet Interv 2019 Dec;18:100266 [FREE Full text] [doi: 10.1016/j.invent.2019.100266] [Medline: 31890619]
56. Stentzel U, Grabe HJ, Strobel L, Penndorf P, Langosch J, Freyberger HJ, et al. Tecla: a telephone- and text-message based telemedical concept for patients with severe mental health disorders--study protocol for a controlled, randomized, study. BMC Psychiatry 2015 Nov 04;15(1):273 [FREE Full text] [doi: 10.1186/s12888-015-0659-7] [Medline: 26537570]
57. Hardy A, Wojdecka A, West J, Matthews E, Golby C, Ward T, et al. How inclusive, user-centered design research can improve psychological therapies for psychosis: development of SlowMo. JMIR Ment Health 2018 Dec 05;5(4):e11222 [FREE Full text] [doi: 10.2196/11222] [Medline: 30518514]
58. Ben-Zeev D, Kaiser SM, Brenner CJ, Begale M, Duffecy J, Mohr DC. Development and usability testing of FOCUS: a smartphone system for self-management of schizophrenia. Psychiatr Rehabil J 2013 Dec;36(4):289-296 [FREE Full text] [doi: 10.1037/prj0000019] [Medline: 24015913]
59. Fortuna KL, Myers AL, Ferron J, Kadakia A, Bianco C, Bruce ML, et al. Assessing a digital peer support self-management intervention for adults with serious mental illness: feasibility, acceptability, and preliminary effectiveness. J Ment Health 2022 Dec 28;31(6):833-841 [FREE Full text] [doi: 10.1080/09638237.2021.2022619] [Medline: 35088619]
60. Gumley AI, Bradstreet S, Ainsworth J, Allan S, Alvarez-Jimenez M, Birchwood M, et al. Digital smartphone intervention to recognise and manage early warning signs in schizophrenia to prevent relapse: the EMPOWER feasibility cluster RCT. Health Technol Assess 2022 May;26(27):1-174. [doi: 10.3310/hlze0479]
61. Ben-Zeev D, Brian RM, Aschbrenner KA, Jonathan G, Steingard S. Video-based mobile health interventions for people with schizophrenia: bringing the "pocket therapist" to life. Psychiatr Rehabil J 2018 Mar;41(1):39-45. [doi: 10.1037/prj0000197] [Medline: 27295133]
62. Ben-Zeev D, Brenner CJ, Begale M, Duffecy J, Mohr DC, Mueser KT. Feasibility, acceptability, and preliminary efficacy of a smartphone intervention for schizophrenia. Schizophr Bull 2014 Nov;40(6):1244-1253. [doi: 10.1093/schbul/sbu033] [Medline: 24609454]
63. Ben-Zeev D, Scherer EA, Gottlieb JD, Rotondi AJ, Brunette MF, Achtyes ED, et al. mHealth for schizophrenia: patient engagement with a mobile phone intervention following hospital discharge. JMIR Ment Health 2016 Jul 27;3(3):e34 [FREE Full text] [doi: 10.2196/mental.6348] [Medline: 27465803]

64. Ben-Zeev D, Brian RM, Jonathan G, Razzano L, Pashka N, Carpenter-Song E, et al. Mobile health (mHealth) versus clinic-based group intervention for people with serious mental illness: a randomized controlled trial. Psychiatr Serv 2018 Sep 01;69(9):978-985. [doi: 10.1176/appi.ps.201800063] [Medline: 29793397]
65. Buck B, Nguyen J, Porter S, Ben-Zeev D, Reger GM. FOCUS mHealth intervention for veterans with serious mental illness in an outpatient department of veterans affairs setting: feasibility, acceptability, and usability study. JMIR Ment Health 2022 Jan 28;9(1):e26049 [FREE Full text] [doi: 10.2196/26049] [Medline: 35089151]
66. Kidd S, McKenzie K, Wang W, Agrawal S, Voineskos A. Examining a digital health approach for advancing schizophrenia illness self-management and provider engagement: protocol for a feasibility trial. JMIR Res Protoc 2021 Jan 25;10(1):e24736 [FREE Full text] [doi: 10.2196/24736] [Medline: 33492235]
67. Garety PA, Ward T, Freeman D, Fowler D, Emsley R, Dunn G, et al. SlowMo, a digital therapy targeting reasoning in paranoia, versus treatment as usual in the treatment of people who fear harm from others: study protocol for a randomised controlled trial. Trials 2017 Nov 02;18(1):510 [FREE Full text] [doi: 10.1186/s13063-017-2242-7] [Medline: 29096681]
68. Price GD, Heinz MV, Nemesure MD, McFadden J, Jacobson NC. Predicting symptom response and engagement in a digital intervention among individuals with schizophrenia and related psychoses. Front Psychiatry 2022 Aug 11;13:807116 [FREE Full text] [doi: 10.3389/fpsyt.2022.807116] [Medline: 36032242]
69. Allan S, Beedie S, McLeod HJ, Farhall J, Gleeson J, Bradstreet S, et al. Using EMPOWER in daily life: a qualitative investigation of implementation experiences. BMC Psychiatry 2023 Aug 17;23(1):597 [FREE Full text] [doi: 10.1186/s12888-023-05096-x] [Medline: 37592231]
70. Krzystanek M, Krysta K, Borkowski M, Skałacka K, Przybyło J, Pałasz A, et al. The effect of smartphone-based cognitive training on the functional/cognitive markers of schizophrenia: a one-year randomized study. J Clin Med 2020 Nov 16;9(11):3681 [FREE Full text] [doi: 10.3390/jcm9113681] [Medline: 33207811]
71. Arnold C, Williams A, Thomas N. Engaging with a web-based psychosocial intervention for psychosis: qualitative study of user experiences. JMIR Ment Health 2020 Jun 19;7(6):e16730 [FREE Full text] [doi: 10.2196/16730] [Medline: 32558659]
72. Jonathan G, Carpenter-Song EA, Brian RM, Ben-Zeev D. Life with FOCUS: a qualitative evaluation of the impact of a smartphone intervention on people with serious mental illness. Psychiatr Rehabil J 2019 Jun;42(2):182-189. [doi: 10.1037/prj0000337] [Medline: 30589278]
73. Ben-Zeev D, Razzano LA, Pashka NJ, Levin CE. Cost of mHealth versus clinic-based care for serious mental illness: same effects, half the price tag. Psychiatr Serv 2021 Apr 01;72(4):448-451. [doi: 10.1176/appi.ps.202000349] [Medline: 33557599]
74. Lüdtke T, Platow-Kohlschein H, Rüegg N, Berger T, Moritz S, Westermann S. Mindfulness mediates the effect of a psychological online intervention for psychosis on self-reported hallucinations: a secondary analysis of voice hearers from the EviBaS trial. Front Psychiatry 2020 Apr 3;11:228 [FREE Full text] [doi: 10.3389/fpsyt.2020.00228] [Medline: 32308631]
75. Ben-Zeev D, Buck B, Vu Chu P, Razzano L, Pashka N, Hallgren KA. Transdiagnostic mobile health: smartphone intervention reduces depressive symptoms in people with mood and psychotic disorders. JMIR Ment Health 2019 Apr 12;6(4):e13202 [FREE Full text] [doi: 10.2196/13202] [Medline: 30977736]
76. Simões de Almeida RF, Sousa TJ, Couto AS, Marques AJ, Queiros CM, Martins CL. Development of weCope, a mobile app for illness self-management in schizophrenia. Arch Clin Psychiatry 2019 Feb;46(1):1-4 [FREE Full text] [doi: 10.1590/0101-60830000000182]
77. Fulford D, Mote J, Gard DE, Mueser KT, Gill K, Leung L, et al. Development of the Motivation and Skills Support (MASS) social goal attainment smartphone app for (and with) people with schizophrenia. J Behav Cogn Ther 2020 Apr;30(1):23-32 [FREE Full text] [doi: 10.1016/j.jbct.2020.03.016] [Medline: 33437970]
78. Bucci S, Barrowclough C, Ainsworth J, Machin M, Morris R, Berry K, et al. Actissist: proof-of-concept trial of a theory-driven digital intervention for psychosis. Schizophr Bull 2018 Aug 20;44(5):1070-1080 [FREE Full text] [doi: 10.1093/schbul/sby032] [Medline: 29566206]
79. Fulford D, Gard DE, Mueser KT, Mote J, Gill K, Leung L, et al. Preliminary outcomes of an ecological momentary intervention for social functioning in schizophrenia: pre-post study of the motivation and skills support app. JMIR Ment Health 2021 Jun 15;8(6):e27475 [FREE Full text] [doi: 10.2196/27475] [Medline: 34128812]
80. Hanssen E, Balvert S, Oorschot M, Borkelmans K, van Os J, Delespaul P, et al. An ecological momentary intervention incorporating personalised feedback to improve symptoms and social functioning in schizophrenia spectrum disorders. Psychiatry Res 2020 Feb;284:112695 [FREE Full text] [doi: 10.1016/j.psychres.2019.112695] [Medline: 31831201]
81. Kim SW, Lee GY, Yu HY, Jung EI, Lee JY, Kim SY, et al. Development and feasibility of smartphone application for cognitive-behavioural case management of individuals with early psychosis. Early Interv Psychiatry 2018 Dec;12(6):1087-1093. [doi: 10.1111/eip.12418] [Medline: 28516480]
82. Lim MH, Gleeson JF, Rodebaugh TL, Eres R, Long KM, Casey K, et al. A pilot digital intervention targeting loneliness in young people with psychosis. Soc Psychiatry Psychiatr Epidemiol 2020 Jul 14;55(7):877-889. [doi: 10.1007/s00127-019-01681-2] [Medline: 30874828]
83. Moitra E, Park HS, Gaudiano BA. Development and initial testing of an mHealth transitions of care intervention for adults with schizophrenia-spectrum disorders immediately following a psychiatric hospitalization. Psychiatr Q 2021 Mar;92(1):259-272 [FREE Full text] [doi: 10.1007/s11126-020-09792-9] [Medline: 32613525]

84. Bell IH, Rossell SL, Farhall J, Hayward M, Lim MH, Fielding-Smith SF, et al. Pilot randomised controlled trial of a brief coping-focused intervention for hearing voices blended with smartphone-based ecological momentary assessment and intervention (SAVVy): feasibility, acceptability and preliminary clinical outcomes. Schizophr Res 2020 Feb;216:479-487. [doi: 10.1016/j.schres.2019.10.026] [Medline: 31812327]
85. Han M, Lee K, Kim M, Heo Y, Choi H. Effects of a metacognitive smartphone intervention with weekly mentoring sessions for individuals with schizophrenia: a quasi-experimental study. J Psychosoc Nurs Ment Health Serv 2023 Feb;61(2):27-37 [FREE Full text] [doi: 10.3928/02793695-20220706-01] [Medline: 35858205]
86. Bucci S, Barrowclough C, Ainsworth J, Morris R, Berry K, Machin M, et al. Using mobile technology to deliver a cognitive behaviour therapy-informed intervention in early psychosis (Actissist): study protocol for a randomised controlled trial. Trials 2015 Sep 10;16(1):404 [FREE Full text] [doi: 10.1186/s13063-015-0943-3] [Medline: 26357943]
87. Steare T, O'Hanlon P, Eskinazi M, Osborn D, Lloyd-Evans B, Jones R, et al. App to support recovery in early intervention services (ARIES) study: protocol of a feasibility randomised controlled trial of a self-management Smartphone application for psychosis. BMJ Open 2019 Mar 20;9(3):e025823 [FREE Full text] [doi: 10.1136/bmjopen-2018-025823] [Medline: 30898825]
88. Barbeito S, Sánchez-Gutiérrez T, Mayoral M, Moreno M, Ríos-Aguilar S, Arango C, et al. Mobile app-based intervention for adolescents with first-episode psychosis: study protocol for a pilot randomized controlled trial. Front Psychiatry 2019 Feb 5;10:27 [FREE Full text] [doi: 10.3389/fpsyt.2019.00027] [Medline: 30804818]
89. Husain N, Gire N, Kelly J, Duxbury J, McKeown M, Riley M, et al. TechCare: mobile assessment and therapy for psychosis - an intervention for clients in the Early Intervention Service: A feasibility study protocol. SAGE Open Med 2016 Oct 18;4:2050312116669613 [FREE Full text] [doi: 10.1177/2050312116669613] [Medline: 27790373]
90. Jongeneel A, Scheffers D, Tromp N, Nuij C, Delespaul P, Riper H, et al. Reducing distress and improving social functioning in daily life in people with auditory verbal hallucinations: study protocol for the 'Temstem' randomised controlled trial. BMJ Open 2018 Mar 06;8(3):e020537 [FREE Full text] [doi: 10.1136/bmjopen-2017-020537] [Medline: 29511020]
91. Steare T, O'Hanlon P, Eskinazi M, Osborn D, Lloyd-Evans B, Jones R, et al. Smartphone-delivered self-management for first-episode psychosis: the ARIES feasibility randomised controlled trial. BMJ Open 2020 Aug 26;10(8):e034927 [FREE Full text] [doi: 10.1136/bmjopen-2019-034927] [Medline: 32847902]
92. van Aubel E, Bakker JM, Batink T, Michielse S, Goossens L, Lange I, et al. Blended care in the treatment of subthreshold symptoms of depression and psychosis in emerging adults: a randomised controlled trial of acceptance and commitment therapy in daily-life (ACT-DL). Behav Res Ther 2020 May;128:103592. [doi: 10.1016/j.brat.2020.103592] [Medline: 32146218]
93. Eisner E, Berry N, Morris R, Emsley R, Haddock G, Machin M, et al. Exploring engagement with the CBT-informed actissist smartphone application for early psychosis. J Ment Health 2023 Jun 27;32(3):643-654. [doi: 10.1080/09638237.2023.2182429] [Medline: 36850040]
94. Hassan L, Eisner E, Berry K, Emsley R, Ainsworth J, Lewis S, et al. User engagement in a randomised controlled trial for a digital health intervention for early psychosis (Actissist 2.0 trial). Psychiatry Res 2023 Nov;329:115536 [FREE Full text] [doi: 10.1016/j.psychres.2023.115536] [Medline: 37857132]
95. Steare T, Giorgalli M, Free K, Harju-Seppänen J, Akther S, Eskinazi M, et al. A qualitative study of stakeholder views on the use of a digital app for supported self-management in early intervention services for psychosis. BMC Psychiatry 2021 Jun 19;21(1):311 [FREE Full text] [doi: 10.1186/s12888-021-03317-9] [Medline: 34147075]
96. Moore E, Williams A, Bell I, Thomas N. Client experiences of blending a coping-focused therapy for auditory verbal hallucinations with smartphone-based ecological momentary assessment and intervention. Internet Interv 2020 Mar;19:100299 [FREE Full text] [doi: 10.1016/j.invent.2019.100299] [Medline: 31890641]
97. Fulford D, Schupbach E, Gard DE, Mueser KT, Mow J, Leung L. Do cognitive impairments limit treatment gains in a standalone digital intervention for psychosis? A test of the digital divide. Schizophr Res Cogn 2022 Jun;28:100244 [FREE Full text] [doi: 10.1016/j.scog.2022.100244] [Medline: 35242612]
98. de Almeida RS, Couto A, Marques A, Queirós C, Martins C. Mobile application for self-management in schizophrenia: a pilot study. J Technol Hum Serv 2018 Jul 31;36(4):179-190. [doi: 10.1080/15228835.2018.1483859]
99. Ybarra ML, Rodriguez K, Madison H, Mojtabai R, Cullen BA. Developing texting for relapse prevention: a scalable mHealth program for people with schizophrenia and schizoaffective disorder. J Nerv Ment Dis 2019 Oct;207(10):854-862 [FREE Full text] [doi: 10.1097/NMD.0000000000001037] [Medline: 31503175]
100. Polillo A, Foussias G, Wong AH, Ampofo A, Stergiopoulos V, Anderson KK, et al. ED to EPI: protocol for a pragmatic randomised controlled trial of an SMS (text) messaging intervention to improve the transition from the emergency department to early psychosis intervention for young people with psychosis. BMJ Open 2020 Dec 17;10(12):e042751 [FREE Full text] [doi: 10.1136/bmjopen-2020-042751] [Medline: 33334839]
101. Ben-Zeev D, Kaiser SM, Krzos I. Remote "hovering" with individuals with psychotic disorders and substance use: feasibility, engagement, and therapeutic alliance with a text-messaging mobile interventionist. J Dual Diagn 2014 Nov 03;10(4):197-203 [FREE Full text] [doi: 10.1080/15504263.2014.962336] [Medline: 25391277]

102. Granholm E, Ben-Zeev D, Link PC, Bradshaw KR, Holden JL. Mobile Assessment and Treatment for Schizophrenia (MATS): a pilot trial of an interactive text-messaging intervention for medication adherence, socialization, and auditory hallucinations. Schizophr Bull 2012 May;38(3):414-425 [FREE Full text] [doi: 10.1093/schbul/sbr155] [Medline: 22080492]
103. Ben-Zeev D, Buck B, Meller S, Hudenko WJ, Hallgren KA. Augmenting evidence-based care with a texting mobile interventionist: a pilot randomized controlled trial. Psychiatr Serv 2020 Dec 01;71(12):1218-1224 [FREE Full text] [doi: 10.1176/appi.ps.202000239] [Medline: 32631130]
104. Luther L, Fischer MW, Johnson-Kwochka AV, Minor KS, Holden R, Lapish CL, et al. Mobile enhancement of motivation in schizophrenia: a pilot randomized controlled trial of a personalized text message intervention for motivation deficits. J Consult Clin Psychol 2020 Oct;88(10):923-936 [FREE Full text] [doi: 10.1037/ccp0000599] [Medline: 32790451]
105. Cai Y, Gong W, He H, Hughes JP, Simoni J, Xiao S, et al. Mobile texting and lay health supporters to improve schizophrenia care in a resource-poor community in rural China (LEAN trial): randomized controlled trial extended implementation. J Med Internet Res 2020 Dec 01;22(12):e22631 [FREE Full text] [doi: 10.2196/22631] [Medline: 33258788]
106. Aschbrenner KA, Naslund JA, Gill LE, Bartels SJ, Ben-Zeev D. A qualitative study of client-clinician text exchanges in a mobile health intervention for individuals with psychotic disorders and substance use. J Dual Diagn 2016 Jan 30;12(1):63-71 [FREE Full text] [doi: 10.1080/15504263.2016.1145312] [Medline: 26829356]
107. Ybarra ML, Rodriguez KM, Fehmie DA, Mojtabai R, Cullen B. Acceptability of texting 4 relapse prevention, text messaging-based relapse prevention program for people with schizophrenia and schizoaffective disorder. J Nerv Ment Dis 2022 Feb 01;210(2):123-128. [doi: 10.1097/NMD.0000000000001421] [Medline: 34570061]
108. Yu Y, Li Y, Li T, Xi S, Xiao X, Xiao S, et al. New path to recovery and well-being: cross-sectional study on WeChat use and endorsement of WeChat-based mHealth among people living with schizophrenia in China. J Med Internet Res 2020 Sep 18;22(9):e18663 [FREE Full text] [doi: 10.2196/18663] [Medline: 32945774]
109. Ludwig KA, Browne JW, Nagendra A, Gleeson JF, D'Alfonso S, Penn DL, et al. Horyzons USA: a moderated online social intervention for first episode psychosis. Early Interv Psychiatry 2021 Apr 17;15(2):335-343. [doi: 10.1111/eip.12947] [Medline: 32067415]
110. Thomas N, Farhall J, Foley F, Leitan ND, Villagonzalo K, Ladd E, et al. Promoting personal recovery in people with persisting psychotic disorders: development and pilot study of a novel digital intervention. Front Psychiatry 2016 Dec;7:196 [FREE Full text] [doi: 10.3389/fpsyt.2016.00196] [Medline: 28066271]
111. Hargreaves A, Dillon R, Anderson-Schmidt H, Corvin A, Fitzmaurice B, Castorina M, et al. Computerised working-memory focused cognitive remediation therapy for psychosis--a preliminary study. Schizophr Res 2015 Dec;169(1-3):135-140. [doi: 10.1016/j.schres.2015.09.004] [Medline: 26421692]
112. Moura BM, Avila A, Chendo I, Frade P, Barandas R, Vian J, et al. Facilitating the delivery of cognitive remediation in first-episode psychosis: pilot study of a home-delivered web-based intervention. J Nerv Ment Dis 2019 Nov;207(11):951-957. [doi: 10.1097/NMD.0000000000001055] [Medline: 31503184]
113. Alvarez-Jimenez M, Bendall S, Koval P, Rice S, Cagliarini D, Valentine L, et al. HORYZONS trial: protocol for a randomised controlled trial of a moderated online social therapy to maintain treatment effects from first-episode psychosis services. BMJ Open 2019 Feb 19;9(2):e024104 [FREE Full text] [doi: 10.1136/bmjopen-2018-024104] [Medline: 30782893]
114. Thomas N, Farhall J, Foley F, Rossell SL, Castle D, Ladd E, et al. Randomised controlled trial of a digitally assisted low intensity intervention to promote personal recovery in persisting psychosis: SMART-Therapy study protocol. BMC Psychiatry 2016 Sep 07;16(1):312 [FREE Full text] [doi: 10.1186/s12888-016-1024-1] [Medline: 27604363]
115. Valentine L, McEnery C, O'Sullivan S, D'Alfonso S, Gleeson J, Bendall S, et al. Young people's experience of online therapy for first-episode psychosis: a qualitative study. Psychol Psychother 2022 Mar 12;95(1):155-172. [doi: 10.1111/papt.12356] [Medline: 34252267]
116. Lüdtke T, Rüegg N, Moritz S, Berger T, Westermann S. Insight and the number of completed modules predict a reduction of positive symptoms in an internet-based intervention for people with psychosis. Psychiatry Res 2021 Dec;306:114223. [doi: 10.1016/j.psychres.2021.114223] [Medline: 34826711]
117. Gumley AI, Bradstreet S, Ainsworth J, Allan S, Alvarez-Jimenez M, Aucott L, et al. The EMPOWER blended digital intervention for relapse prevention in schizophrenia: a feasibility cluster randomised controlled trial in Scotland and Australia. Lancet Psychiatry 2022 Jun;9(6):477-486. [doi: 10.1016/s2215-0366(22)00103-1]
118. Xu DR, Gong W, Caine ED, Xiao S, Hughes JP, Ng M, et al. Lay health supporters aided by a mobile phone messaging system to improve care of villagers with schizophrenia in Liuyang, China: protocol for a randomised control trial. BMJ Open 2016 Jan 20;6(1):e010120 [FREE Full text] [doi: 10.1136/bmjopen-2015-010120] [Medline: 26792221]
119. Pijnenborg GH, Withaar FK, Evans JJ, van den Bosch RJ, Brouwer WH. SMS text messages as a prosthetic aid in the cognitive rehabilitation of schizophrenia. Rehabil Psychol 2007 May;52(2):236-240. [doi: 10.1037/0090-5550.52.2.236]
120. Biagianti B, Fisher M, Howard L, Rowlands A, Vinogradov S, Woolley J. Feasibility and preliminary efficacy of remotely delivering cognitive training to people with schizophrenia using tablets. Schizophr Res Cogn 2017 Dec;10:7-14 [FREE Full text] [doi: 10.1016/j.scog.2017.07.003] [Medline: 28824850]
121. Batra S, Baker R, Wang T, Forma F, DiBiasi F, Peters-Strickland T. Digital health technology for use in patients with serious mental illness: a systematic review of the literature. Med Devices (Auckl) 2017 Oct;10:237-251 [FREE Full text] [doi: 10.2147/MDER.S144158] [Medline: 29042823]

122. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018 Aug 05;21(3):116-119 [FREE Full text] [doi: 10.1136/eb-2018-102891] [Medline: 29871870]
123. Jacob C, Sezgin E, Sanchez-Vazquez A, Ivory C. Sociotechnical factors affecting patients' adoption of mobile health tools: systematic literature review and narrative synthesis. JMIR Mhealth Uhealth 2022 May 05;10(5):e36284 [FREE Full text] [doi: 10.2196/36284] [Medline: 35318189]
124. Wang T, Wang W, Liang J, Nuo M, Wen Q, Wei W, et al. Identifying major impact factors affecting the continuance intention of mHealth: a systematic review and multi-subgroup meta-analysis. NPJ Digit Med 2022 Sep 15;5(1):145 [FREE Full text] [doi: 10.1038/s41746-022-00692-9] [Medline: 36109594]
125. Nwosu A, Boardman S, Husain MM, Doraiswamy PM. Digital therapeutics for mental health: is attrition the Achilles heel? Front Psychiatry 2022 Aug 5;13:900615 [FREE Full text] [doi: 10.3389/fpsyt.2022.900615] [Medline: 35982936]
126. Wu A, Scult MA, Barnes ED, Betancourt JA, Falk A, Gunning FM. Smartphone apps for depression and anxiety: a systematic review and meta-analysis of techniques to increase engagement. NPJ Digit Med 2021 Feb 11;4(1):20 [FREE Full text] [doi: 10.1038/s41746-021-00386-8] [Medline: 33574573]
127. Gagnon M, Ngangue P, Payne-Gagnon J, Desmartis M. m-Health adoption by healthcare professionals: a systematic review. J Am Med Inform Assoc 2016 Jan;23(1):212-220. [doi: 10.1093/jamia/ocv052] [Medline: 26078410]
128. Taher R, Hsu CW, Hampshire C, Fialho C, Heaysman C, Stahl D, et al. The safety of digital mental health interventions: systematic review and recommendations. JMIR Ment Health 2023 Oct 09;10:e47433 [FREE Full text] [doi: 10.2196/47433] [Medline: 37812471]
129. Anmella G, Faurholt-Jepsen M, Hidalgo-Mazzei D, Radua J, Passos IC, Kapczinski F, et al. Smartphone-based interventions in bipolar disorder: systematic review and meta-analyses of efficacy. A position paper from the International Society for Bipolar Disorders (ISBD) big data task force. Bipolar Disord 2022 Sep 11;24(6):580-614 [FREE Full text] [doi: 10.1111/bdi.13243] [Medline: 35839276]
130. Zhong S, Yang X, Pan Z, Fan Y, Chen Y, Yu X, et al. The usability, feasibility, acceptability, and efficacy of digital mental health services in the COVID-19 pandemic: scoping review, systematic review, and meta-analysis. JMIR Public Health Surveill 2023 Feb 13;9:e43730 [FREE Full text] [doi: 10.2196/43730] [Medline: 36634261]
131. Morales-Pillado C, Fernández-Castilla B, Sánchez-Gutiérrez T, González-Fraile E, Barbeito S, Calvo A. Efficacy of technology-based interventions in psychosis: a systematic review and network meta-analysis. Psychol Med 2022 Dec 06:1-12. [doi: 10.1017/s0033291722003610]
132. Gan DZ, McGillivray L, Han J, Christensen H, Torok M. Effect of engagement with digital interventions on mental health outcomes: a systematic review and meta-analysis. Front Digit Health 2021;3:764079 [FREE Full text] [doi: 10.3389/fdgth.2021.764079] [Medline: 34806079]
133. Sawyer C, Hassan L, Guinart D, Agulleiro LM, Firth J. Smoking cessation apps for people with schizophrenia: how feasible are m-health approaches? Behav Sci (Basel) 2022 Aug 01;12(8):265 [FREE Full text] [doi: 10.3390/bs12080265] [Medline: 36004836]
134. Naslund JA, Aschbrenner KA. Technology use and interest in digital apps for mental health promotion and lifestyle intervention among young adults with serious mental illness. J Affect Disord Rep 2021 Dec;6:100227. [doi: 10.1016/j.jadr.2021.100227]
135. Sin J, Henderson C, Elkes J, Cornelius V, Woodham LA, Batchelor R, et al. Effect of digital psychoeducation and peer support on the mental health of family carers supporting individuals with psychosis in England (COPe-support): a randomised clinical trial. Lancet Digit Health 2022 May;4(5):e320-e329. [doi: 10.1016/s2589-7500(22)00031-0]

Abbreviations

AE: adverse event
CBT: cognitive behavioral therapy
EMA: ecological momentary assessment
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized control trial
SAE: serious adverse event
SMI: serious mental illness
SSD: schizophrenia spectrum disorder
TAU: treatment as usual

Edited by A Mavragani; submitted 08.02.24; peer-reviewed by C Sawyer, A Hassan, A Lustig; comments to author 28.05.24; revised version received 16.07.24; accepted 25.07.24; published 30.09.24.

Please cite as:
D'Arcey J, Torous J, Asuncion TR, Tackaberry-Giddens L, Zahid A, Ishak M, Foussias G, Kidd S
Leveraging Personal Technologies in the Treatment of Schizophrenia Spectrum Disorders: Scoping Review
JMIR Ment Health 2024;11:e57150
URL: https://mental.jmir.org/2024/1/e57150
doi:10.2196/57150
PMID:39348196

Review

Large Language Models for Mental Health Applications: Systematic Review

Zhijun Guo[1], MSc; Alvina Lai[1], DPhil; Johan H Thygesen[1], DPhil; Joseph Farrington[1], MSc; Thomas Keen[1,2], MMathPhil; Kezhi Li[1], DPhil

[1]Institute of Health Informatics University College, London, London, United Kingdom

[2]Great Ormond Street Institute of Child Health, University College London, London, United Kingdom

Corresponding Author:
Kezhi Li, DPhil
Institute of Health Informatics University College, London
222 Euston Road
London, NW1 2DA
United Kingdom
Phone: 44 7859 995590
Email: ken.li@ucl.ac.uk

Abstract

Background: Large language models (LLMs) are advanced artificial neural networks trained on extensive datasets to accurately understand and generate natural language. While they have received much attention and demonstrated potential in digital health, their application in mental health, particularly in clinical settings, has generated considerable debate.

Objective: This systematic review aims to critically assess the use of LLMs in mental health, specifically focusing on their applicability and efficacy in early screening, digital interventions, and clinical settings. By systematically collating and assessing the evidence from current studies, our work analyzes models, methodologies, data sources, and outcomes, thereby highlighting the potential of LLMs in mental health, the challenges they present, and the prospects for their clinical use.

Methods: Adhering to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, this review searched 5 open-access databases: MEDLINE (accessed by PubMed), IEEE Xplore, Scopus, JMIR, and ACM Digital Library. Keywords used were (*mental health* OR *mental illness* OR *mental disorder* OR *psychiatry*) AND (*large language models*). This study included articles published between January 1, 2017, and April 30, 2024, and excluded articles published in languages other than English.

Results: In total, 40 articles were evaluated, including 15 (38%) articles on mental health conditions and suicidal ideation detection through text analysis, 7 (18%) on the use of LLMs as mental health conversational agents, and 18 (45%) on other applications and evaluations of LLMs in mental health. LLMs show good effectiveness in detecting mental health issues and providing accessible, destigmatized eHealth services. However, assessments also indicate that the current risks associated with clinical use might surpass their benefits. These risks include inconsistencies in generated text; the production of hallucinations; and the absence of a comprehensive, benchmarked ethical framework.

Conclusions: This systematic review examines the clinical applications of LLMs in mental health, highlighting their potential and inherent risks. The study identifies several issues: the lack of multilingual datasets annotated by experts, concerns regarding the accuracy and reliability of generated content, challenges in interpretability due to the "black box" nature of LLMs, and ongoing ethical dilemmas. These ethical concerns include the absence of a clear, benchmarked ethical framework; data privacy issues; and the potential for overreliance on LLMs by both physicians and patients, which could compromise traditional medical practices. As a result, LLMs should not be considered substitutes for professional mental health services. However, the rapid development of LLMs underscores their potential as valuable clinical aids, emphasizing the need for continued research and development in this area.

Trial Registration: PROSPERO CRD42024508617; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=508617

(JMIR Ment Health 2024;11:e57400) doi:10.2196/57400

KEYWORDS

large language models; mental health; digital health care; ChatGPT; Bidirectional Encoder Representations from Transformers; BERT

Introduction

Mental Health

Mental health, a critical component of overall well-being, is at the forefront of global health challenges [1]. In 2019, an estimated 970 million individuals worldwide experienced mental illness, accounting for 12.5% of the global population [2]. Anxiety and depression are among the most prevalent psychological conditions, affecting 301 million and 280 million individuals, respectively [2]. In addition, 40 million people experienced bipolar disorder, 24 million experienced schizophrenia, and 14 million experienced eating disorders [3]. These mental disorders collectively contribute to an estimated US $5 trillion in global economic losses annually [4]. Despite the staggering prevalence, many cases remain undetected or untreated, with the resources allocated to the diagnosis and treatment of mental illness far less than the negative impact it has on society [5]. Globally, untreated mental illnesses affect 5% of the population in high-income countries and 19% of the population in low- and middle-income countries [3]. The COVID-19 pandemic has further exacerbated the challenges faced by mental health services worldwide [6], as the demand for these services increased while access was decreased [7]. This escalating crisis underscores the urgent need for more innovative and accessible mental health care approaches.

Mental illness treatment encompasses a range of modalities, including medication, psychotherapy, support groups, hospitalization, and complementary and alternative medicine [8]. However, the societal stigma attached to mental illnesses often deters people from seeking appropriate care [9]. Influenced by the fear of judgment and concerns about costly, ineffective treatments [10], many people with mental illness avoid or delay psychotherapy [11]. The COVID-19 crisis and other global pandemics have underscored the importance of digital tools, such as telemedicine and mobile apps, in delivering care during critical times [12]. In this evolving context, large language models (LLMs) present new possibilities for enhancing the delivery and effectiveness of mental health care.

Recent technological advancements have revealed some unique advantages of LLMs in mental health. These models, capable of processing and generating text akin to human communication, provide accessible support directly to users [13]. A study analyzing 2917 Reddit (Reddit, Inc) user reviews found that conversational agents (CAs) powered by LLMs are valued for their nonjudgmental listening and effective problem-solving advice. This aspect is particularly beneficial for individuals considered socially marginalized, as it enables them to be heard and understood without the need for direct social interaction [14]. Moreover, LLMs enhance the accessibility of mental health services, which are notably undersupplied globally [15]. Recent data reveals substantial delays in traditional mental health care delivery; 23% of individuals with mental illnesses report waiting for >12 weeks for face-to-face psychotherapy sessions [16], with 12% waiting for >6 months and 6% waiting for >1 year [16]. In addition, 43% of adults with mental illness indicate that such long waits have exacerbated their conditions [16].

Telemedicine, enhanced by LLMs, offers a practical alternative that expedites service delivery and could flatten traditional health care hierarchies [17]. This includes real-time counseling sessions through CAs that are not only cost-effective but also accessible anytime and from any location. By reducing the reliance on physical visits to traditional health care settings, telemedicine has the potential to decentralize access to medical expertise and diminish the hierarchical structures within the health care system [17]. Mental health chatbots developed using language models, such as Woebot [18] and Wysa [19], have been gaining recognition. Both chatbots follow the principles of cognitive behavioral therapy and are designed to equip users with self-help tools for managing their mental health issues [20]. In clinical practice, LLMs hold the potential to support the automatic assessment of therapists' adherence to evidence-based practices and the development of systems that offer real-time feedback and support for patient homework between sessions [21]. These models also have the potential to provide feedback on psychotherapy or peer support sessions, which is especially beneficial for clinicians with less training and experience [21]. Currently, these applications are still in the proposal stage. Although promising, they are not yet widely used in routine clinical settings, and further evaluation of their feasibility and effectiveness is necessary.

The deployment of LLMs in mental health also poses several risks, particularly for groups considered vulnerable. Challenges such as inconsistencies in the content generated and the production of "hallucinatory" content may mislead or harm users [22], raising serious ethical concerns. In response, authorities such as the World Health Organization have developed ethical guidelines for artificial intelligence (AI) research in health care, emphasizing the importance of data privacy; human oversight; and the principle that AI tools should augment, rather than replace, human practitioners [23]. These potential problems with LLMs in health care have gained considerable industry attention, underscoring the need for a comprehensive and responsible evaluation of LLMs' applications in mental health. The following section will further explore the workings of LLMs and their potential applications in mental health and critically evaluate the opportunities and challenges they introduce.

LLMs in Mental Health

LLMs represent advancements in machine learning, characterized by their ability to understand and generate human-like text with high accuracy [24]. The efficacy of these models is typically evaluated using benchmarks designed to assess their linguistic fidelity and contextual relevance. Common metrics include Bilingual Evaluation Understudy for translation accuracy and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) for summarization tasks [25]. LLMs are characterized by their scale, often encompassing billions of parameters, setting them apart from traditional language models [26]. This breakthrough is largely due to the transformer architecture, a deep neural network structure that uses a "self-attention" mechanism developed by Vaswani et al [27]. This allows LLMs to process information in parallel rather than sequentially, greatly enhancing speed and contextual understanding [27]. To clearly define the scope of this study concerning LLMs, we specify that an LLM must use the transformer architecture and contain a high number of parameters, traditionally at least 1 billion, to qualify as "large" [28]. This criterion encompasses models such as GPT (OpenAI) and Bidirectional Encoder Representations from Transformers (BERT; Google AI). Although the standard BERT model, with only 0.34 billion parameters [29], does not meet the traditional criteria for "large," its sophisticated bidirectional design and pivotal role in establishing new natural language processing (NLP) benchmarks justify its inclusion among notable LLMs [30]. The introduction of ChatGPT (OpenAI) in 2022 generated substantial public and academic interest in LLMs, underlining their transformative potential within the field of AI [31]. Other state-of-the-art LLMs include Large Language Model Meta AI (LLaMA; Meta AI) and Pathways Language Model (PaLM; Google AI), as illustrated in Table 1 [32-35].

Table 1. Comparative analysis of large language models (LLMs) by parameter size and developer entity. Data were summarized with the latest models up to June 2024, with data for parameters and developers from GPT (OpenAI) to Large Language Model Meta AI (LLaMA; Meta AI) adapted from the study by Thirunavukarasu et al [32].

Model name	Publication date	Parameters (billion)	Developer
Generative Pretrained Transformer (GPT)	June 2018	0.117	OpenAI
Bidirectional Encoder Representations from Transformers (BERT)	October 2018	0.34	Google
GPT-2	January 2019	1.5	OpenAI
Enhanced Representation through Knowledge Integration (ERNIE)	September 2019	0.114	Baidu
Conditional Transformer Language Model (CTRL)	September 2019	1.63	OpenAI
Megatron	September 2019	3.9	NVIDIA
Bidirectional and Auto-Regressive Transformers (BART)	October 2019	0.374	Meta
Turing Natural Language Generation (Turing-NLG)	January 2020	530	Microsoft
GPT-3	June 2020	175	OpenAI
Vision Transformer (ViT)	October 2020	0.632	Google
Inspired by artist Salvador Dalí and Pixar's WALL·E (DALL-E)	October 2020	1.2	OpenAI
Swin Transformer	March 2021	0.197	Microsoft
Wu Dao 2.0	June 2021	1750	Huawei
Jurassic-1	August 2021	178	AI21 Labs
Megatron-Turing Natural Language Generation (MT-NLG)	October 2021	530	Microsoft & Nvidia
Claude	December 2021	52	Anthropic
Generalist Language Model (GLAM)	December 2021	1200	Google
ERNIE 3.0	December 2021	260	Baidu
Guided Language-to-Image Diffusion for Generation and Editing (GLIDE)	December 2021	3.5	OpenAI
Gopher	December 2021	280	DeepMind
Causal Masked Modeling 3 (CM3)	January 2022	13	Meta
Language Model for Dialogue Applications (LaMDA)	January 2022	137	Google
GPT-NeoX	February 2022	20	EleutherAI
Chinchilla	March 2022	70	DeepMind
GopherCite	March 2022	280	DeepMind
DALL-E 2	April 2022	3.5	OpenAI
Flamingo	April 2022	80	DeepMind
Pathways Language Model (PaLM)	April 2022	540	Google
Gato	May 2022	1.2	DeepMind
Open Pretrained Transformer (OPT)	May 2022	175	Meta
Yet Another Language Model (YaLM)	June 2022	100	Yandex
Minerva	June 2022	540	Google
BigScience Large Open-science Open-access Multilingual Language Model (BLOOM)	July 2022	175	Hugging Face
Galactica	November 2022	120	Meta
Alexa Teacher Model (Alexa TM)	November 2022	20	Amazon
Large Language Model Meta AI (LLAMA)	February 2023	65	Meta
GPT-4	March 2023	1760	OpenAI
Cerebras-GPT	March 2023	13	Cerebras
Falcon	March 2023	40	Technology Innovation Institute

Model name	Publication date	Parameters (billion)	Developer
Bloomberg Generative Pretrained Transformer (BloombergGPT)	March 2023	50	Bloomberg
PanGu-2	March 2023	1085	Huawei
OpenAssistant	March 2023	17	LAION
PaLM 2	May 2023	340	Google
Llama 2	July 2023	70	Meta
Falcon 180B	September 2023	180	Technology Innovation Institute
Mistral 7B	September 2023	7.3	Mistral
Claude 2.1	November 2023	200	Anthropic
Grok-1	November 2023	314	xAI
Mixtral 8x7B	December 2023	46.7	Mistral
Phi-2	December 2023	2.7	EleutherAI
Gemma	February 2024	7	Google
DBRX	March 2024	136	Databricks
Llama 3	April 2024	70	Meta AI
Fugaku-LLM	May 2024	13	Fujitsu, Tokyo Institute of Technology, etc
Nemotron-4	June 2024	340	Nvidia

LLMs are primarily designed to learn fundamental statistical patterns of language [36]. Initially, these models were used as the basis for fine-tuning task-specific models rather than training those models from scratch, offering a more resource-efficient approach [37]. This fine-tuning process involves adjusting a pretrained model to a specific task by further training it on a smaller, task-specific dataset [38]. However, developments in larger and more sophisticated models have reduced the need for extensive fine-tuning in some cases. Notably, some advanced LLMs can now effectively understand and execute tasks specified through natural language prompts without extensive task-specific fine-tuning [39]. Instruction fine-tuned models undergo additional training on pairs of user requests and appropriate responses. This training allows them to generalize across various complex tasks, such as sentiment analysis, which previously required explicit fine-tuning by researchers or developers [40]. A key part of the input to these models, such as ChatGPT and Gemini (Google AI), includes a system prompt, often hidden from the user, which guides the model on how to interpret and respond to user prompts. For example, it might direct the model to act as a helpful mental health assistant. In addition, "prompt engineering" has emerged as a crucial technique in optimizing model performance. Prompt engineering involves crafting input texts that guide the model to produce the desired output without additional training. For example, refining a prompt from "Tell me about current events in health care" to "Summarize today's top news stories about technology in health care" provides the model with more specific guidance, which can enhance the relevance and accuracy of its responses [41]. While prompt engineering can be highly effective and reduce the need to retrain the model, it is important to be wary of "hallucinations," a phenomenon where models confidently generate incorrect or irrelevant outputs [42]. This can be particularly challenging in high-accuracy scenarios, such as health care and medical applications [43-46]. Thus, while prompt engineering reduces the reliance on extensive fine-tuning, it underscores the need for thorough evaluation and testing to ensure the reliability of model outputs in sensitive applications.

The existing literature includes a review of the application of machine learning and NLP in mental health [47], analyses of LLMs in medicine [32], and a scoping review of LLMs in mental health. These studies have demonstrated the effectiveness of NLP for tasks such as text categorization and sentiment analysis [47] and provided a broad overview of LLM applications in mental health [48]. However, a gap remains in systematically reviewing state-of-the-art LLMs in mental health, particularly in the comprehensive assessment of literature published since the introduction of the transformer architecture in 2017.

This systematic review addresses these gaps by providing a more in-depth analysis; evaluating the quality and applicability of studies; and exploring ethical challenges specific to LLMs, such as data privacy, interpretability, and clinical integration. Unlike previous reviews, this study excludes preprints, follows a rigorous search strategy with clear inclusion and exclusion criteria (using Cohen κ to assess the interreviewer agreement), and uses a detailed assessment of study quality and bias (using the Risk of Bias 2 tool) to ensure the reliability and reproducibility of the findings.

Guided by specific research questions, this systematic review critically assesses the use of LLMs in mental health, focusing on their applicability and efficacy in early screening, digital interventions, and clinical settings, as well as the methodologies and data sources used. The findings of this study highlight the potential of LLMs in enhancing mental health diagnostics and interventions while also identifying key challenges such as inconsistencies in model outputs and the lack of robust ethical guidelines. These insights suggest that, while LLMs hold promise, their use should be supervised by physicians, and they are not yet ready for widespread clinical implementation.

Methods

This systematic review followed the PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analyses) guidelines [49]. The protocol was registered on PROSPERO (CRD42024508617). A PRISMA checklist is available in Multimedia Appendix 1.

Search Strategies

The search was initiated on August 3, 2024, and completed on August 6, 2024, by 1 author (ZG). ZG systematically searched 5 databases: MEDLINE, IEEE Xplore, Scopus, JMIR, and ACM Digital Library using the following search keywords: (*mental health* OR *mental illness* OR *mental disorder* OR *psychiatry*) and (*large language models*). These keywords were consistently applied across each database to ensure a uniform search strategy. To conduct a comprehensive and precise search for relevant literature, strategies were tailored for different databases. All metadata were searched in MEDLINE and IEEE Xplore, whereas the search in Scopus was confined to titles, abstracts, and keywords. The JMIR database used the criteria *exact match* feature to refine search results and enhance precision. In the ACM Digital Library database, the search focused on full text. The screening of all citations involved four steps:

1. Initial search. All relevant citations were imported into a Zotero (Corporation for Digital Scholarship) citation manager library.
2. Preliminary inclusion. Citations were initially screened based on predefined inclusion criteria.
3. Duplicate removal. Citations were consolidated into a single group, from which duplicates were eliminated.
4. Final inclusion. The remaining references were carefully evaluated against the inclusion criteria to determine their suitability.

Study Selection and Eligibility Criteria

All the articles that matched the search criteria were double screened by 2 independent reviewers (ZG and KL) to ensure that each article fell within the scope of LLMs in mental health. This process involved the removal of duplicates followed by a detailed manual evaluation of each article to confirm adherence to our predefined inclusion criteria, ensuring a comprehensive and focused review. To quantify the agreement level between the reviewers and ensure objectivity, interrater reliability was calculated using Cohen κ, with a score of 0.84 indicating a good level of agreement. In instances of disagreement, a third reviewer (AL) was consulted to achieve consensus.

To assess the risk of bias, we used the Risk of Bias 2 tool, as recommended for Cochrane Reviews. The results have been visualized in Multimedia Appendix 2. We thoroughly examined each study for potential biases that could impact the validity of the results. These included biases from the randomization process, deviations from intended interventions, missing outcome data, inaccuracies in outcome measurement, and selective reporting of results. This comprehensive assessment ensures the credibility of each study.

The criteria for selecting articles were as follows: we limited our search to English-language publications, focusing on articles published between January 1, 2017, and April 30, 2024. This timeframe was chosen considering the substantial developments in the field of LLMs in 2017, marked notably by the introduction of the transformer architecture, which has greatly influenced academic and public interest in this area.

In this review, the original research articles and available full-text papers have been carefully selected, aiming to focus on the application of LLMs in mental health. To comply with the PRISMA guidelines, articles that have not been published in a peer-reviewed venue, including those only available on a preprint server, were excluded. Owing to the limited literature specifically addressing the mental health applications of LLMs, we included review articles to ensure a comprehensive perspective. The selection criteria focused on direct applications, expert evaluations, and ethical considerations related to the use of LLMs in mental health contexts, with the goal of providing a thorough analysis of this rapidly developing field.

Information Extraction

The data extraction process was jointly conducted by 2 reviewers (ZG and KL), focusing on examining the application scenarios, model architecture, data sources, methodologies used, and main outcomes from selected studies on LLMs in mental health.

Initially, we categorized each study to determine its main objectives and applications. The categorization process was conducted in 2 steps. First, after reviewing all the included articles, we grouped them into 3 primary categories: detection of mental health conditions and suicidal ideation through text, LLM use for mental health CAs, and other applications and evaluation of the LLMs in mental health. In the second step, we performed a more detailed categorization. After a thorough, in-depth reading of each article within these broad categories, we refined the classifications based on the specific goals of the studies. Following this, we summarized the main model architectures of the LLMs used and conducted a thorough examination of data sources, covering both public and private datasets. We noted that some review articles lacked detail on dataset content; therefore, we focused on providing comprehensive information on public datasets, including their origins and sample sizes. We also investigated the various methods used across different studies, including data collection strategies and analytic methodologies. We examined their comparative structures and statistical techniques to offer a clear understanding of how these methods are applied in practice.

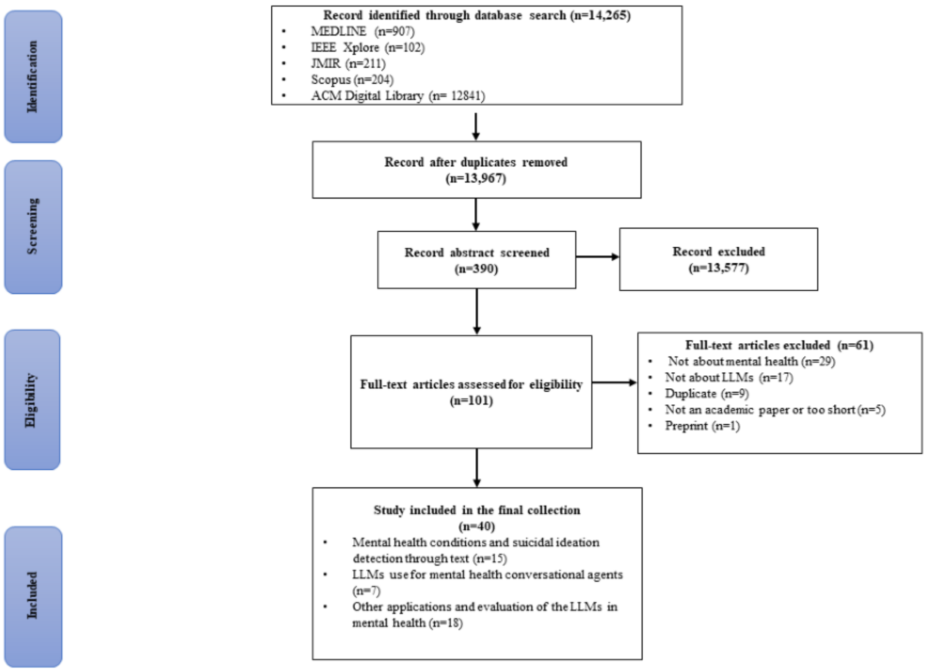
Finally, we documented the main outcomes of each study, recording significant results and aligning them with relevant performance metrics and evaluation criteria. This included providing quantitative data where applicable to underscore these findings. We used a narrative approach to synthesize the information, integrating and comparing results from various studies to emphasize the efficacy and impact of LLMs on mental health. This narrative synthesis allowed us to highlight the efficacy and impact of LLMs in mental health, providing quantitative data where applicable to underscore these findings. The results of this analysis are presented in Tables S1-S3 in Multimedia Appendix 3 [14,50-131], each corresponding to 1 of the primary categories.

Results

Strategy and Screening Process

The PRISMA diagram of the systematic screening process can be seen in Figure 1. Our initial search across 5 academic databases, namely, MEDLINE, IEEE Xplore, Scopus, JMIR, and ACM Digital Library, yielded 14,265 papers: 907 (6.36%) from MEDLINE, 102 (0.72%) from IEEE Xplore, 204 (1.43%) from Scopus, 211 (1.48%) from JMIR, and 12,841 (90.02%) from ACM Digital Library. After duplication, 97.91% (13,967/14,265) of the unique papers were retained. Subsequent screening was based on predefined inclusion and exclusion criteria, narrowing down the selection to 0.29% (40/13,967) of the papers included in this review. The reasons for the full-text exclusion of 61 papers can be found in Multimedia Appendix 4.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow of the selection process. LLM: large language model.



In our review of the literature, we classified the included articles into 3 broad categories: detection of mental health conditions and suicidal ideation through text (15/40, 38%), LLMs' use for mental health CAs (7/40, 18%), and the other applications and evaluation of the LLMs in mental health (18/40, 45%). The first category investigates the potential of LLMs for the early detection of mental illness and suicidal ideation via social media and other textual sources. Early screening is highlighted as essential for preventing the progression of mental disorders and mitigating more severe outcomes. The second category assesses LLM-supported CAs used as teletherapeutic interventions for mental health issues, such as loneliness, with a focus on evaluating their effectiveness and validity. The third category covers a broader range of LLM applications in mental health, including clinical uses such as decision support and therapy enhancement. It aims to assess the overall effectiveness, utility, and ethical considerations associated with LLMs in these settings. All selected articles are summarized in Tables S1-S3 in Multimedia Appendix 3 according to the 3 categories.

Mental Health Conditions and Suicidal Ideation Detection Through Text

Early intervention and screening are crucial in mitigating the global burden of mental health issues [132]. We examined the performance of LLMs in detecting mental health conditions and suicidal ideation through textual analysis. Of 40 articles, 6 (15%) assessed the efficacy of early screening for depression using LLMs [50,57,60,61,66,68], while another (1/40, 2%) simultaneously addressed both depression and anxiety [60]. One comprehensive study examined various psychiatric conditions, including depression, social anxiety, loneliness, anxiety, and other prevalent mental health issues [69]. Two (5%) of the 40 articles assessed and compared the ability of LLMs to perform sentiment and emotion analysis [75,81], and 5 (12%) articles focused on the capability of LLMs to analyze textual content for detecting suicidal ideation [54,65,70,72,78]. Most studies (10/40, 25%) used BERT and its variants as one of the primary models [50,54,57,62,65,66,68,69,75,78], while GPT models were also commonly used (8/40, 20%) [57,60,61,66,70,72,78,81]. Most training data (10/40, 25%) comprised social media posts [50,54,62,65,68,69,72,75,78,81] from platforms such as Twitter (Twitter, Inc), Reddit, and Sina Weibo (Sina corporation), covering languages such as English, Malay dialects, Chinese, and Portuguese. In addition, 5 (12%) of the 40 studies used datasets consisting of clinical transcripts and patient interviews [50,57,60,61,66], providing deeper insights into LLM applications in clinical mental health settings.

In studies focusing on early screening for depression, comparing results horizontally is challenging due to variations in datasets, training methods, and models across different investigations. Nonetheless, substantial evidence supports the significant potential of LLMs in detecting depression from text-based data. For example, Danner et al [57] conducted a comparative analysis using a convolutional neural network on the Distress Analysis Interview Corpus-Wizard of Oz dataset, achieving F_1-scores of 0.53 and 0.59; however, their use of GPT-3.5 demonstrated superior performance, with an F_1-score of 0.78. Another study involving the E-Distress Analysis Interview Corpus dataset (an extension of Distress Analysis Interview Corpus-Wizard of Oz) used the Robustly Optimized BERT Approach for Depression Detection to predict the Patient Health Questionnaire-8 scores from textual data. This approach identified 3 levels of depression and achieved the lowest mean absolute error of 3.65 in Patient Health Questionnaire–8 scores [66].

LLMs play an important role in sentiment analysis [75,81], which categorizes text into overall polarity classes, such as positive, neutral, negative, and occasionally mixed, and emotion classification, which assigns labels such as "joy," "sadness," "anger," and "fear" [75]. These analyses enable the detection of emotional states and potential mental health issues from textual data, facilitating early intervention [133]. Stigall et al [75] demonstrated the efficacy of these models, with their study showing that Emotion-aware BERT Tiny, a fine-tuned variant of BERT, achieved an accuracy of 93.14% in sentiment analysis and 85.46% in emotion analysis. This performance surpasses that of baseline models, including BERT-Base Cased and BERTTiny-Pretrained [75], underscoring the advantages and validity of fine-tuning in enhancing model performance. LLMs have also demonstrated robust accuracy in detecting and classifying a range of mental health syndromes, including social anxiety, loneliness, and generalized anxiety. Vajre et al [69] introduced PsychBERT, developed using a diverse training dataset from both social media texts and academic literature, which achieved an F_1-score of 0.63, outperforming traditional deep learning approaches such as convolutional neural networks and long short-term memory networks, which recorded F_1-scores of 0.57 and 0.51, respectively [69]. In research on detecting suicidal ideation using LLMs, Diniz et al [54] showcased the high efficacy of the BERTimbau large model within a non-English (Portuguese) context, achieving an accuracy of 0.955, precision of 0.961, and an F_1-score of 0.954. The assessment of the BERT model by Metzler et al [65] found that it correctly identified 88.5% of tweets as suicidal or off-topic, performing comparably to human analysts and other leading models. However, Levkovich et al [70] noted that while GPT-4 assessments of suicide risk closely aligned with those by mental health professionals, it overestimated suicidal ideation. These results underscore that while LLMs have the potential to identify tweets reflecting suicidal ideation with accuracy comparable to psychological professionals, extensive follow-up studies are required to establish their practical application in clinical settings.

LLMs in Mental Health CAs

In the growing field of mental health digital support, the implementation of LLMs as CAs has exhibited both promising advantages [14,84,91,96] and significant challenges [92,96]. The studies by Ma et al [14] and Heston [96] demonstrate the effectiveness of CAs powered by LLMs in providing timely, nonjudgmental mental health support. This intervention is particularly important for those who lack ready access to a therapist due to constraints such as time, distance, and work, as well as for certain populations considered socially marginalized, such as older adults who experience chronic loneliness and a lack of companionship [14,97]. The qualitative analysis of user interactions on Reddit by Ma et al [14] highlights that LLMs encourage users to speak up and boost their confidence by providing personalized and responsive interactions. In addition, VHope, a DialoGPT-enabled mental health CA, was evaluated by 3 experts who rated its responses as 67% relevant, 78% human-like, and 79% empathetic [84]. Another study found that after observing 717 evaluations by 100 participants on 239 autism-specific questions, 46.86% of evaluators preferred responses of the chief physicians, whereas 34.87% preferred the responses of GPT-4, and 18.27% favored the responses of Enhanced Representation through Knowledge Integration Bot (ERNIE Bot; version 2.2.3; Baidu, Inc). Moreover, ChatGPT (mean 3.64, 95% CI 3.57-3.71) outperformed physicians (mean 3.13, 95% CI 3.04-3.21) in terms of empathy [98], indicating that LLM-powered CAs are not only effective but also acceptable by users. These findings highlight the potential for LLMs to complement mental health intervention systems and provide valuable medical guidance.

The development and implementation of a non-English CA for emotion capture and categorization was explored in a study by Zygadlo et al [92]. Faced with a scarcity of Polish datasets, the study adapted by translating an existing database of personal conversations from English into Polish, which decreased the accuracy in tasks from 90% in English to 80% in Polish [92]. While the performance remained commendable, it highlighted the challenges posed by the lack of robust datasets in languages other than English, impacting the effectiveness of CAs across different linguistic environments. However, findings by He et al [98] suggest that the availability of language-specific datasets is not the sole determinant of CA performance. In their study, although ERNIE Bot was trained in Chinese and ChatGPT in English, ChatGPT demonstrated greater empathy for Chinese users [98]. This implies that factors beyond the training language and dataset availability, such as model architecture or training methodology, can also affect the empathetic responsiveness of LLMs, underscoring the complexity of human-AI interaction.

Meanwhile, the reliability of LLM-driven CAs in high-risk scenarios remains a concern [14,96]. An evaluation of 25 CAs found that in tests involving suicide scenarios, only 2 included suicide hotline referrals during the conversation [96]. This suggests that while these CAs can detect extreme emotions, few are equipped to take effective preventive measures. Furthermore, CAs often struggle with maintaining consistent communication due to limited memory capacity, leading to disruptions in conversation flow and negatively affecting user experience [14].

Other Applications and Evaluation of the LLMs in Mental Health

ChatGPT has gained attention for its unparalleled ability to generate human-like text and analyze large amounts of textual data, attracting the interest of many researchers and practitioners [100]. Numerous evaluations of LLMs in mental health have focused on ChatGPT, exploring its utility across various scenarios such as clinical diagnosis [100,106,111], treatment planning [106,128,131], medication guidance [105,109,129], patient management [106], psychiatry examinations [118], and psychology education [102], among others [107,110,127,130].

Research has highlighted ChatGPT's accuracy in diagnosing various psychiatric conditions [106,110,111,126]. For example, Franco D'Souza et al [100] evaluated ChatGPT's responses to 100 clinical psychiatric cases, awarding it an "A" rating in 61 cases, with no errors in the diagnoses of different psychiatric disorders and no unacceptable responses, underscoring ChatGPT's expertise and interpretative capacity in psychiatry. Further supporting this, Schubert et al [118] assessed the performance of ChatGPT 4.0 using neurology board-style examination questions, finding that it answered 85% of the questions correctly, surpassing the average human performance of 73.8%. Meanwhile, in a study of LLMs regarding the prognosis and long-term outcomes of depression, GPT-4, Claude (Anthropic), and Bard (Google AI) showed strong agreement with mental health professionals. They all recommended a combination of psychotherapy and antidepressant medication in every case [130]. This not only proves the reliability of LLMs for mental health assessment but also highlights their usefulness in providing valuable support and guidance for individuals seeking information or coping with mental illness.

However, the direct deployment of LLMs, such as ChatGPT, in clinical settings carries inherent risks. The outputs of LLMs are heavily influenced by prompt engineering, which can lead to inconsistencies that undermine clinical reliability [102,105-107,109]. For example, Farhat et al [105] conducted a critical evaluation of ChatGPT's ability to generate medication guidelines through detailed cross-questioning and noted that altering prompts substantially changed the responses. While ChatGPT typically provided helpful advice and recommended seeking expert consultation, it occasionally produced inappropriate medication suggestions. Perlis et al [129] verified this, showing that GPT-4 Turbo suggested medications that were considered less efficient choices or contraindicated by experts in 12% of the cases. Moreover, LLMs often lack the necessary clinical judgment capabilities. This issue was highlighted in the study by Grabb [109], which revealed that despite built-in safeguards, ChatGPT remains susceptible to generating extreme and potentially hazardous recommendations. A particularly alarming example was ChatGPT advising a patient with depression to engage in high-risk activities such as bungee jumping as a means of seeking pleasure [109]. These LLMs depend on prompt engineering [102,105,109], which means their responses can vary widely depending on the wording and context of the prompts given. The system prompts, which are predefined instructions given to the model, and the prompts used by the experimental team, such as those in the study by Farhat et al [105], guide the behavior of ChatGPT and similar LLMs. These prompts are designed to accommodate a variety of user requests within legal and ethical boundaries. However, while these boundaries are intended to ensure safe and appropriate responses, they often fail to align with the nuanced sensitivities required in psychological contexts. This mismatch underscores a significant deficiency in the clinical judgment and control of LLMs within sensitive mental health settings.

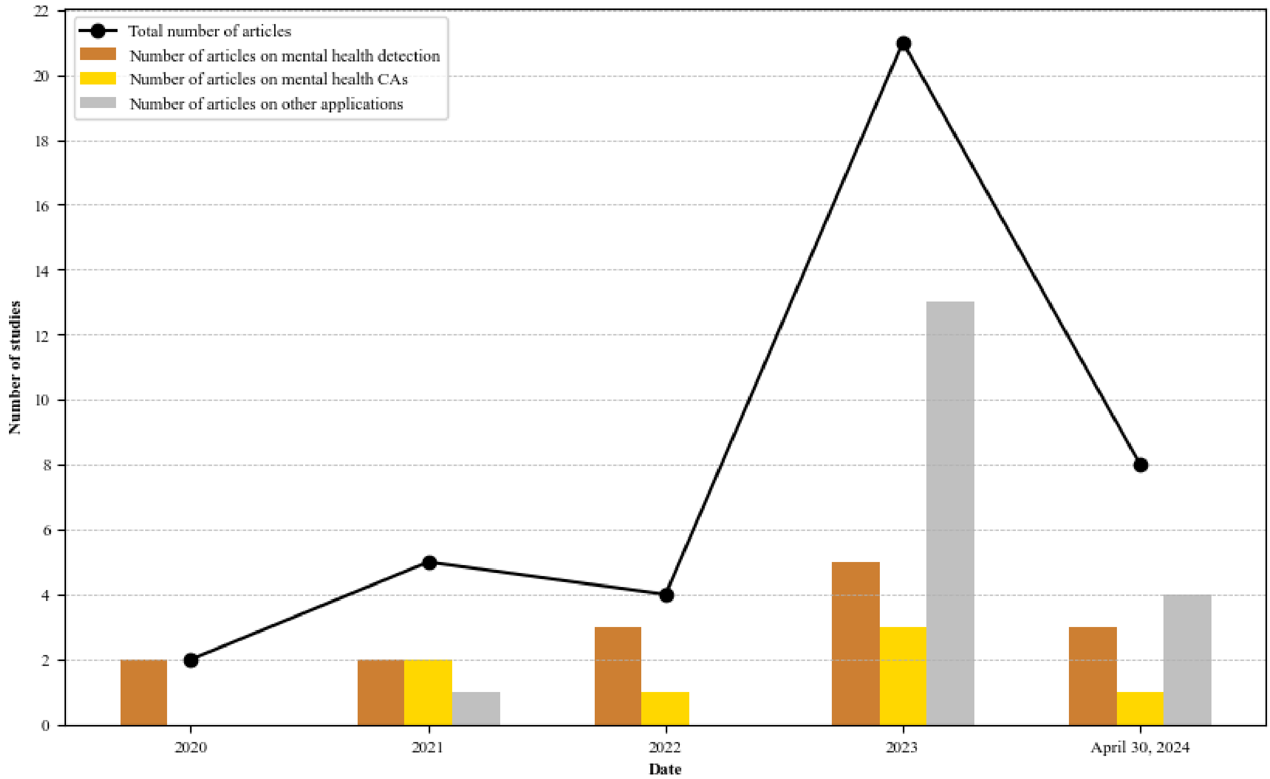
Further research into other LLMs in the mental health sector has shown a range of capabilities and limitations. For example, a study by Sezgin et al [111] highlighted Language Model for Dialogue Applications' (LaMDA's) proficiency in managing complex inquiries about postpartum depression that require medical insight or nuanced understanding; however, they pointed out its challenges with straightforward, factual questions, such as "What are antidepressants?" [111]. Assessments of LLMs such as LLaMA-7B, ChatGLM-6B, and Alpaca, involving 50 interns specializing in mental illness, received favorable feedback regarding the fluency of these models in a clinical context, with scores >9.5 out of 10. However, the results also indicated that the responses of these LLMs often failed to address mental health issues adequately, demonstrated limited professionalism, and resulted in decreased usability [116]. Similarly, a study on psychiatrists' perceptions of using LLMs such as Bard and Bing AI (Microsoft Corp) in mental health care revealed mixed feelings. While 40% of physicians indicated that they would use such LLMs to assist in answering clinical questions, some expressed serious concerns about their reliability, confidentiality, and potential to damage the patient-physician relationship [130].

Discussion

Principal Findings

In the context of the wider prominence of LLMs in the literature [14,50,57,60,61,69,96,130], this study supports the assertion that interest in LLMs is growing in the field of mental health. Figure 2 indicates an increase in the number of mental health studies using LLMs, with a notable surge observed in 2023 following the introduction of ChatGPT in late 2022. Although we included articles only up to the end of April 2024, it is evident that the number of articles related to LLMs in the field of mental health continues to show a steady increase in 2024. This marks a substantial change in the discourse around LLMs, reflecting their broader acceptance and integration into various aspects of mental health research and practice. The progression from text analysis to a diverse range of applications highlights the academic community's recognition of the multifaceted uses of LLMs. LLMs are increasingly used for complex psychological assessments, including early screening, diagnosis, and therapeutic interventions.

Figure 2. Number of articles included in this literature review, grouped by year of publication and application field. The black line indicates the total number of articles in each year. CA: conversational agent.



The findings of this study demonstrate that LLMs are highly effective in analyzing textual data to assess mental states and identify suicidal ideation [50,54,57,60,61,65,66,68,69,72,78], although their categorization often tends to be binary [50,54,65,68,69,72,78]. These LLMs possess extensive knowledge in the field of mental health and are capable of generating empathic responses that closely resemble human interactions [97,98,107]. They show great potential for providing mental health interventions with improved prognoses [50,96,110,127,128,131], with the majority being recognized by psychologists for their appropriateness and accuracy [98,100,129]. The careful and rational application of LLMs can enhance mental health care efficiently and at a lower cost, which is crucial in areas with limited health care capacity. However, there are currently no studies available that provide evaluative evidence to support the clinical use of LLMs.

Limitations

Limitations of Using LLMs in Mental Health

On the basis of the works of literature, the strengths and weaknesses of applying the LLMs in mental health are summarized in Multimedia Appendix 5.

LLMs have a broad range of applications in the mental health field. These models excel in user interaction, provide empathy and anonymity, and help reduce the stigma associated with mental illness [14,107], potentially encouraging more patients to participate in treatment. They also offer a convenient, personalized, and cost-effective way for individuals to access mental health services at any time and from any location, which can be particularly helpful for populations considered socially isolated, especially older adults [60,84,97]. In addition, LLMs can help reduce the burden of care during times of severe health care resource shortages and patient overload, such as during the COVID-19 pandemic [68]. Although previous research has highlighted the potential of LLMs in mental health, it is evident that they are not yet ready for clinical use due to unresolved technical risks and ethical issues.

The use of LLMs in mental health, particularly those fine-tuned for specific tasks such as ChatGPT, reveals clear limitations. The effectiveness of these models heavily depends on the specificity of user-generated prompts. Inappropriate or imprecise prompts can disrupt the conversation's flow and diminish the model's effectiveness [75,96,105,107,109]. Even small changes in the content or tone of prompts can sometimes lead to significant variations in responses, which can be particularly problematic in health care settings where interpretability and consistency are critical [14,105,107]. Furthermore, LLMs lack clinical judgment and are not equipped to handle emergencies [95,108]. While they can generally capture extreme emotions and recognize scenarios requiring urgent action, such as suicide ideation [54,65,70,72,78], they often fail to provide direct, practical measures, typically only advising users to seek professional help [96]. In addition, the inherent bias in LLM training data [66,106] can lead to the propagation of stereotypical, discriminatory, or biased viewpoints. This bias can also give rise to hallucinations, that is, LLMs producing erroneous or misleading information [105,131]. Furthermore, hallucinations may stem from overfitting the training data or a lack of context understanding [134]. Such inaccuracies can have serious consequences, such as providing incorrect medical information, reinforcing harmful stereotypes, or failing to recognize and appropriately respond to mental health crises [131]. For example, an LLM might reinforce a harmful belief held by a user, potentially exacerbating their mental health issues. It could also generate nonfactual, overly optimistic, or pessimistic medical advice, delaying appropriate professional intervention. These issues could undermine the integrity and fairness of social psychology [102,105,106,110].

Another critical concern is the "black box" nature of LLMs [105,107,131]. This lack of interpretability complicates the application of LLMs in mental health, where trustworthiness and clarity are important. When we talk about neural networks as black boxes, we know details such as what they were trained with, how they were trained, and what the weights are. However, with many new LLMs, such as GPT-3.5 and 4, researchers and practitioners often access the models via web interfaces or application programming interfaces without complete knowledge of the training data, methods, and model updates. This situation not only presents the traditional challenges associated with neural networks but also has all these additional problems that come from the "hidden" model.

Ethical concern is another significant challenge associated with applying LLMs in mental health. Debates are emerging around issues such as digital personhood, informed consent, the risk of manipulation, and the appropriateness of AI in mimicking human interactions [60,102,105,106,135]. A primary ethical concern is the potential alteration of the traditional therapist-patient relationship. Individuals may struggle to fully grasp the advantages and disadvantages of LLM derivatives, often choosing these options for their lower cost or greater convenience. This shift could lead to an increased reliance on the emotional support provided by AI [14], inadvertently positioning AI as the primary diagnostician and decision maker for mental health issues, thereby undermining trust in conventional health care settings. Moreover, therapists may become overly reliant on LLM-generated answers and use them in clinical decision-making, overlooking the complexities involved in clinical assessment. This reliance could compromise their professional judgment and reduce opportunities for in-depth engagement with patients [17,129,130]. Furthermore, the dehumanization and technocratic nature of mental health care has the potential to depersonalize and dehumanize patients [136], where decisions are more driven by algorithms than by human insight and empathy. This can lead to decisions becoming mechanized, lacking empathy, and detached from ethics [137]. AI systems may fail to recognize or adequately interpret the subtle and often nonverbal cues, such as the tone of voice, facial expressions, and the emotional weightage behind words, which are critical in traditional therapeutic settings [136]. These cues are essential for comprehensively understanding a patient's condition and providing empathetic care.

In addition, the current roles and accuracy of LLMs in mental health are limited. For instance, while LLMs can categorize a patient's mood or symptoms, most of these categorizations are binary, such as *depressed* or *not depressed* [50,65]. This oversimplification can lead to misdiagnoses. Data security and user privacy in clinical settings are also of utmost concern [14,54,60,96,130]. Although approximately 70% of psychiatrists believe that managing medical documents will be more efficient using LLMs, many still have concerns about their reliability and privacy [97,130,131]. These concerns could have a devastating impact on patient privacy and undermine the trust between physicians and patients if confidential treatment records stored in LLM databases are compromised. Beyond the technical limitations of AI, the current lack of an industry-benchmarked ethical framework and accountability system hinders the true application of LLMs in clinical practice [131].

Limitations of the Selected Articles

Several limitations were identified in the literature review. A significant issue is the age bias present in the social media data used for depression and mental health screening. Social media platforms tend to attract younger demographics, leading to an underrepresentation of older age groups [65]. Furthermore, most studies have focused on social media platforms, such as Twitter, primarily used by English-speaking populations, which may result in a lack of insight into mental health patterns in non–English-speaking regions. Our review included studies in Polish, Chinese, Portuguese, and Malay, all of which highlighted the significant limitations of LLMs caused by the availability and size of databases [54,61,92,98,116]. For instance, due to the absence of a dedicated Polish-language mental health database, a Polish study had to rely on machine-translated English databases [92]. While the LLMs achieve 80% accuracy in categorizing emotions and moods in Polish, this is still lower than the 90% accuracy observed in the original English dataset. This discrepancy highlights that the accuracy of LLMs can be affected by the quality of the database.

Another limitation of this study is the low diversity of LLMs studied. Although we used "large language models" as keywords in our search phase, the vast majority of identified studies (39/40, 98%) focused on BERT and its variants, as well as the GPT model, as one of the models studied. Therefore, this review provides only a limited picture of the variability expected in applicability between different LLMs. In addition, the rapid development of LLM technologies presents a limitation; this study can only reflect current advancements and may not encompass future advances or the full potential of LLMs. For instance, in tests involving psychologically relevant questions and answers, GPT-3.5 achieved an accuracy of 66.8%, while GPT-4.0 reached an accuracy of 85%, compared to the average human score of 73.8% [118]. Evaluating ChatGPT at different stages separately and comparing its performance to that of humans can lead to varied conclusions. In the assessment of prognosis and treatment planning for depression using LLMs, GPT 3.5 demonstrated a distinctly pessimistic prognosis that differed significantly from those of GPT-4, Claude, Bard, and mental health professionals [128]. Therefore, continuous monitoring and evaluation are essential to fully understand and effectively use the advancements in LLM technologies.

Opportunities and Future Work

Implementing technologies involving LLMs within the health care provision of real patients demands thorough and multifaceted evaluations. It is imperative for both industry and researchers to not let rollout exceed proportional requirements for evidence on safety and efficacy. At the level of the service provider, this includes providing explicit warnings to the public to discourage mistaking LLM functionality for clinical reliability. For example, GPT-4 introduced the ability to process and interpret image inputs within conversational contexts, leading OpenAI to issue an official warning that GPT-4 is not approved for analyzing specialized medical images such as computed tomography scans [138].

A key challenge to address in LLM research is the tendency to produce incoherent text or hallucinations. Future efforts could focus on training LLMs specifically for mental health applications, using datasets with expert labeling to reduce bias and create specialized mental health lexicons [84,102,116]. The creation of specialized datasets could take advantage of the customizable nature of LLMs, fostering the development of models that cater to the distinct needs of varied demographic groups. For instance, unlike models designed for health care professionals that assist in tasks such as data documentation, symptom analysis, medication management, and postoperative care, LLMs intended for patient interaction might be trained with an emphasis on empathy and comfortable dialogue.

Another critical concern is the problem of outdated training data in LLMs. Traditional LLMs, such as GPT-4 (with a cutoff date up to October 2023), rely on potentially outdated training data, limiting their ability to incorporate recent events or information. This can compromise the accuracy and relevance of their responses, leading to the generation of uninformative or incorrect answers, known as "hallucinations" [139]. Retrieval-augmented generation (RAG) technology offers a solution by retrieving facts from external knowledge bases, ensuring that LLMs use the most accurate and up-to-date information [140]. By searching for relevant information from numerous documents, RAG enhances the generation process with the most recent and contextually relevant content [141]. In addition, RAG includes evidence-based information, increasing the reliability and credibility of LLM responses [139].

To further enhance the reliability of LLM content and minimize hallucinations, recent studies suggest adjusting model parameters, such as the "temperature" setting [142-144]. The temperature parameter influences the randomness and predictability of outputs [145]. Lowering the temperature typically results in more deterministic outputs, enhancing coherence and reducing irrelevant content [146]. However, this adjustment can also limit the model's creativity and adaptability, potentially making it less effective in scenarios requiring diverse or nuanced responses. In mental therapy, where nuanced and sensitive responses are essential, maintaining an optimal balance is crucial. While a lower temperature can ensure accuracy, which is important for tasks such as clinical documentation, it may not suit therapeutic dialogues where personalized engagement is key. Low temperatures can lead to repetitive and impersonal responses, reducing patient engagement and therapeutic effectiveness. To mitigate these risks, regular updates of the model incorporating the latest therapeutic practices and clinical feedback are essential. Such updates could refine the model's understanding and response mechanisms, ensuring it remains a safe and effective tool for mental health care. Nevertheless, determining the "optimal" temperature setting is challenging, primarily due to the variability in tasks and interaction contexts, which require different levels of creativity and precision.

Data privacy is another important area of concern. Many LLMs, such as ChatGPT and Claude, involve sending data to third-party servers, which poses the risk of data leakage. Current studies have found that LLMs can be enhanced by privacy-enhancing techniques, such as zero-knowledge proofs, differential privacy, and federated learning [147]. In addition, privacy can be preserved by replacing identifying information in textual data with generic tokens. For example, when recording sensitive information (eg, names, addresses, or credit card numbers), using alternatives to mask tokens can help protect user data from unauthorized access [148]. This obfuscation technique ensures that sensitive user information is not stored directly, thereby enhancing data security.

The lack of interpretability in LLM decision-making is another crucial area for future research on health care applications. Future research should examine the models' architecture, training, and inferential processes for clearer understanding. Detailed documentation of training datasets, sharing of model architectures, and third-party audits would ideally form part of this undertaking. Investigating techniques such as attention mechanisms and modular architectures could illuminate aspects of neural network processing. The implementation of knowledge graphs might help in outlining logical relationships and facts [149]. In addition, another promising approach involves creating a dedicated embedding space during training, guided by an LLM. This space aligns with a causal graph and aids in identifying matches that approximate counterfactuals [146].

Before deploying LLMs in mental health settings, a comprehensive assessment of their reliability, safety, fairness, abuse resistance, interpretability, compliance with social norms, robustness, performance, linguistic accuracy, and cognitive ability is essential. It is also crucial to foster collaborative relationships among mental health professionals, patients, AI researchers, and policy makers. LLMs, for instance, have demonstrated initial competence in providing medication advice; however, their responses can sometimes be inconsistent or include inappropriate suggestions. As such, LLMs require professional oversight and should not be used independently. Nevertheless, when used as decision aids, LLMs have the potential to enhance health care efficiency. This study calls on developers of LLMs to collaborate with authoritative regulators in actively developing ethical guidelines for AI research in health care. These guidelines should aim to adopt a balanced approach that considers the multifaceted nature of LLMs and ensures their responsible integration into medical practice. They are expected to become industry benchmarks, facilitating the future development of LLMs in mental health.

Conclusions

This review examines the use of LLMs in mental health applications, including text-based screening for mental health conditions, detection of suicidal ideation, CAs, clinical use, and other related applications. Despite the potential of LLMs, challenges such as the production of hallucinatory or harmful information, output inconsistency, and ethical concerns remain. Nevertheless, as technology advances and ethical guidelines improve, LLMs are expected to become increasingly integral and valuable in mental health services, providing alternative solutions to this global health care issue.

Acknowledgments

This work was funded by the UK Research and Innovation (UKRI) Centre for Doctoral Training in artificial intelligence–enabled health care systems (grant EP/S021612/1). The funders were not involved in the study design, data collection, analysis, publication decisions, or manuscript writing. The views expressed in the text are those of the authors and not those of the funder.

Data Availability

The authors ensure that all pertinent data have been incorporated in the manuscript and the multimedia appendices. For access to the research data, interested parties may contact the corresponding author (KL) subject to a reasonable request.

Authors' Contributions

ZG and KL contributed to the conception and design of the study. ZG, KL, and AL contributed to the development of the search strategy. Database search outputs were screened by ZG, and data were extracted by ZG and KL. An assessment of the risk of bias in the included studies was performed by ZG and KL. ZG completed the literature review, collated the data, performed the data analysis, interpreted the results, and wrote the first draft of the manuscript. KL, AL, JHT, JF, and TK reviewed the manuscript and provided multiple rounds of guidance in the writing of the manuscript. All authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[DOCX File , 27 KB - mental_v11i1e57400_app1.docx]

Multimedia Appendix 2
Risk of bias assessment.
[DOCX File , 559 KB - mental_v11i1e57400_app2.docx]

Multimedia Appendix 3
Summary of the 40 selected articles from the literature on large language models in mental applications, categorized into each group.
[DOCX File , 78 KB - mental_v11i1e57400_app3.docx]

Multimedia Appendix 4
List of the studies excluded at the full-text screening stage.
[DOCX File , 30 KB - mental_v11i1e57400_app4.docx]

Multimedia Appendix 5
Summary of the strengths and weaknesses of applying the large language models in mental health.
[DOCX File , 24 KB - mental_v11i1e57400_app5.docx]

References

1. Mental health. World Health Organization. 2022 Jun 17. URL: https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response [accessed 2024-04-15]
2. Mental disorders. World Health Organization. 2022 Jun 8. URL: https://www.who.int/news-room/fact-sheets/detail/mental-disorders [accessed 2024-04-15]
3. MHPSS worldwide: facts and figures. Government of the Netherlands. URL: https://www.government.nl/topics/mhpss/mhpss-worldwide-facts-and-figures [accessed 2024-04-15]
4. Arias D, Saxena S, Verguet S. Quantifying the global burden of mental disorders and their economic value. EClinicalMedicine 2022 Dec;54:101675 [FREE Full text] [doi: 10.1016/j.eclinm.2022.101675] [Medline: 36193171]

5. Zhang W, Yang C, Cao Z, Li Z, Zhuo L, Tan Y, et al. Detecting individuals with severe mental illness using artificial intelligence applied to magnetic resonance imaging. EBioMedicine 2023 Apr;90:104541 [FREE Full text] [doi: 10.1016/j.ebiom.2023.104541] [Medline: 36996601]
6. Mental health and COVID-19: early evidence of the pandemic's impact: scientific brief, 2 March 2022. World Health Organization. 2022 Mar 2. URL: https://www.who.int/publications/i/item/WHO-2019-nCoV-Sci_Brief-Mental_health-2022.1 [accessed 2024-04-15]
7. Duden GS, Gersdorf S, Stengler K. Global impact of the COVID-19 pandemic on mental health services: a systematic review. J Psychiatr Res 2022 Oct;154:354-377 [FREE Full text] [doi: 10.1016/j.jpsychires.2022.08.013] [Medline: 36055116]
8. Mental health treatments. Mental Health America. URL: https://mhanational.org/mental-health-treatments [accessed 2024-04-15]
9. Stigma, prejudice and discrimination against people with mental illness. American Psychiatric Association. URL: https://www.psychiatry.org/patients-families/stigma-and-discrimination [accessed 2024-04-15]
10. Nietzel MT. Almost half of Americans don't seek professional help for mental disorders. Forbes. 2021 May 24. URL: https://www.forbes.com/sites/michaeltnietzel/2021/05/24/why-so-many-americans-do-not-seek-professional-help-for-mental-disorders/?sh=55b4ec4b3de7 [accessed 2024-04-15]
11. Why do people avoid mental health treatment? Thriveworks. 2022 Aug 9. URL: https://thriveworks.com/blog/why-people-avoid-mental-health-treatment/ [accessed 2024-04-15]
12. Torous J, Jän Myrick K, Rauseo-Ricupero N, Firth J. Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow. JMIR Ment Health 2020 Mar 26;7(3):e18848 [FREE Full text] [doi: 10.2196/18848] [Medline: 32213476]
13. Kumar V, Srivastava P, Dwivedi A, Budhiraja I, Ghosh D, Goyal V, et al. Large-language-models (LLM)-based AI chatbots: architecture, in-depth analysis and their performance evaluation. In: Santosh KC, Makkar A, Conway M, Singh AK, Vacavant A, Abou el Kalam A, et al, editors. Recent Trends in Image Processing and Pattern Recognition. Cham, Switzerland: Springer; 2024.
14. Ma Z, Mei Y, Su Z. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. AMIA Annu Symp Proc 2023;2023:1105-1114 [FREE Full text] [Medline: 38222348]
15. Yang K, Ji S, Zhang T, Xie Q, Kuang Z, Ananiadou S. Towards interpretable mental health analysis with large language models. arXiv. Preprint posted online on April 6, 2023 [FREE Full text] [doi: 10.18653/v1/2023.emnlp-main.370]
16. Patients turning to A and E as wait times for NHS mental health treatment spiral. The Guardian. 2022 Oct 10. URL: https://www.theguardian.com/society/2022/oct/10/nhs-mental-health-patients-wait-times [accessed 2024-04-16]
17. Elyoseph Z, Gur T, Haber Y, Simon T, Angert T, Navon Y, et al. An ethical perspective on the democratization of mental health with generative artificial intelligence. JMIR Preprints. Preprint posted online on March 2, 2024 [FREE Full text] [doi: 10.2196/preprints.58011]
18. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 06;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
19. Wysa - everyday mental health. Wysa. URL: https://www.wysa.com/ [accessed 2024-04-16]
20. Haque MD, Rubya S. An overview of chatbot-based mobile mental health apps: insights from app description and user reviews. JMIR Mhealth Uhealth 2023 May 22;11:e44838 [FREE Full text] [doi: 10.2196/44838] [Medline: 37213181]
21. Stade EC, Stirman SW, Ungar LH, Boland CL, Schwartz HA, Yaden DB, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. Npj Ment Health Res 2024 Apr 02;3(1):12 [FREE Full text] [doi: 10.1038/s44184-024-00056-z] [Medline: 38609507]
22. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine 2023 Apr;90:104512 [FREE Full text] [doi: 10.1016/j.ebiom.2023.104512] [Medline: 36924620]
23. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. World Health Organization. 2024. URL: https://iris.who.int/bitstream/handle/10665/375579/9789240084759-eng.pdf? [accessed 2024-04-16]
24. What are large language models (LLMs)? IBM. URL: https://www.ibm.com/topics/large-language-models [accessed 2024-04-16]
25. LLM evaluation: key metrics and best practices. AISERA. URL: https://aisera.com/blog/llm-evaluation/ [accessed 2024-04-16]
26. Better language models and their implications. OpenAI. 2019 Feb 14. URL: https://openai.com/research/better-language-models [accessed 2024-04-16]
27. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv. Preprint posted online on June 12, 2017 [FREE Full text]
28. Kerner SM. What are large language models (LLMs)? TechTarget. URL: https://www.techtarget.com/whatis/definition/large-language-model-LLM [accessed 2024-05-16]
29. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online on October 11, 2018 [FREE Full text] [doi: 10.5260/chara.21.2.8]
30. Rogers A, Kovaleva O, Rumshisky A. A primer in BERTology: what we know about how BERT works. arXiv. Preprint posted online on February 27, 2020 [FREE Full text] [doi: 10.1162/tacl_a_00349]

31. ChatGPT a year on: 3 ways the AI chatbot has completely changed the world in 12 months. Euronews. 2023 Nov 30. URL: https://www.euronews.com/next/2023/11/30/chatgpt-a-year-on-3-ways-the-ai-chatbot-has-completely-changed-the-world-in-12-months [accessed 2024-04-16]
32. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med 2023 Aug 17;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
33. Hickey G. The best Large Language Models (LLMs) of 2024. TechRadar. 2024. URL: https://www.techradar.com/computing/artificial-intelligence/best-llms [accessed 2024-08-08]
34. Dilmegani C. 10+ large language model examples – benchmark and use cases in '24. AIMultiple. URL: https://research.aimultiple.com/large-language-models-examples/ [accessed 2024-04-16]
35. Timeline of AI and language models. Life Architect. URL: https://lifearchitect.ai/timeline/ [accessed 2024-04-16]
36. Priest M. Large language models explained. boost.ai. 2024 Feb 20. URL: https://boost.ai/blog/llms-large-language-models [accessed 2024-04-16]
37. Vucetic D, Tayaranian M, Ziaeefard M, Clark JJ, Meyer BH, Gross WJ. Efficient fine-tuning of BERT models on the edge. In: Proceedings of the IEEE International Symposium on Circuits and Systems. 2022 Presented at: ISCAS 2022; May 27-June 1, 2022; Austin, TX. [doi: 10.1109/iscas48785.2022.9937567]
38. Kumar M. Understanding large language models and fine-tuning for business scenarios: a simple guide. Medium. 2023 Oct 27. URL: https://medium.com/@careerInAI/understanding-large-language-models-and-fine-tuning-for-business-scenarios-a-simple-guide-42f44cb687f0 [accessed 2024-04-16]
39. Mishra S, Khashabi D, Baral C, Hajishirzi H. Cross-task generalization via natural language crowdsourcing instructions. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022 Presented at: ACL 2022; May 22-27, 2022; Dublin, Ireland. [doi: 10.18653/v1/2022.acl-long.244]
40. Zhang S, Dong L, Li X, Zhang S, Sun X, Wang S, et al. Instruction tuning for large language models: a survey. arXiv. Preprint posted online on August 21, 2023 . [doi: 10.48550/arXiv.2308.10792]
41. Berryman J, Ziegler A. A developer's guide to prompt engineering and LLMs. GitHub. 2024. URL: https://github.blog/2023-07-17-prompt-engineering-guide-generative-ai-llms/ [accessed 2024-04-15]
42. Bender EM, Koller A. Climbing towards NLU: on meaning, form, and understanding in the age of data. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020 Presented at: ACL 2020; July 5-10, 2020; Online. [doi: 10.18653/v1/2020.acl-main.463]
43. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
44. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv. Preprint posted online on April 10, 2019 . [doi: 10.48550/arXiv.1904.05342]
45. Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell 2020 Sep 03;182(5):1360 [FREE Full text] [doi: 10.1016/j.cell.2020.08.029] [Medline: 32888496]
46. Trengove M, Vandersluis R, Goetz L. Response to "attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine". EBioMedicine 2023 Jul;93:104671 [FREE Full text] [doi: 10.1016/j.ebiom.2023.104671] [Medline: 37327676]
47. Le Glaz A, Haralambous Y, Kim-Dufor DH, Lenca P, Billot R, Ryan TC, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res 2021 May 04;23(5):e15708 [FREE Full text] [doi: 10.2196/15708] [Medline: 33944788]
48. Hua Y, Liu F, Yang K, Li Z, Sheu YH, Zhou P, et al. Large language models in mental health care: a scoping review. arXiv. Preprint posted online on January 1, 2024 . [doi: 10.48550/arXiv.2401.02984]
49. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009 Aug 18;151(4):264-9, W64 [FREE Full text] [doi: 10.7326/0003-4819-151-4-200908180-00135] [Medline: 19622511]
50. Verma S, Vishal, Joshi RC, Dutta MK, Jezek S, Burget R. AI-enhanced mental health diagnosis: leveraging transformers for early detection of depression tendency in textual data. In: Proceedings of the 15th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops. 2023 Presented at: ICUMT 2023; October 30-November 1, 2023; Ghent, Belgium. [doi: 10.1109/icumt61075.2023.10333301]
51. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. Preprint posted online on July 26, 2019 . [doi: 10.48550/arXiv.1907.11692]
52. Namdari R. Mental health corpus. Kaggle. URL: https://www.kaggle.com/datasets/reihanenamdari/mental-health-corpus [accessed 2024-04-17]
53. Depression: Reddit dataset (cleaned). Kaggle. URL: https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned [accessed 2024-04-17]

54. Diniz EJ, Fontenele JE, de Oliveira AC, Bastos VH, Teixeira S, Rabêlo RL, et al. Boamente: a natural language processing-based digital phenotyping tool for smart monitoring of suicidal ideation. Healthcare (Basel) 2022 Apr 08;10(4):698 [FREE Full text] [doi: 10.3390/healthcare10040698] [Medline: 35455874]
55. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020 Presented at: EMNLP 2020 - Demos; November 16-20, 2020; Online. [doi: 10.18653/v1/2020.emnlp-demos.6]
56. Souza F, Nogueira R, Lotufo R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: Proceedings of the 9th Brazilian Conference on Intelligent Systems. 2020 Presented at: BRACIS 2020; October 20-23, 2020; Rio Grande, Brazil. [doi: 10.1007/978-3-030-61377-8_28]
57. Danner M, Hadzic B, Gerhardt S, Ludwig S, Uslu I, Shao P, et al. Advancing mental health diagnostics: GPT-based method for depression detection. In: Proceedings of the 62nd Annual Conference of the Society of Instrument and Control Engineers. 2023 Presented at: SICE 2023; September 6-9, 2023; Tsu, Japan. [doi: 10.23919/sice59929.2023.10354236]
58. Gratch J, Artstein R, Lucas G, Stratou G, Scherer S, Nazarian A, et al. The distress analysis interview corpus of human and computer interviews. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation. 2014 Presented at: LREC 2014; May 26-31, 2014; Reykjavik, Iceland.
59. Extended DAIC database. University of Southern California. URL: https://dcapswoz.ict.usc.edu/extended-daic-database-download/ [accessed 2024-04-17]
60. Tao Y, Yang M, Shen H, Yang Z, Weng Z, Hu B. Classifying anxiety and depression through LLMs virtual interactions: a case study with ChatGPT. In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine. 2023 Presented at: BIBM 2023; December 5-8, 2023; Istanbul, Turkiye. [doi: 10.1109/bibm58861.2023.10385305]
61. Hayati MF, Md. Ali MA, Md. Rosli AN. Depression detection on Malay dialects using GPT-3. In: Proceedings of the IEEE-EMBS Conference on Biomedical Engineering and Sciences. 2022 Presented at: IECBES 2022; December 7-9, 2022; Kuala Lumpur, Malaysia. [doi: 10.1109/iecbes54088.2022.10079554]
62. Wang X, Chen S, Li T, Li W, Zhou Y, Zheng J, et al. Depression risk prediction for Chinese microblogs via deep-learning methods: content analysis. JMIR Med Inform 2020 Jul 29;8(7):e17958 [FREE Full text] [doi: 10.2196/17958] [Medline: 32723719]
63. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. XLNet: generalized autoregressive pretraining for language understanding. arXiv. Preprint posted online on June 19, 2019 . [doi: 10.48550/arXiv.1906.08237]
64. Wang X, Chen S, Li T, Li W, Zhou Y, Zheng J, et al. Assessing depression risk in Chinese microblogs: a corpus and machine learning methods. In: Proceedings of the IEEE International Conference on Healthcare Informatics. 2019 Presented at: ICHI 2019; June 10-13, 2019; Xi'an, China. [doi: 10.1109/ichi.2019.8904506]
65. Metzler H, Baginski H, Niederkrotenthaler T, Garcia D. Detecting potentially harmful and protective suicide-related content on Twitter: machine learning approach. J Med Internet Res 2022 Aug 17;24(8):e34705 [FREE Full text] [doi: 10.2196/34705] [Medline: 35976193]
66. Sadeghi M, Egger B, Agahi R, Richer R, Capito K, Rupp LH, et al. Exploring the capabilities of a language model-only approach for depression detection in text data. In: Proceedings of the IEEE EMBS International Conference on Biomedical and Health Informatics. 2023 Presented at: BHI 2023; October 15-18, 2023; Pittsburgh, PA. [doi: 10.1109/bhi58575.2023.10313367]
67. DeVault D, Artstein R, Benn G, Dey T, Fast E, Gainer A, et al. SimSensei kiosk: a virtual human interviewer for healthcare decision support. In: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems. 2014 Presented at: AAMAS '14; May 5-9, 2014; Paris, France. [doi: 10.1609/aaai.v29i1.9777]
68. Zhang Y, Lyu H, Liu Y, Zhang X, Wang Y, Luo J. Monitoring depression trends on Twitter during the COVID-19 pandemic: observational study. JMIR Infodemiology 2021 Jul 18;1(1):e26769 [FREE Full text] [doi: 10.2196/26769] [Medline: 34458682]
69. Vajre V, Naylor M, Kamath U, Shehu A. PsychBERT: a mental health language model for social media mental health behavioral analysis. In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine. 2021 Presented at: BIBM 2021; December 9-12, 2021; Houston, TX. [doi: 10.1109/BIBM52615.2021.9669469]
70. Levkovich I, Elyoseph Z. Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: vignette study. JMIR Ment Health 2023 Sep 20;10:e51232 [FREE Full text] [doi: 10.2196/51232] [Medline: 37728984]
71. Levi-Belz Y, Gamliel E. The effect of perceived burdensomeness and thwarted belongingness on therapists' assessment of patients' suicide risk. Psychother Res 2016 Jul;26(4):436-445. [doi: 10.1080/10503307.2015.1013161] [Medline: 25751580]
72. Howard D, Maslej MM, Lee J, Ritchie J, Woollard G, French L. Transfer learning for risk classification of social media posts: model evaluation study. J Med Internet Res 2020 May 13;22(5):e15371 [FREE Full text] [doi: 10.2196/15371] [Medline: 32401222]
73. Overview DeepMoji. Massachusetts Institute of Technology Media Lab. URL: https://www.media.mit.edu/projects/deepmoji/overview/ [accessed 2024-04-17]
74. Cer D, Yang Y, Kong SY, Hua N, Limtiaco N, St. John R, et al. Universal sentence encoder. arXiv. Preprint posted online on March 29, 2018 [FREE Full text] [doi: 10.18653/v1/d18-2029]

75. Stigall W, Khan MA, Attota D, Nweke F, Pei Y. Large language models performance comparison of emotion and sentiment classification. In: Proceedings of the 2024 ACM Southeast Conference. 2024 Presented at: ACMSE '24; April 18-20, 2024; Marietta, GA. [doi: 10.1145/3603287.3651183]
76. dair-ai / emotion. Hugging Face. URL: https://huggingface.co/datasets/dair-ai/emotion/tree/main [accessed 2024-08-06]
77. Twitter tweets sentiment dataset. Kaggle. URL: https://www.kaggle.com/datasets/yasserh/twitter-tweets-sentiment-dataset [accessed 2024-08-06]
78. Ghanadian H, Nejadgholi I, Osman HA. Socially aware synthetic data generation for suicidal ideation detection using large language models. IEEE Access 2024;12:14350-14363. [doi: 10.1109/access.2024.3358206]
79. FLAN-T5. Hugging Face. URL: https://huggingface.co/docs/transformers/model_doc/flan-t5 [accessed 2024-08-06]
80. Ghanadian H, Nejadgholi I, Osman HA. ChatGPT for suicide risk assessment on social media: quantitative evaluation of model performance, potentials and limitations. In: Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis. 2023 Presented at: WASSA@ACL 2023; July 14, 2023; Toronto, ON. [doi: 10.18653/v1/2023.wassa-1.16]
81. Lossio-Ventura JA, Weger R, Lee AY, Guinee EP, Chung J, Atlas L, et al. A comparison of ChatGPT and fine-tuned open pre-trained transformers (OPT) against widely used sentiment analysis tools: sentiment analysis of COVID-19 survey data. JMIR Ment Health 2024 Jan 25;11:e50150 [FREE Full text] [doi: 10.2196/50150] [Medline: 38271138]
82. Chung JY, Gibbons A, Atlas L, Ballard E, Ernst M, Japee S, et al. COVID-19 and mental health: predicted mental health status is associated with clinical symptoms and pandemic-related psychological and behavioral responses. medRxiv. Preprint posted online on October 14, 2021 [FREE Full text] [doi: 10.1101/2021.10.12.21264902] [Medline: 34671781]
83. Nelson LM, Simard JF, Oluyomi A, Nava V, Rosas LG, Bondy M, et al. US public concerns about the COVID-19 pandemic from results of a survey given via social media. JAMA Intern Med 2020 Jul 01;180(7):1020-1022 [FREE Full text] [doi: 10.1001/jamainternmed.2020.1369] [Medline: 32259192]
84. Beredo JL, Ong EC. A hybrid response generation model for an empathetic conversational agent. In: Proceedings of the International Conference on Asian Language Processing (IALP). 2022 Presented at: IALP 2022; October 27-28, 2022; Singapore, Singapore. [doi: 10.1109/ialp57159.2022.9961311]
85. Santos KA, Ong E, Resurreccion R. Therapist vibe: children's expressions of their emotions through storytelling with a chatbot. In: Proceedings of the Interaction Design and Children Conference. 2020 Presented at: IDC '20; June 21-24, 2020; London, UK. [doi: 10.1145/3392063.3394405]
86. Ong E, Go MJ, Lao R, Pastor J, To LB. Towards building mental health resilience through storytelling with a chatbot. In: Proceedings of the 29th International Conference on Computers in Education. 2021 Presented at: ICCE 2021; November 22-26, 2021; Online.
87. PERMA model. Corporate Finance Institute. URL: https://corporatefinanceinstitute.com/resources/management/perma-model/ [accessed 2024-04-17]
88. Rashkin H, Smith EM, Li M, Boureau YL. Towards empathetic open-domain conversation models: a new benchmark and dataset. arXiv. Preprint posted online on November 1, 2018 [FREE Full text] [doi: 10.18653/v1/p19-1534]
89. Sia DE, Yu MJ, Daliva JL, Montenegro J, Ong E. Investigating the acceptability and perceived effectiveness of a chatbot in helping students assess their well-being. In: Proceedings of the Asian CHI Symposium 2021. 2021 Presented at: Asian CHI '21; May 8-13, 2021; Yokohama, Japan. [doi: 10.1145/3429360.3468177]
90. Schwartz HA, Sap M, Kern ML, Eichstaedt JC, Kapelner A, Agrawal M, et al. Predicting individual well-being through the language of social media. Pac Symp Biocomput 2016;21:516-527 [FREE Full text] [Medline: 26776214]
91. Crasto R, Dias L, Miranda D, Kayande D. CareBot: a mental health ChatBot. In: Proceedings of the 2nd International Conference for Emerging Technology. 2021 Presented at: INCET 2021; May 21-23, 2021; Belagavi, India. [doi: 10.1109/incet51464.2021.9456326]
92. Zygadlo A. A therapeutic dialogue agent for Polish language. In: Proceedings of the 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos. 2021 Presented at: ACIIW 2021; September 28-October 1, 2021; Nara, Japan. [doi: 10.1109/aciiw52867.2021.9666281]
93. Conversational AI platform. Rasa Technologies Inc. URL: https://rasa.com/ [accessed 2024-04-18]
94. Industrial-strength natural language processing in Python. spaCy. URL: https://spacy.io/ [accessed 2024-04-18]
95. Li Y, Su H, Shen X, Li W, Cao Z, Niu S. DailyDialog: a manually labelled multi-turn dialogue dataset. arXiv. Preprint posted online on October 11, 2017 . [doi: 10.48550/arXiv.1710.03957]
96. Heston TF. Safety of large language models in addressing depression. Cureus 2023 Dec 18;15(12):e50729 [FREE Full text] [doi: 10.7759/cureus.50729] [Medline: 38111813]
97. Alessa A, Al-Khalifa H. Towards designing a ChatGPT conversational companion for elderly people. In: Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments. 2023 Presented at: PETRA '23; July 5-7, 2023; Corfu, Greece. [doi: 10.1145/3594806.3596572]
98. He W, Zhang W, Jin Y, Zhou Q, Zhang H, Xia Q. Physician versus large language model chatbot responses to web-based questions from autistic patients in Chinese: cross-sectional comparative analysis. J Med Internet Res 2024 Apr 30;26:e54706 [FREE Full text] [doi: 10.2196/54706] [Medline: 38687566]

99. Deng Z, Deng Z, Liu S, Evans R. Knowledge transfer between physicians from different geographical regions in China's online health communities. Inf Technol Manag 2023 May 19:1-18 [FREE Full text] [doi: 10.1007/s10799-023-00400-3] [Medline: 37359990]
100. Franco D'Souza R, Amanullah S, Mathew M, Surapaneni KM. Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. Asian J Psychiatr 2023 Nov;89:103770. [doi: 10.1016/j.ajp.2023.103770] [Medline: 37812998]
101. Wright B, Dave S, Dogra N. 100 Cases in Psychiatry, Second Edition. Boca Raton, FL: CRC Press; 2017.
102. Spallek S, Birrell L, Kershaw S, Devine EK, Thornton L. Can we use ChatGPT for mental health and substance use education? Examining its quality and potential harms. JMIR Med Educ 2023 Nov 30;9:e51243 [FREE Full text] [doi: 10.2196/51243] [Medline: 38032714]
103. Evidence-based information for the community. Cracks in the Ice. URL: https://cracksintheice.org.au/ [accessed 2024-04-18]
104. Positive choices: drug and alcohol education - get informed, stay smart, stay safe. Positive Choices. URL: https://positivechoices.org.au/ [accessed 2024-04-18]
105. Farhat F. ChatGPT as a complementary mental health resource: a boon or a bane. Ann Biomed Eng 2024 May;52(5):1111-1114. [doi: 10.1007/s10439-023-03326-7] [Medline: 37477707]
106. Wei Y, Guo L, Lian C, Chen J. ChatGPT: opportunities, risks and priorities for psychiatry. Asian J Psychiatr 2023 Dec;90:103808. [doi: 10.1016/j.ajp.2023.103808] [Medline: 37898100]
107. Yongsatianchot N, Torshizi PG, Marsella S. Investigating large language models' perception of emotion using appraisal theory. In: Proceedings of the 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos. 2023 Presented at: ACIIW 2023; September 10-13, 2023; Cambridge, MA. [doi: 10.1109/aciiw59127.2023.10388194]
108. Xenova / text-davinci-003. Hugging Face. URL: https://huggingface.co/Xenova/text-davinci-003 [accessed 2024-04-18]
109. Grabb D. The impact of prompt engineering in large language model performance: a psychiatric example. J Med Artif Intell 2023 Oct;6. [doi: 10.21037/jmai-23-71]
110. Hadar-Shoval D, Elyoseph Z, Lvovsky M. The plasticity of ChatGPT's mentalizing abilities: personalization for personality structures. Front Psychiatry 2023 Sep 01;14:1234397 [FREE Full text] [doi: 10.3389/fpsyt.2023.1234397] [Medline: 37720897]
111. Sezgin E, Chekeni F, Lee J, Keim S. Clinical accuracy of large language models and Google search responses to postpartum depression questions: cross-sectional study. J Med Internet Res 2023 Sep 11;25:e49240 [FREE Full text] [doi: 10.2196/49240] [Medline: 37695668]
112. Collins E, Ghahramani Z. LaMDA: our breakthrough conversation technology. Google. 2021 May 18. URL: https://blog.google/technology/ai/lamda/ [accessed 2024-04-18]
113. Tanana MJ, Soma CS, Kuo PB, Bertagnolli NM, Dembe A, Pace BT, et al. How do you feel? Using natural language processing to automatically rate emotion in psychotherapy. Behav Res Methods 2021 Oct;53(5):2069-2082 [FREE Full text] [doi: 10.3758/s13428-020-01531-z] [Medline: 33754322]
114. Welcome to LIWC-22. LIWC. URL: https://www.liwc.app/ [accessed 2024-04-18]
115. Publisher of streaming video, audio, and text library databases that promote research, teaching, and learning across disciplines, including music, counseling, history, business and more. Alexander Street. URL: https://alexanderstreet.com/ [accessed 2024-04-18]
116. Wang X, Liu K, Wang C. Knowledge-enhanced pre-training large language model for depression diagnosis and treatment. In: Proceedings of the IEEE 9th International Conference on Cloud Computing and Intelligent Systems. 2023 Presented at: CCIS 2023; August 12-13, 2023; Dali, China. [doi: 10.1109/ccis59572.2023.10263217]
117. ChineseNLP/docs /language_modeling.md. GitHub. URL: https://github.com/didi/ChineseNLP/blob/master/docs/language_modeling.md [accessed 2024-04-18]
118. Schubert MC, Wick W, Venkataramani V. Performance of large language models on a neurology board-style examination. JAMA Netw Open 2023 Dec 01;6(12):e2346721 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.46721] [Medline: 38060223]
119. Neurology board review questions and practice tests. BoardVitals. URL: https://www.boardvitals.com/neurology-board-review [accessed 2024-04-18]
120. Friedman SF, Ballentine G. Trajectories of sentiment in 11,816 psychoactive narratives. Hum Psychopharmacol 2024 Jan;39(1):e2889. [doi: 10.1002/hup.2889] [Medline: 38117133]
121. Taylor WL. "Cloze" readability scores as indices of individual differences in comprehension and aptitude. J Appl Psychol 1957;41(1):19-26. [doi: 10.1037/h0040591]
122. Erowid homepage. Erowid. URL: https://www.erowid.org/ [accessed 2024-04-18]
123. Besnard J, Ruda GF, Setola V, Abecassis K, Rodriguiz RM, Huang XP, et al. Automated design of ligands to polypharmacological profiles. Nature 2012 Dec 13;492(7428):215-220 [FREE Full text] [doi: 10.1038/nature11691] [Medline: 23235874]
124. Sunkin SM, Ng L, Lau C, Dolbeare T, Gilbert TL, Thompson CL, et al. Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. Nucleic Acids Res 2013 Jan;41(Database issue):D996-1008 [FREE Full text] [doi: 10.1093/nar/gks1042] [Medline: 23193282]

125. Demszky D, Movshovitz-Attias D, Ko J, Cowen A, Nemade G, Ravi S. GoEmotions: a dataset of fine-grained emotions. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020 Presented at: ACL 2020; July 5-10, 2020; Online. [doi: 10.18653/v1/2020.acl-main.372]
126. Wu Y, Chen J, Mao K, Zhang Y. Automatic post-traumatic stress disorder diagnosis via clinical transcripts: a novel text augmentation with large language models. In: Proceedings of the IEEE Biomedical Circuits and Systems Conference. 2023 Presented at: BioCAS 2023; October 19-21, 2023; Toronto, ON. [doi: 10.1109/biocas58349.2023.10388714]
127. Kumar H, Wang Y, Shi J, Musabirov I, Farb NA, Williams JJ. Exploring the use of large language models for improving the awareness of mindfulness. In: Proceedings of the Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems. 2023 Presented at: CHI EA '23; April 23-28, 2023; Hamburg, Germany. [doi: 10.1145/3544549.3585614]
128. Elyoseph Z, Levkovich I, Shinan-Altman S. Assessing prognosis in depression: comparing perspectives of AI models, mental health professionals and the general public. Fam Med Community Health 2024 Jan 09;12(Suppl 1):e002583 [FREE Full text] [doi: 10.1136/fmch-2023-002583] [Medline: 38199604]
129. Perlis RH, Goldberg JF, Ostacher MJ, Schneck CD. Clinical decision support for bipolar depression using large language models. Neuropsychopharmacology 2024 Aug;49(9):1412-1416. [doi: 10.1038/s41386-024-01841-2] [Medline: 38480911]
130. Blease C, Worthen A, Torous J. Psychiatrists' experiences and opinions of generative artificial intelligence in mental healthcare: an online mixed methods survey. Psychiatry Res 2024 Mar;333:115724 [FREE Full text] [doi: 10.1016/j.psychres.2024.115724] [Medline: 38244285]
131. Berrezueta-Guzman S, Kandil M, Martín-Ruiz ML, Pau de la Cruz I, Krusche S. Future of ADHD care: evaluating the efficacy of ChatGPT in therapy enhancement. Healthcare (Basel) 2024 Mar 19;12(6):683 [FREE Full text] [doi: 10.3390/healthcare12060683] [Medline: 38540647]
132. Colizzi M, Lasalvia A, Ruggeri M. Prevention and early intervention in youth mental health: is it time for a multidisciplinary and trans-diagnostic model for care? Int J Ment Health Syst 2020 Mar 24;14:23 [FREE Full text] [doi: 10.1186/s13033-020-00356-9] [Medline: 32226481]
133. Nandwani P, Verma R. A review on sentiment analysis and emotion detection from text. Soc Netw Anal Min 2021;11(1):81 [FREE Full text] [doi: 10.1007/s13278-021-00776-6] [Medline: 34484462]
134. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. arXiv. Preprint posted online on November 9, 2023 . [doi: 10.48550/arXiv.2311.05232]
135. Egli A. ChatGPT, GPT-4, and other large language models: the next revolution for clinical microbiology? Clin Infect Dis 2023 Nov 11;77(9):1322-1328 [FREE Full text] [doi: 10.1093/cid/ciad407] [Medline: 37399030]
136. Palmer A, Schwan D. Beneficent dehumanization: employing artificial intelligence and carebots to mitigate shame-induced barriers to medical care. Bioethics 2022 Feb;36(2):187-193. [doi: 10.1111/bioe.12986] [Medline: 34942057]
137. Haque OS, Waytz A. Dehumanization in medicine: causes, solutions, and functions. Perspect Psychol Sci 2012 Mar;7(2):176-186. [doi: 10.1177/1745691611429706] [Medline: 26168442]
138. Image inputs for ChatGPT - FAQ. OpenAI. URL: https://help.openai.com/en/articles/8400551-image-inputs-for-chatgpt-faq [accessed 2024-04-18]
139. What is RAG (Retrieval Enhanced Generation)? Amazon Web Services. URL: https://aws.amazon.com/cn/what-is/retrieval-augmented-generation/ [accessed 2024-08-08]
140. Models. OpenAI Platform. URL: https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4 [accessed 2024-08-08]
141. What is retrieval-augmented generation (RAG)? Google Cloud. URL: https://cloud.google.com/use-cases/retrieval-augmented-generation [accessed 2024-08-08]
142. Mündler N, He J, Jenko S, Vechev M. Self-contradictory hallucinations of large language models: evaluation, detection and mitigation. arXiv. Preprint posted online on May 25, 2023 . [doi: 10.48550/arXiv.2305.15852]
143. Forbes GC, Katlana P, Ortiz Z. Metric ensembles for hallucination detection. arXiv. Preprint posted online on October 16, 2023 . [doi: 10.48550/arXiv.2310.10495]
144. Bhayana R. Chatbots and large language models in radiology: a practical primer for clinical and research applications. Radiology 2024 Jan;310(1):e232756. [doi: 10.1148/radiol.232756] [Medline: 38226883]
145. What is LLM temperature? Iguazio. URL: https://www.iguazio.com/glossary/llm-temperature/ [accessed 2024-04-27]
146. LLM optimization parameters. Attri. URL: https://attri.ai/generative-ai-wiki/llm-optimization-parameters [accessed 2024-04-27]
147. Yao Y, Duan J, Xu K, Cai Y, Sun Z, Zhang Y. A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. High Confid Comput 2024 Jun;4(2):100211. [doi: 10.1016/j.hcc.2024.100211]
148. Vats A, Liu Z, Su P, Paul D, Ma Y, Pang Y, et al. Recovering from privacy-preserving masking with large language models. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. 2024 Presented at: ICASSP 2024; April 14-19, 2024; Seoul, Republic of Korea. [doi: 10.1109/icassp48485.2024.10448234]
149. Ramlochan S. The black box problem: opaque inner workings of large language models. Prompt Engineering. URL: https://promptengineering.org/the-black-box-problem-opaque-inner-workings-of-large-language-models/ [accessed 2024-04-18]

Abbreviations

AI: artificial intelligence
BERT: Bidirectional Encoder Representations from Transformers
CA: conversational agent
ERNIE Bot: Enhanced Representation through Knowledge Integration Bot
LaMDA: Language Model for Dialogue Application
LLM: large language model
NLP: natural language processing
PRISMA: Preferred Reporting Items for Systematic Review and Meta-Analyses
RAG: retrieval-augmented generation
ROUGE: Recall-Oriented Understudy for Gisting Evaluation

Edited by J Torous; submitted 18.02.24; peer-reviewed by A Hassan, Z Elyoseph, M Hasnain, Y Hua, M Larsen; comments to author 25.03.24; revised version received 17.05.24; accepted 03.09.24; published 18.10.24.

Please cite as:
Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K
Large Language Models for Mental Health Applications: Systematic Review
JMIR Ment Health 2024;11:e57400
URL: https://mental.jmir.org/2024/1/e57400
doi:10.2196/57400
PMID:

Development of Recommendations for the Digital Sharing of Notes With Adolescents in Mental Health Care: Delphi Study

Martine Stecher Nielsen[1], MSc; Aslak Steinsbekk[2,3], PhD; Torunn Hatlen Nøst[1,4], PhD

1

2

3

4

Corresponding Author:
Martine Stecher Nielsen, MSc

Abstract

Background: In many countries, health care professionals are legally obliged to share information from electronic health records with patients. However, concerns have been raised regarding the sharing of notes with adolescents in mental health care, and health care professionals have called for recommendations to guide this practice.

Objective: The aim was to reach a consensus among authors of scientific papers on recommendations for health care professionals' digital sharing of notes with adolescents in mental health care and to investigate whether staff at child and adolescent specialist mental health care clinics agreed with the recommendations.

Methods: A Delphi study was conducted with authors of scientific papers to reach a consensus on recommendations. The process of making the recommendations involved three steps. First, scientific papers meeting the eligibility criteria were identified through a PubMed search where the references were screened. Second, the results from the included papers were coded and transformed into recommendations in an iterative process. Third, the authors of the included papers were asked to provide feedback and consider their agreement with each of the suggested recommendations in two rounds. After the Delphi process, a cross-sectional study was conducted among staff at specialist child and adolescent mental health care clinics to assess whether they agreed with the recommendations that reached a consensus.

Results: Of the 84 invited authors, 27 responded. A consensus was reached on 17 recommendations on areas related to digital sharing of notes with adolescents in mental health care. The recommendations considered how to introduce digital access to notes, write notes, and support health care professionals, and when to withhold notes. Of the 41 staff members at child and adolescent specialist mental health care clinics, 60% or more agreed with the 17 recommendations. No consensus was reached regarding the age at which adolescents should receive digital access to their notes and the timing of digitally sharing notes with parents.

Conclusions: A total of 17 recommendations related to key aspects of health care professionals' digital sharing of notes with adolescents in mental health care achieved consensus. Health care professionals can use these recommendations to guide their practice of sharing notes with adolescents in mental health care. However, the effects and experiences of following these recommendations should be tested in clinical practice.

(JMIR Ment Health 2024;11:e57965) doi:10.2196/57965

KEYWORDS

electronic health record; EHR; electronic health records; EHRs; electronic medical record; EMR; electronic medical records; EMRs; patient record; health record; health records; personal health record; PHR; online access to electronic health records; open notes; clinical notes; adolescent mental health care; adolescent mental health; child mental health; mental health; mental illness; mental illnesses; mental disorder; mental disorders; recommendations; Delphi study; digital mental health; e-health; eHealth; e–mental health; health care professionals; digital health care

Introduction

In many countries, health care professionals are legally obligated to share information from electronic health records (EHRs), including clinical notes, medications, and test results with patients [1,2]. This information is often shared through patient portals and aligns with the growing focus on patient-centered care and patient engagement to improve health care services and individual health outcomes, such as quality of life and mental health status [3,4]. However, this practice may pose challenges for health care professionals working with adolescents in mental health care, such as preventing potential harm arising from accessing mental health notes or limiting adolescents' confidentiality [5,6].

Although health care professionals' experience of sharing notes with adolescents in mental health care has not been studied, health care professionals in adult mental health care have expressed concerns about the sensitive nature of notes in mental health care and whether reading them can be harmful to the patient or damaging for the therapeutic relationship [7-9].

Moreover, it has been stated that different levels of autonomy and maturity among adolescents can pose challenges regarding the consequences of having access to notes about themselves [6,10]. Additionally, the possibility of parents or guardians accessing notes meant for the adolescent can compromise the confidentiality of what is discussed between the health care professional and the adolescent and impede the adolescent's autonomy [5,10,11]. These potential harms make it challenging for health care professionals to determine what type of information from the EHR should be shared and when [6,12].

Health care professionals have called for recommendations and support on how to handle challenges with sharing notes with adolescents in mental health care or their guardians [8,10]. The World Health Organization has proposed that recommendations should be based on formal consensus methods such as the Delphi method [13]. Over the last decade, such recommendations have been made in related areas, for example, by providing guidance on supporting self-management and the transition to adult health care for adolescents with chronic somatic diseases [14-16] and communicating with young people in mental health care about their online behavior [17]. Such recommendations for health care professionals typically include providing adolescents and young people with information about the specific topic or treatment program early on and relevant topics to cover when giving information [14-17]. While these recommendations are often either created based on or evaluated by members of medical societies and associations, the views of diverse staff members working in the specific field are not always included, potentially excluding some experiences [16,18]

To our knowledge, no recommendations are available to guide health care professionals' digital sharing of notes with adolescents in mental health care despite the specific challenges associated with this practice. Therefore, the aim was to reach a consensus among authors of scientific papers on recommendations for health care professionals' digital sharing of notes with adolescents in mental health care and to investigate whether staff at child and adolescent specialist mental health care clinics agreed with the recommendations.

Methods

To address the aims, a Delphi study and a cross-sectional study were conducted. The CREDES (Guidance on Conducting and Reporting Delphi Studies) [19] and STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for cross-sectional studies [20] were consulted during the planning and reporting of the study.

Delphi Study With Authors of Scientific Papers

The Delphi method [21] was used to reach a consensus on recommendations for the digital sharing of notes with adolescents in mental health care. The Delphi study involved creating suggestions for recommendations, receiving feedback on the recommendations from Delphi participants, and creating final recommendations based on consensus. A literature search was performed to develop suggestions for recommendations and to identify participants for the Delphi study.

Literature Search

The aim was to include scientific papers indexed in PubMed about the digital sharing of clinical notes with adolescents in general, somatic, and mental health care, and adults in mental health care. Moreover, relevant papers identified in the references of the papers from the PubMed search were invited to participate. The search for papers was performed in PubMed in April 2023 by searching the references of the papers meeting the eligibility criteria (Table 1).

The publication date range was selected because the field is rapidly evolving, and most research relevant to our aim has been conducted within the last couple of years. An initial review of older references showed that the digital solutions, particularly those from before 2016, differed from later publications (eg, by being prototypes made for the research project). Moreover, they were, to a greater extent, considering perceptions and expectations toward potential digital solutions. Peer review in international journals indexed in PubMed was used as a proxy for quality, and hence, authors of protocols, unpublished papers, and gray literature were excluded due to the challenge of evaluating their quality.

Table . Inclusion and exclusion criteria for identifying relevant publications.

Area		Inclusion criteria	Exclusion criteria
Criteria			
	Population	• Adolescents (in general) • Adolescents in psychiatry/mental health care • Adolescents in somatic health care • Adults in psychiatry/mental health care	All other study populations
	Phenomenon of interest	• Electronic health record • Patient-accessible electronic health record • Access to clinical notes • Patient portal use	All other phenomena
Information sources and methods			
	Identifications of papers via database: database name, search string, and time period	• Relevant papers from PubMed search: "(((Adolescent[MeSH Terms]) OR (Mental Health[MeSH Terms])) OR (Psychiatr*[MeSH Terms])) AND (electronic health record*[MeSH Terms])" from 2021 to 2023	Papers from the PubMed search published before 2021
	Identification of papers via other methods: search method and time period	• Relevant references from the papers identified in the PubMed search 2016 - 2020	Papers published before 2016
Limits and restrictions			
	Language	• English	Languages other than English
	Type of publications	• Papers from empirical studies and reviews published in peer-reviewed journals	Gray literature, protocols, and unpublished papers

Participants

All authors of the scientific papers identified in the literature search were invited to participate in the Delphi study if their email addresses could be identified through the publication, their institutions' web pages, or a Google search.

Together with the invitation to participate, the authors received a description of the task and the project's purpose. Additionally, they received information explaining how they had been selected, why their participation was requested, what participation involved, and details regarding the storage and use of their data.

Data Collection

The Delphi study started with screening the scientific papers included from the literature search to identify findings that could be turned into suggestions for recommendations (Multimedia Appendix 1 [7-10,22-53]). This process is further reported in the Analysis and Results sections.

The authors who created the suggestions have experience from research on patient portals, adolescents, and mental health. Care was taken to ensure that the created suggestions were based on the reviewed literature. Before sending out the suggested recommendations, they were discussed with a research group at a university that possesses expertise in research related to digital health, patient participation, and patient education. The suggestions for recommendations were sent to the Delphi participants between June and September 2023.

The Delphi participants' responses to the recommendations were collected in two rounds, referred to as round 1 and round 2. Due to the summer holiday, the Delphi participants were given 8 weeks to respond to round 1 and 6 weeks in round 2. Two reminders were sent 2 weeks apart for both rounds.

In both rounds, the Delphi participants received an email with a link to a web-based questionnaire with the suggested recommendations for the digital sharing of notes with adolescents in mental health care. Adolescence was defined as being legally old enough to access their notes digitally. The participants were asked to score their agreement with each recommendation on a Likert scale from 1 to 5 ("strongly agree," "agree," "don't know," "disagree," and "strongly disagree"). In round 1, the participants could also comment on each recommendation and give overall comments (Multimedia Appendix 2).

In round 2 of the Delphi study, participants who responded in round 1 were asked to score the recommendations supported in round 1 (in either their original or a modified form based on comments) and the recommendations generated based on comments in round 1. Moreover, the participants were informed about the proportion of participants who agreed with each of these recommendations in round 1, in accordance with recommendations for Delphi studies [19,54-56]. For both rounds, demographic data were collected on participants' gender, profession, age, and years of working experience.

Analysis

The process of making the recommendations involved three steps: identifying scientific papers, creating suggestions for recommendations, and receiving feedback from the Delphi participants. In the first step, the eligibility criteria were first used to screen the titles and the abstracts, and then the full paper of those not possible to classify based on the title or abstract. Finally, the references of the included papers were screened to identify additional relevant studies.

The second step involved creating suggestions for recommendations. The results of all scientific papers were first coded in NVivo 14 (Lumivero) by the first author, who identified codes related to experiences, effects, views, and expectations regarding access to EHRs. After this, the coded results were discussed with the other authors who read the coded results and some of the papers. The authors prepared the recommendations in an iterative process, where several discussions were held on how the coded results could be transformed into recommendations. The first author proposed suggestions, which were then discussed until there was agreement among all authors that the suggested recommendations accurately represented the coded results. Unambiguous results from the papers were used to suggest recommendations, while conflicting or hesitant results were used to propose mutually exclusive recommendations.

The third step involved receiving feedback on the recommendations from the Delphi participants, where the scoring and comments were considered. A predetermined criterion for agreement was set for both rounds; a consensus was reached if 70% of participants responded with "strongly agree" or "agree" with the recommendation [54,55]. The recommendations that did not achieve consensus were dropped unless comments suggested modifications. The authors reviewed all the free-text comments and incorporated suggestions from comments that proposed alternative formulations of recommendations or suggested new recommendations for the preparation of round 2. The authors considered the wording of the recommendations, achieving consensus and suggestions for new recommendations in several meetings and rounds of review before round 2 was sent out. Answers to round 2 from authors who were health care professionals and those who were not were compared in a subgroup analysis to determine potential differences in the proportions agreeing. The responses from round 2 were used to create the final list of recommendations.

Cross-Sectional Study With Staff

A cross-sectional study was used to investigate if staff at specialist child and adolescent mental health care clinics agreed with the recommendations for the digital sharing of notes with adolescents in mental health care. For this purpose, a survey with a web-based questionnaire was conducted in October 2023.

Participants

Participants in the cross-sectional study were staff working at child and adolescent specialist mental health care clinics geographically spread in Norway. Inclusion criteria were staff at child and adolescent specialist mental health care clinics who had contact with patients, including administrative staff. This choice was made because different types of staff can provide information about digital access to mental health notes at various times. Staff without patient contact, such as cleaning personnel, were excluded.

To recruit staff, four departments of child and adolescent mental health at university hospitals, geographically spread in Norway, were asked to distribute invitations to relevant clinics in their region. The total number of staff at their affiliated clinics was approximately 690. Eligible staff received information about the study and a link to the questionnaire.

Data Collection

The questionnaire sent to the staff at the specialist child and adolescent mental health care clinics included the recommendations that achieved consensus in the Delphi process translated to Norwegian. Moreover, questions about their gender, years of work experience in child and adolescent specialist mental health care, profession, and role at the clinic were asked. The staff were informed about the origin of the recommendations and instructed to think about adolescents as legally old enough to access their clinical notes digitally. For each recommendation, they were asked whether they "agreed" or "disagreed." The link for the questionnaire was open for 1 month.

Analysis

To investigate whether staff at child and adolescent specialist mental health care clinics agreed with the recommendations developed in the Delphi study, the frequencies of agreement and disagreement with each statement were calculated. Moreover, descriptive statistics were performed on the sociodemographic data to present the characteristics of the staff.

Ethical Considerations

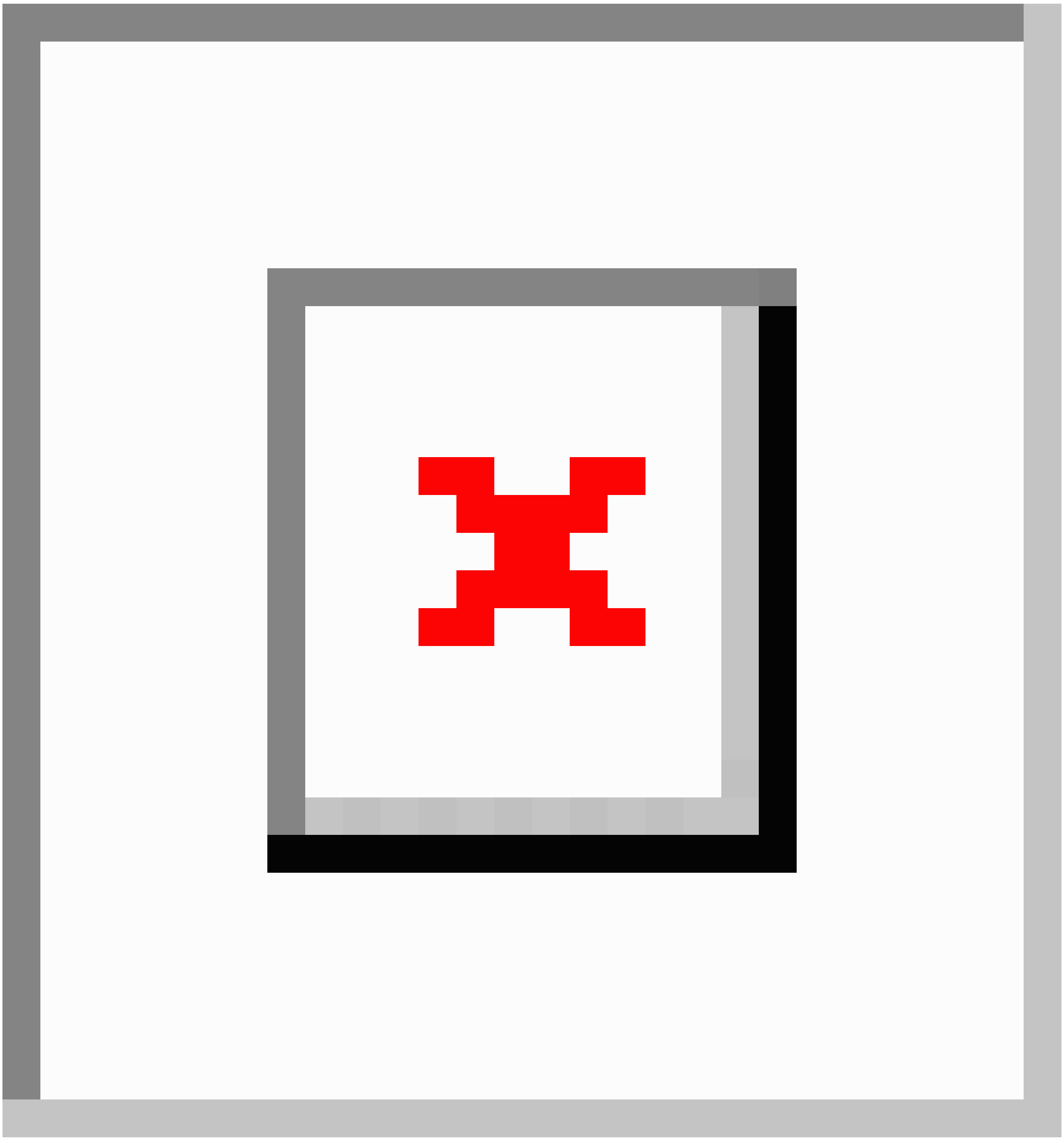
This study was approved by the Norwegian Agency for Shared Services in Education and Research (717040), which ensures that the data processing is performed in accordance with national data protection legislation. All methods were carried out in accordance with the Declaration of Helsinki. Informed consent was obtained from all participants. None of the participants received any compensation for their participation.

Results

Results From the Delphi Study

From the 1199 papers identified in the PubMed search and their relevant references, 37 scientific papers published between 2016 and 2023 were identified (Figure 1). The majority of the papers were from the United States (n=21), while 10 were from Europe, 4 were from Canada, and 1 was from Australia. Most papers used a quantitative method (n=14); 11 were qualitative, 4 were systematic and scoping reviews, and 7 were other studies such as mixed methods studies and a Delphi study (Multimedia Appendix 1). The identified papers had a total of 121 authors, of which email addresses were identified for 84 authors (Figure 1).

Figure 1. Flowchart for identification of scientific papers and participants for the Delphi study.



Among the 84 invited Delphi participants (ie, the authors of the identified papers), 27 (32%) responded in round 1; of whom, 21 (78%) responded in round 2. Most participants were researchers, and approximately one-third were health care professionals. Among the 7 health care professionals, 2 were exclusively health care professionals, while the remaining were also researchers. Most of the authors were female (Table 2).

Table . Characteristics of the Delphi participants.

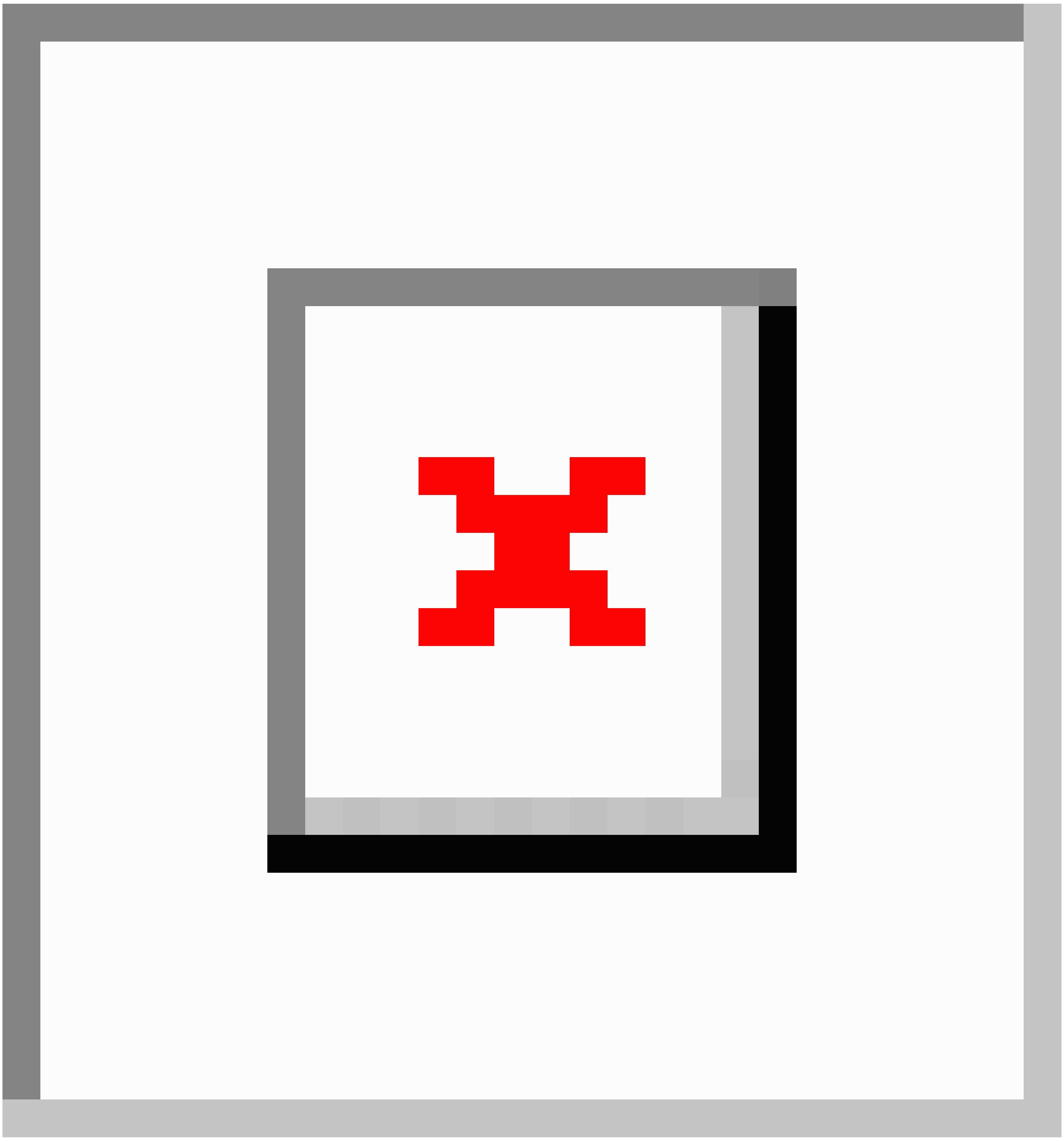
		Round 1 (n=27), n (%)	Round 2 (n=21), n (%)
Gender			
	Female	22 (81)	17 (81)
	Male	5 (19)	4 (19)
	Other	0 (0)	0 (0)
Age (years)			
	31 - 40	7 (26)	8 (38)
	41 - 50	12 (44)	7 (33)
	51 - 60	5 (19)	5 (24)
	61 - 70	3 (11)	1 (5)
Profession			
	Researcher	23 (85)	18 (90)
	Health care provider	10 (37)	7 (35)
	Patient, informal caregiver, or user representative	2 (7)	2 (10)
	Student	1 (4)	0 (0)
	Other	1 (4)	1 (5)

Development and Adjustment of Recommendations

Initially, 43 suggestions for recommendations were formulated focusing on various aspects of the digital sharing of notes, including informing about digital access to notes, the forms of the notes that are being shared, training and support for professionals, and the possibility of withholding notes from the adolescents (Multimedia Appendix 2).

The responses from round 1 were analyzed for consensus (70% agreement), and the suggested recommendations that did not reach a consensus were dropped (Figure 2). Some recommendations were intentionally formulated to be mutually exclusive; thus, support for some was not anticipated. This included recommendations to abstain from any action, while others recommended specific actions. As a result, the Delphi participants were expected to support only one of these options.

Figure 2. Flow diagram illustrating the two rounds of the Delphi study.



The recommendations that did not reach a consensus in round 1 included mutually exclusive recommendations (n=9), at what age adolescents should receive digital access to their notes or be informed about their access (n=9), the timing of digitally sharing notes with parents or guardians (n=5), and some of the points on how to inform the adolescent about their digital access to their notes (n=3). Three recommendations regarding withholding notes from adolescents did not reach a consensus in round 1, yet comments on these were used to formulate 3 new recommendations for round 2 (Table 3).

Table . Recommendations sent out in the final round (round 2), with the proportion of agreement for each recommendation in rounds 1 and 2.

Recommendations		Participants who agreed, n (%)	
		Round 1 (n=27)	Round 2 (n=21)
Information about digital access to mental health notes should be given:			
	1a ...only if considered appropriate by the health care provider.[a]	19 (70)	5 (25)
	1b ...between the first contact with the service and the first clinical consultation.	—[b]	17 (81)
	1c ...when having a consultation with the adolescent for the first time.[a]	24 (88)	20 (95)
	1d ...at later occasions if the adolescent has not comprehended the information initially provided.	—	12 (57)
	1e ...if requested by the adolescent.[a]	23 (85)	19 (90)
When informing the adolescent about digital access to mental health notes:			
	2a ...information should be provided on where the adolescent can learn more.[a]	24 (88)	18 (86)
	2b ...the sensitive nature of the notes should be discussed with the adolescent (eg, that they should not uncritically share information on social media).[a]	22 (81)	19 (90)
	2c ...parents' or guardians' potential access should be discussed.[a]	25 (93)	20 (95)
	2d ...the adolescent should be encouraged to ask questions.[a]	26 (96)	20 (95)
Mental health notes shared with both other health care professionals and adolescents:			
	3a ...should be written in a respectful language.	21 (77)	19 (90)
	3b ...should primarily be written to be useful for adolescents (eg, in plain language and avoiding or explaining medical terms).[a]	19 (70)	10 (48)
	3c ...should primarily be written to be useful for other health care professionals (eg, by using objective descriptions and medical terms).[a]	22 (81)	16 (76)
Training and/or support should be provided:			
	4a ...on how to write mental health notes.[a]	23 (85)	19 (90)
	4b ...with information about the legal and/or formal regulations on digital access to mental health notes for adolescents.[a]	25 (93)	20 (95)
	4c ...on how to digitally share mental health notes with adolescents.[a]	20 (74)	20 (95)
	4d ...on how to demonstrate to adolescents how they can access their mental health notes digitally.	—	16 (76)

Recommendations		Participants who agreed, n (%)	
		Round 1 (n=27)	Round 2 (n=21)
	4e ...with allocated time for discussion about the practice of sharing mental health notes for adolescents.[a]	20 (74)	14 (67)
	4f ...on the routines for withholding mental health notes from the adolescent.	—	18 (86)
It should be possible to withhold notes from the adolescent:			
	5a ...if it endangers the adolescent's life or causes serious harm to their health.	—	21 (100)
	5b ...if it endangers the next of kin's life or causes serious harm to their health.	—	20 (95)
	5c ...after having done a case-by-case assessment following explicitly stated criteria with a process of review by others.	—	19 (90)

[a]Statement is reformulated or modified based on comments from round 1.
[b]New in round 2.

The Final List of Recommendations

A consensus was reached on recommendations concerning how to introduce digital access to notes, write notes, and support health care professionals, and when to withhold notes (Multimedia Appendix 3). Additionally, a consensus was reached on areas where professionals should receive support and training, and on three situations where it should be possible to withhold notes from adolescents. A subgroup analysis was also conducted comparing round 2 responses between participants who were also health care professionals and those who were not, but this did not reveal any noteworthy differences.

Results From the Cross-Sectional Study

The 17 recommendations that achieved a consensus in the Delphi study were sent to staff at 4 child and adolescent specialist mental health care clinics in Norway to assess whether they agreed with the recommendations.

A total of 41 staff members responded, 90% (n=37) of whom were female. In total, 80% (n=33) of the respondents currently worked as health care professionals, 10% (n=4) worked in management positions, and 10% (n=4) worked as administrative staff. The largest groups of health care professionals were psychologists (n=21, 51%), medical doctors (n=5, 12%), and clinical social workers (n=3, 7%). Most informants (n=29, 70%) had worked at a specialist mental health care clinic for at least 5 years.

At child and adolescent specialist mental health care clinics, 70% (n=29) or more of staff agreed on 14 recommendations, and 60% (n=25) or more agreed on all 17 (Multimedia Appendix 4). The recommendations that the fewest staff agreed on were recommendation 1b (that information about digital access to notes should be given between the first contact with the service and the first clinical consultation; n=24, 59% agreed), recommendation 1e (that it should be given if requested by the adolescent; n=25, 61% agreed), and recommendation 3c (that notes should primarily be written to be useful for other health care professionals; n=27, 66% agreed).

Almost all staff agreed that training should be provided on how to write notes to be shared with the adolescent (n=40, 98%) and about the legal or formal regulations on digital access to notes for adolescents (n=40, 98%). Additionally, staff agreed it should be possible to withhold notes from the adolescent if it endangers their life or causes serious harm to the health of the adolescent or their next of kin (n=40, 98%).

Discussion

Principal Findings

A consensus was reached on 17 recommendations regarding central areas related to the digital sharing of notes with adolescents in mental health care who are legally old enough to access their notes digitally. The recommendations considered how to introduce digital access to notes, write notes, and support health care professionals, and when to withhold notes. At child and adolescent specialist mental health care clinics, 60% or more of the 41 staff members agreed with the 17 recommendations.

The findings of this study contribute to the knowledge about which recommendations related to the digital sharing of notes with adolescents are agreed on by both the authors of scientific papers and staff at child and adolescent specialist mental health care clinics. While similar recommendations for the digital sharing of notes with adolescents or with adults in mental health care have not been identified, the current recommendations resonate with the advice given to health care professionals regarding communicating with young people in mental health care about online behavior [17] and for adolescents' transition into adult care [14-16]. The similarities include initiating a conversation about the specific topic during an initial meeting [17] or at an early stage [14-16]; encouraging adolescents to ask questions [15-17]; giving advice on relevant legal regulations [14]; and addressing confidentiality, parents' role, and privacy [15,16].

Most recommendations also aligned with results from prior studies on this subject identified in the literature search. For instance, several studies in adult mental health care have highlighted the significance of communicating with patients about their access to clinical notes [9,22], the importance of how clinical notes are written when they are shared [22-24,57], and consideration of situations where some information should not be shared with the patient [22]. In addition, the need to offer support and training for health care professionals in sharing notes with adolescents [12,25] and in mental health care [8,10,26,58] has been reported in previous studies. Although consistency with prior studies was anticipated since the recommendations were based on these studies, some of the recommendations also stemmed from areas characterized by contradictory research findings [8,10,26,58].

Areas where no consensus was reached may indicate opposing opinions in the field [8,10] and geographical differences in regulations and practices [1,27,28]. In this study, a consensus was not reached on when adolescents or parents should have access to notes. This is consistent with the findings of a scoping review about sharing EHR information with children, adolescents, and parents, which reported inconclusive results and complexity related to both manual adjustments for sharing and set age limits for automatic access [10]. Moreover, it aligns with the multifaceted experiences of health care professionals with experiences of sharing clinical notes with adolescents and parents [5,10,29,59-61]. Studies have reported that health care professionals appreciate how sharing clinical notes with adolescents and parents encouraged them to ask questions about what they read [59] and improved communication [10,60], but it also could pose a threat to the adolescents' autonomy and confidentiality [5,29]. Furthermore, health care professionals face ethical challenges in preventing adolescents from accessing information that their parents have shared and from reading confidential information about their adolescent [6,10,61]. Such individual aspects can challenge set age limits for access while at the same time making decisions for manual adjustments a complex and time-consuming process.

Both authors of scientific papers and staff from child and adolescent specialist mental health care clinics agreed that notes should be written in respectful language. However, they were less certain about who the main receiver of the notes should be, with three-quarters of the authors and two-thirds of the staff agreeing that notes should primarily be written to be useful for health care professionals. This is interesting considering the increased focus on writing in a language that patients can understand, with the aim of, for example, improving their understanding of and engagement with their health and treatment [30,62,63]. This may indicate that we are still in the early stages of implementing this practice, and there are some challenges associated with the multiple audiences of the notes.

A previous study reported that mental health care professionals working with adults experienced an overall improvement in the quality of their notes even though they adapted their note writing to make it more understandable to the patient [31]. However, mental health care professionals in another study were concerned about the consequences of omitting clinically relevant information that was considered inappropriate for the patient to read [7]. A study found that health care professionals in psychiatry were more likely to perform off-the-record journaling and underreporting than those in somatic health care when sharing clinical notes with adult patients [64]. Although these studies focused on mental health care professionals working with adults, their diverse results align with this study's finding that both the authors of scientific papers and the staff reported that health professionals should receive training on note writing.

Implications

The recommendations developed in this study need to be tested, and future research should be conducted to assess the consequences of following the recommendations. However, health care professionals currently do not have recommendations that can be used in clinical practice. The implementation should be accompanied by training and support.

Strengths and Limitations

To the best of our knowledge, this is the first study to create recommendations for the digital sharing of notes with adolescents in mental health care. A strength of the study was the systematic process of developing recommendations based on research in the field. Moreover, international authors in the field evaluated and agreed on these recommendations, including researchers and health care professionals. Additionally, the recommendations underwent external validation, as most recommendations were supported by staff working at child and adolescent specialist mental health care clinics.

One limitation of the study was that the response rate for the Delphi study was relatively low (32%). This means that it is unclear whether the participants in the Delphi study were a selected group or representative of the researchers in the field. A limitation might be that most of the invited participants were authors from studies in the United States and Europe. While this might reflect that the practice of digital access to notes is more developed and common in these regions, we might have missed experiences from other settings and cultures. Similarly, selection bias could be connected to staff recruitment at child and adolescent specialist mental health care clinics. The most likely consequence is that those with a special interest will reply. Still, it is not possible to know whether they have opinions regarding the recommendations that diverge from others.

Another possible limitation is the external validity of the recommendations since they were based on studies from the United States; fewer were from Europe, and no studies were from South America, Asia, or Africa. The results may not be generalizable to these areas.

Conclusions

A total of 17 recommendations related to key aspects of health care professionals' digital sharing of notes with adolescents in mental health care achieved consensus. Health care professionals can use these recommendations to guide their practice of sharing notes with adolescents in mental health care. However, the effects and experiences of following these recommendations should be tested in clinical practice.

Acknowledgments

The authors thank the participants in the Delphi study and the staff from the specialist child and adolescent mental health care clinics for participating in this study. Moreover, we acknowledge the management of the four departments of child and adolescent mental health at university hospitals for distributing the questionnaires. This study was funded by the Faculty of Medicine and Health Sciences at the Norwegian University of Science and Technology.

Authors' Contributions

MSN, THN, and AS were responsible for designing and conceptualizing the study. MSN performed the literature search and data collection and drafted the manuscript. MSN, THN, and AS were involved in the formulation of the recommendations, the data analysis, and the writing of the manuscript. All the authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Scientific papers used to make recommendations and identify Delphi participants.

[DOC File, 36 KB - mental_v11i1e57965_app1.doc]

Multimedia Appendix 2

The 43 statements to be ranked by the participants in round 1 of the Delphi study.

[DOC File, 36 KB - mental_v11i1e57965_app2.doc]

Multimedia Appendix 3

Recommendations for digitally sharing notes with adolescents in mental health care.

[DOC File, 31 KB - mental_v11i1e57965_app3.doc]

Multimedia Appendix 4

Recommendations for digitally sharing notes with adolescents in mental health care: agreement among authors and staff.

[DOC File, 24 KB - mental_v11i1e57965_app4.doc]

References

1. Essén A, Scandurra I, Gerrits R, et al. Patient access to electronic health records: differences across ten countries. Health Policy Technol 2018 Mar;7(1):44-56. [doi: 10.1016/j.hlpt.2017.11.003]
2. Arvisais-Anhalt S, Lau M, Lehmann CU, et al. The 21st Century Cures Act and multiuser electronic health record access: potential pitfalls of information release. J Med Internet Res 2022 Feb 17;24(2):e34085. [doi: 10.2196/34085] [Medline: 35175207]
3. Benjamins J, Haveman-Nies A, Gunnink M, Goudkuil A, de Vet E. How the use of a patient-accessible health record contributes to patient-centered care: scoping review. J Med Internet Res 2021 Jan 11;23(1):e17655. [doi: 10.2196/17655] [Medline: 33427683]
4. People-centred and integrated health services: an overview of the evidence: interim report. World Health Organization. 2015. URL: https://iris.who.int/handle/10665/155002 [accessed 2024-05-30]
5. Sprackling CM, Kieren MQ, Nacht CL, Moreno MA, Wooldridge A, Kelly MM. Adolescent access to clinicians' notes: adolescent, parent, and clinician perspectives. J Adolesc Health 2024 Jan;74(1):155-160. [doi: 10.1016/j.jadohealth.2023.08.008] [Medline: 37831050]
6. Sharko M, Wilcox L, Hong MK, Ancker JS. Variability in adolescent portal privacy features: how the unique privacy needs of the adolescent patient create a complex decision-making process. J Am Med Inform Assoc 2018 Aug 1;25(8):1008-1017. [doi: 10.1093/jamia/ocy042] [Medline: 29788423]
7. van Rijt AM, Hulter P, Weggelaar-Jansen AM, Ahaus K, Pluut B. Mental health care professionals' appraisal of patients' use of web-based access to their electronic health record: qualitative study. J Med Internet Res 2021 Aug 27;23(8):e28045. [doi: 10.2196/28045] [Medline: 34448705]
8. Schwarz J, Bärkås A, Blease C, et al. Sharing clinical notes and electronic health records with people affected by mental health conditions: scoping review. JMIR Ment Health 2021 Dec 14;8(12):e34170. [doi: 10.2196/34170] [Medline: 34904956]

9. Denneson LM, Cromer R, Williams HB, Pisciotta M, Dobscha SK. A qualitative analysis of how online access to mental health notes is changing clinician perceptions of power and the therapeutic relationship. J Med Internet Res 2017 Jun 14;19(6):e208. [doi: 10.2196/jmir.6915] [Medline: 28615152]
10. Hagström J, Blease C, Haage B, Scandurra I, Hansson S, Hägglund M. Views, use, and experiences of web-based access to pediatric electronic health records for children, adolescents, and parents: scoping review. J Med Internet Res 2022 Nov 22;24(11):e40328. [doi: 10.2196/40328] [Medline: 36413382]
11. Drinkwater J, Stanley N, Szilassy E, Larkins C, Hester M, Feder G. Juggling confidentiality and safety: a qualitative study of how general practice clinicians document domestic violence in families with children. Br J Gen Pract 2017 Jun;67(659):e437-e444. [doi: 10.3399/bjgp17X689353] [Medline: 28137783]
12. Nielsen BA. Confidentiality and electronic health records: keeping up with advances in technology and expectations for access. Clin Pract Pediatric Psychol 2015;3(2):175-178. [doi: 10.1037/cpp0000096]
13. WHO handbook for guideline development, 2nd edition. World Health Organization. 2014 Dec 18. URL: https://www.who.int/publications/i/item/9789241548960 [accessed 2024-05-30]
14. Pape L, Ernst G. Health care transition from pediatric to adult care: an evidence-based guideline. Eur J Pediatr 2022 May;181(5):1951-1958. [doi: 10.1007/s00431-022-04385-z] [Medline: 35084548]
15. Kubota W, Honda M, Okada H, et al. A consensus statement on health-care transition of patients with childhood-onset chronic kidney diseases: providing adequate medical care in adolescence and young adulthood. Clin Exp Nephrol 2018 Aug;22(4):743-751. [doi: 10.1007/s10157-018-1589-8] [Medline: 29869191]
16. Roberts G, Vazquez-Ortiz M, Knibb R, et al. EAACI guidelines on the effective transition of adolescents and young adults with allergy and asthma. Allergy 2020 Nov;75(11):2734-2752. [doi: 10.1111/all.14459] [Medline: 32558994]
17. Biddle L, Rifkin-Zybutz R, Derges J, et al. Developing good practice indicators to assist mental health practitioners to converse with young people about their online activities and impact on mental health: a two-panel mixed-methods Delphi study. BMC Psychiatry 2022 Jul 19;22(1):485. [doi: 10.1186/s12888-022-04093-w] [Medline: 35854237]
18. Hoerbst A, Kohl CD, Knaup P, Ammenwerth E. Attitudes and behaviors related to the introduction of electronic health records among Austrian and German citizens. Int J Med Inform 2010 Feb;79(2):81-89. [doi: 10.1016/j.ijmedinf.2009.11.002] [Medline: 20031482]
19. Jünger S, Payne SA, Brine J, Radbruch L, Brearley SG. Guidance on Conducting and Reporting Delphi Studies (CREDES) in palliative care: recommendations based on a methodological systematic review. Palliat Med 2017 Sep;31(8):684-706. [doi: 10.1177/0269216317690685] [Medline: 28190381]
20. Cuschieri S. The STROBE guidelines. Saudi J Anaesth 2019 Apr;13(Suppl 1):S31-S34. [doi: 10.4103/sja.SJA_543_18] [Medline: 30930717]
21. Jones J, Hunter D. Consensus methods for medical and health services research. BMJ 1995 Aug 5;311(7001):376-380. [doi: 10.1136/bmj.311.7001.376] [Medline: 7640549]
22. Pisciotta M, Denneson LM, Williams HB, Woods S, Tuepker A, Dobscha SK. Providing mental health care in the context of online mental health notes: advice from patients and mental health clinicians. J Ment Health 2019 Feb;28(1):64-70. [doi: 10.1080/09638237.2018.1521924] [Medline: 30468100]
23. O'Neill S, Chimowitz H, Leveille S, Walker J. Embracing the new age of transparency: mental health patients reading their psychotherapy notes online. J Ment Health 2019 Oct;28(5):527-535. [doi: 10.1080/09638237.2019.1644490] [Medline: 31364902]
24. Cromer R, Denneson LM, Pisciotta M, Williams H, Woods S, Dobscha SK. Trust in mental health clinicians among patients who access clinical notes online. Psychiatr Serv 2017 May 1;68(5):520-523. [doi: 10.1176/appi.ps.201600168] [Medline: 28142383]
25. Nielsen MS, Steinsbekk A, Nøst TH. Views on patient portal use for adolescents in mental health care - a qualitative study. BMC Health Serv Res 2023 Feb 9;23(1):1-9. [doi: 10.1186/s12913-023-09156-6] [Medline: 36759825]
26. Kariotis TC, Prictor M, Chang S, Gray K. Impact of electronic health records on information practices in mental health contexts: scoping review. J Med Internet Res 2022 May 4;24(5):e30405. [doi: 10.2196/30405] [Medline: 35507393]
27. Hagström J, Scandurra I, Moll J, et al. Minor and parental access to electronic health records: differences across four countries. In: Seroussi B, Weber P, Dhombres F, editors. Challenges of Trustable AI and Added-Value on Health - Proceedings of MIE 2022: IOS Press; 2022:495-499.
28. Bärkås A, Hägglund M, Moll J, et al. Patients' access to their psychiatric records-a comparison of four countries. In: Seroussi B, Weber P, Dhombres F, editors. Challenges of Trustable AI and Added-Value on Health - Proceedings of MIE 2022: IOS Press; 2022:510-514.
29. Goldstein RL, Anoshiravani A, Svetaz MV, Carlson JL. Providers' perspectives on adolescent confidentiality and the electronic health record: a state of transition. J Adolesc Health 2020 Mar;66(3):296-300. [doi: 10.1016/j.jadohealth.2019.09.020] [Medline: 31831320]
30. Klein JW, Peacock S, Tsui JI, O'Neill SF, DesRoches CM, Elmore JG. Perceptions of primary care notes by patients with mental health diagnoses. Ann Fam Med 2018 Jul;16(4):343-345. [doi: 10.1370/afm.2287] [Medline: 29987083]

31. Zanaboni P, Kristiansen E, Lintvedt O, et al. Impact on patient-provider relationship and documentation practices when mental health patients access their electronic health records online: a qualitative study among health professionals in an outpatient setting. BMC Psychiatry 2022 Jul 28;22(1):508. [doi: 10.1186/s12888-022-04123-7] [Medline: 35902841]
32. Fagerlund AJ, Kristiansen E, Simonsen RA. Experiences from using patient accessible electronic health records - a qualitative study within Sámi mental health patients in Norway. Int J Circumpolar Health 2022 Dec;81(1):2025682. [doi: 10.1080/22423982.2022.2025682] [Medline: 35037572]
33. Hochwarter S, Fehler G, Muente C, et al. Design of a patient-accessible electronic health record system in mental health. In: Séroussi B, Weber P, Dhombres F, editors. Challenges of Trustable AI and Added-Value on Health - Proceedings of MIE 2022: IOS Press; 2022.
34. Sipanoun P, Oulton K, Gibson F, Wray J. The experiences and perceptions of users of an electronic patient record system in a pediatric hospital setting: a systematic review. Int J Med Inform 2022 Apr;160:104691. [doi: 10.1016/j.ijmedinf.2022.104691] [Medline: 35091287]
35. Blease C, Kharko A, Hägglund M, et al. The benefits and harms of open notes in mental health: a Delphi survey of international experts. PLoS One 2021 Oct 13;16(10):e0258056. [doi: 10.1371/journal.pone.0258056] [Medline: 34644320]
36. Bruni T, LaLonde L, Maragakis A, et al. The use of electronic health record tools to improve evidence-based treatment of adolescent depression in primary care. Acad Pediatr 2021;21(7):1195-1202. [doi: 10.1016/j.acap.2021.05.022] [Medline: 34098172]
37. Blease CR, O'Neill SF, Torous J, DesRoches CM, Hagglund M. Patient access to mental health notes: motivating evidence-informed ethical guidelines. J Nerv Ment Dis 2021 Apr 1;209(4):265-269. [doi: 10.1097/NMD.0000000000001303] [Medline: 33764954]
38. Pageler NM, Webber EC, Lund DP. Implications of the 21st Century Cures Act in pediatrics. Pediatrics 2021 Mar;147(3):e2020034199. [doi: 10.1542/peds.2020-034199] [Medline: 33293349]
39. Kelly MM, Smith CA, Hoonakker PLT, et al. Stakeholder perspectives in anticipation of sharing physicians' notes with parents of hospitalized children. Acad Pediatr 2021 Mar;21(2):259-264. [doi: 10.1016/j.acap.2020.11.018] [Medline: 33259951]
40. Dohil I, Cruz R, Sweet H, Huang JS. Sharing notes with adolescents and young adults admitted to an inpatient psychiatry unit. J Am Acad Child Adolesc Psychiatry 2021 Mar;60(3):317-320. [doi: 10.1016/j.jaac.2020.09.016] [Medline: 33035620]
41. Strudwick G, Booth RG, McLean D, et al. Identifying indicators of meaningful patient portal use by psychiatric populations. Inform Health Soc Care 2020 Oct 1;45(4):396-409. [doi: 10.1080/17538157.2020.1776291] [Medline: 32603617]
42. Strudwick G, Yeung A, Gratzer D. Easy access, difficult consequences? Providing psychiatric patients with access to their health records electronically. Front Psychiatry 2019 Dec 16;10:917. [doi: 10.3389/fpsyt.2019.00917] [Medline: 31920763]
43. Miklin DJ, Vangara SS, Delamater AM, Goodman KW. Understanding of and barriers to electronic health record patient portal access in a culturally diverse pediatric population. JMIR Med Inform 2019 Apr 26;7(2):e11570. [doi: 10.2196/11570] [Medline: 31066681]
44. Leung K, Clark C, Sakal M, Friesen M, Strudwick G. Patient and family member readiness, needs, and perceptions of a mental health patient portal: a mixed methods study. Stud Health Technol Inform 2019;257:266-270. [Medline: 30741207]
45. Ancker JS, Sharko M, Hong M, Mitchell H, Wilcox L. Should parents see their teen's medical record? Asking about the effect on adolescent-doctor communication changes attitudes. J Am Med Inform Assoc 2018 Dec 1;25(12):1593-1599. [doi: 10.1093/jamia/ocy120] [Medline: 30247699]
46. Petersson L, Erlingsdóttir G. Open notes in Swedish psychiatric care (part 2): survey among psychiatric care professionals. JMIR Ment Health 2018 Jun 21;5(2):e10521. [doi: 10.2196/10521] [Medline: 29929946]
47. Denneson LM, Chen JI, Pisciotta M, Tuepker A, Dobscha SK. Patients' positive and negative responses to reading mental health clinical notes online. Psychiatr Serv 2018 May 1;69(5):593-596. [doi: 10.1176/appi.ps.201700353] [Medline: 29493408]
48. Stablein T, Loud KJ, DiCapua C, Anthony DL. The catch to confidentiality: the use of electronic health records in adolescent health care. J Adolesc Health 2018 May;62(5):577-582. [doi: 10.1016/j.jadohealth.2017.11.296] [Medline: 29422435]
49. Petersson L, Erlingsdóttir G. Open notes in Swedish psychiatric care (part 1): survey among psychiatric care professionals. JMIR Ment Health 2018 Feb 2;5(1):e11. [doi: 10.2196/mental.9140] [Medline: 29396386]
50. Peck P, Torous J, Shanahan M, Fossa A, Greenberg W. Patient access to electronic psychiatric records: a pilot study. Health Policy Technol 2017 Sep;6(3):309-315. [doi: 10.1016/j.hlpt.2017.06.003]
51. Kipping S, Stuckey MI, Hernandez A, Nguyen T, Riahi S. A web-based patient portal for mental health care: benefits evaluation. J Med Internet Res 2016 Nov 16;18(11):e294. [doi: 10.2196/jmir.6483] [Medline: 27852556]
52. Hong MK, Wilcox L, Feustel C, Wasileski-Masker K, Olson TA, Simoneaux SF. Adolescent and caregiver use of a tethered personal health record system. AMIA Annu Symp Proc 2017 Feb 10;2016:628-637. [Medline: 28269859]
53. Thompson LA, Martinko T, Budd P, Mercado R, Schentrup AM. Meaningful use of a confidential adolescent patient portal. J Adolesc Health 2016 Feb;58(2):134-140. [doi: 10.1016/j.jadohealth.2015.10.015] [Medline: 26802988]
54. Vogel C, Zwolinsky S, Griffiths C, Hobbs M, Henderson E, Wilkins E. A Delphi study to build consensus on the definition and use of big data in obesity research. Int J Obes (Lond) 2019 Dec;43(12):2573-2586. [doi: 10.1038/s41366-018-0313-9] [Medline: 30655580]

55. Slade SC, Dionne CE, Underwood M, Buchbinder R. Standardised method for reporting exercise programmes: protocol for a modified Delphi study. BMJ Open 2014 Dec 30;4(12):e006682. [doi: 10.1136/bmjopen-2014-006682] [Medline: 25550297]
56. Spranger J, Homberg A, Sonnberger M, Niederberger M. Reporting guidelines for Delphi techniques in health sciences: a methodological review. Z Evid Fortbild Qual Gesundhwes 2022 Aug;172:1-11. [doi: 10.1016/j.zefq.2022.04.025] [Medline: 35718726]
57. Meier-Diedrich E, Davidge G, Hägglund M, et al. Changes in documentation due to patient access to electronic health records: protocol for a scoping review. JMIR Res Protoc 2023 Aug 28;12:e46722. [doi: 10.2196/46722] [Medline: 37639298]
58. Blease CR, O'Neill S, Walker J, Hägglund M, Torous J. Sharing notes with mental health patients: balancing risks with respect. Lancet Psychiatry 2020 Nov;7(11):924-925. [doi: 10.1016/S2215-0366(20)30032-8] [Medline: 32059796]
59. Kelly MM, Coller RJ, Hoonakker PLT, Nacht CL, Dean SM. Provider experiences with offering families bedside health record access across a children's hospital. Hosp Pediatr 2020 Nov;10(11):1002-1005. [doi: 10.1542/hpeds.2020-0044] [Medline: 33109520]
60. Janssen A, Keep M, Selvadurai H, et al. Factors that influence use of a patient portal by health professionals. Int J Environ Res Public Health 2021 Feb 15;18(4):1877. [doi: 10.3390/ijerph18041877] [Medline: 33671953]
61. Bourgeois FC, DesRoches CM, Bell SK. Ethical challenges raised by opennotes for pediatric and adolescent patients. Pediatrics 2018 Jun;141(6):e20172745. [doi: 10.1542/peds.2017-2745] [Medline: 29776979]
62. Smolle C, Schwarz CM, Hoffmann M, Kamolz LP, Sendlhofer G, Brunner G. Design and preliminary evaluation of a newly designed patient-friendly discharge letter - a randomized, controlled participant-blind trial. BMC Health Serv Res 2021 May 12;21(1):450. [doi: 10.1186/s12913-021-06468-3] [Medline: 33975590]
63. Lin R, Gallagher R, Spinaze M, et al. Effect of a patient-directed discharge letter on patient understanding of their hospitalisation. Intern Med J 2014 Sep;44(9):851-857. [doi: 10.1111/imj.12482] [Medline: 24863954]
64. Kristiansen E, Johansen M, Zanaboni P. Healthcare personnels' experience with patients' online access to electronic health records: differences between professions, regions, and somatic and psychiatric healthcare. : Linköping University Electronic Press Presented at: 17th Scandinavian Conference on Health Informatics; Nov 12 to 13, 2019:; Oslo, Norway p. 93-98 URL: https://ep.liu.se/ecp/161/016/ecp19161016.pdf [accessed 2023-05-30]

Abbreviations

CREDES: Guidance on Conducting and Reporting Delphi Studies
EHR: electronic health record
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by J Torous; submitted 01.03.24; peer-reviewed by E Meier-Diedrich, J Moll; revised version received 18.04.24; accepted 24.04.24; published 06.06.24.

Please cite as:
Nielsen MS, Steinsbekk A, Nøst TH
Development of Recommendations for the Digital Sharing of Notes With Adolescents in Mental Health Care: Delphi Study
JMIR Ment Health 2024;11:e57965
URL: https://mental.jmir.org/2024/1/e57965
doi:10.2196/57965

Clinical Use of Mental Health Digital Therapeutics in a Large Health Care Delivery System: Retrospective Patient Cohort Study and Provider Survey

Samuel J Ridout[1], MD, PhD; Kathryn K Ridout[2,3], MD, PhD; Teresa Y Lin[2], MPH; Cynthia I Campbell[2,4], MPH, PhD

1

2

3

4

Corresponding Author:
Cynthia I Campbell, MPH, PhD

Abstract

Background: While the number of digital therapeutics (DTx) has proliferated, there is little real-world research on the characteristics of providers recommending DTx, their recommendation behaviors, or the characteristics of patients receiving recommendations in the clinical setting.

Objective: The aim of this study was to characterize the clinical and demographic characteristics of patients receiving DTx recommendations and describe provider characteristics and behaviors regarding DTx.

Methods: This retrospective cohort study used electronic health record data from a large, integrated health care delivery system. Demographic and clinical characteristics of adult patients recommended versus not recommended DTx by a mental health provider between May 2020 and December 2021 were examined. A cross-sectional survey of mental health providers providing these recommendations was conducted in December 2022 to assess the characteristics of providers and recommendation behaviors related to DTx. Parametric and nonparametric tests were used to examine statistical significance between groups.

Results: Of 335,250 patients with a mental health appointment, 53,546 (16%) received a DTx recommendation. Patients recommended to DTx were younger, were of Asian or Hispanic race or ethnicity, were female, were without medical comorbidities, and had commercial insurance compared to those without a DTx recommendation (P<.001). More patients receiving a DTx recommendation had anxiety or adjustment disorder diagnoses, but less had depression, bipolar, or psychotic disorder diagnoses (P<.001) versus matched controls not recommended to DTx. Overall, depression and anxiety symptom scores were lower in patients recommended to DTx compared to matched controls not receiving a recommendation, although female patients had a higher proportion of severe depression and anxiety scores compared to male patients. Provider survey results indicated a higher proportion of nonprescribers recommended DTx to patients compared to prescribers (P=.008). Of all providers, 29.4% (45/153) reported using the suggested internal electronic health record–based tools (eg, smart text) to recommend DTx, and of providers recommending DTx resources to patients, 64.1% (98/153) reported they follow up with patients to inquire on DTx benefits. Only 38.4% (58/151) of respondents report recommending specific DTx modules, and of those, 58.6% (34/58) report following up on the impact of these specific modules.

Conclusions: DTx use in mental health was modest and varied by patient and provider characteristics. Providers do not appear to actively engage with these tools and integrate them into treatment plans. Providers, while expressing interest in potential benefits from DTx, may view DTx as a passive strategy to augment traditional treatment for select patients.

(JMIR Ment Health 2024;11:e56574) doi:10.2196/56574

KEYWORDS

digital therapeutics; depression; anxiety; mental health; retrospective cohort; electronic health record; adults; survey; recommendation; mobile phone

Introduction

The digital therapeutics (DTx) market has experienced rapid growth in recent years, with the global DTx market valued around US $5.2 billion in 2022 and market research analysis projecting double-digit compounded annual growth from 2023 through 2030 [1]. In the first half of 2021 alone, the DTx business sector raised US $1.6 billion in venture capital [2]. Evidence suggests that DTx resources may be acceptable and scalable and therefore are a logical adjunct for mental health treatment, particularly given the growing need for mental health care [3].

However, despite DTx proliferation and potential for application in the mental health care space, patient engagement has been minimal [4]. Reasons for this poor engagement may include a lack of personal support in DTx (ie, integration of the tool into treatment rather than passive deployment of digital resources), which has been shown as important for the acceptability and usability of mental health DTx [5]. Further, there is a notable evidence gap for DTx's mental health treatment effectiveness, with most evidence-based DTx having minimal market share [6], which may impact provider recommendations of DTx. Small, randomized controlled trials (RCTs) of DTx have shown small to moderate effect sizes for smartphone interventions in reducing depression and anxiety symptoms compared to controls [7,8]. A recent meta-analysis of 22 RCTs of mindfulness DTx showed small to moderate positive effects on depression and stress symptoms but not on anxiety [9]. An RCT of 88 college students using the Calm app (CALM.com, Inc) showed an improvement in perceived stress, mindfulness, and self-compassion in the users compared to controls [10]. Finally, an RCT of 146 employees with depressive symptoms using the myStrength app (Livongo Health, Inc) showed a more rapid reduction in depressive symptoms compared to controls [11]. The literature is not yet robust on which clinical populations are most appropriate for DTx, but those with mild to moderate symptoms of depression and anxiety, rather than severe, appear to benefit most [12,13]. DTx that are based on cognitive behavioral therapy may also be more effective, but much depends on the patient population as well as DTx design [13].

A mental health provider recommendation could be an influential factor for patient engagement with DTx tools. A 2016 survey of physicians found broad interest in DTx, but physicians also reported concerns over limited experience with such tools as well as lack of DTx data availability in the electronic medical record as barriers to DTx implementation [14]. Provider lack of knowledge regarding DTx, lack of investment and infrastructure, technical issues, workflow, and workload concerns have been identified as barriers to mobile health app adoption and implementation by providers [15-17]. In mental health, DTx tool adoption may vary by provider type and scope of practice, with psychiatrists reporting less bandwidth to focus on therapeutic DTx tools and therapists reporting that caseload, familiarity with DTx content and evidence, and limits on access to DTx data impact the ability to engage with patients around these tools [18]. Despite these descriptions of barriers to DTx tool adoption, there are little data on who providers recommend DTx tools to, the characteristics of providers recommending DTx tools, or their behaviors in recommending DTx tools. Understanding the characteristics of providers and their behaviors in recommending DTx tools can inform how these tools are integrated into mental health care, which is of particular importance given the increasing patient need for mental health resources and the dominant role mental health is assuming in the DTx space.

This study sought to characterize the clinical use of DTx resources within a large, integrated health care delivery system in two ways: (1) using electronic health record (EHR) data to identify the clinical and demographic characteristics of patients receiving recommendations from mental health providers to use DTx tools and (2) using an anonymous provider survey to determine characteristics and DTX recommending behaviors of mental health providers in clinical practice. Findings can inform the integration of DTx into mental health clinical care.

Methods

Setting

Kaiser Permanente Northern California (KPNC) is a large, diverse, integrated health care delivery system with approximately 4.6 million members. The membership is largely representative of the region, with less representation at the extreme ends of the income distribution [19]. Mental health services are primarily provided internally, both individually and in group format. Services are also available by referral to external community providers as needed.

DTx Resources

KPNC offers multiple DTx tools to its members at no cost. During the study time frame, the portfolio included Calm, myStrength, Headspace (Headspace Health, Inc), SilverCloud (Amwell), Thrive (Waypoint), and Whil (RethinkFirst) [20]. Providers can recommend DTx tools to patients based on their clinical judgment using smart text elements in the EHR, which are inserted into electronic secure messages to the patients with an associated link to the DTx. This electronic secure message contact is documented in the after-visit summary if providers use these smart text elements.

Study Design and Data Sources

Overview

This study had 2 components. The first used a retrospective cohort design with KPNC EHR data to identify adult patients who were seen in the mental health department, diagnosed with a mental health disorder, and received a provider recommendation to a DTx, as well as a matched cohort of patients who were also seen in the mental health department with a diagnosed mental health disorder but did not receive a recommendation to a DTx. The second component consisted of an anonymous web-based survey of health system mental health providers. This study followed the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines for observational studies [21] and the CHERRIES (Checklist for Reporting Results of Internet E-Surveys) for web-based surveys (Checklists 1 and 2) [22].

EHR Cohort

Adult patients aged 18 years or older with a mental health department visit and mental health diagnosis between May 1, 2020, and December 31, 2021, were included (Multimedia Appendix 1). Qualifying mental health disorders based on the *International Classification of Diseases, 10th Revision* (*ICD-10*) diagnosis codes included generalized anxiety disorder, bipolar disorder, major depressive disorder, adjustment disorders, posttraumatic stress disorder, psychosis, and sleep disorders (Multimedia Appendix 2). The index date was the first mental health visit; subsequent visits were excluded. Patients were excluded if they had less than 1-year health plan membership prior to their index mental health visit, in order to ensure the ability to capture baseline diagnoses assigned at patient encounters.

EHR Measures

The primary outcomes were the number and proportion of patients who received a DTx recommendation within 30 days of their index date. Within this time frame, we assumed that the primary diagnosis of their index mental health visit was the reason for a DTx recommendation. The recommendation was determined through text string searches (Multimedia Appendix 3) using relevant phrases within provider progress notes, smart data element capturing questionnaires, and secure messages from the patient's provider.

Additional EHR-based measures included patient demographic variables (sex, race, ethnicity, and age group), type of psychiatric provider seen (nurse, therapist, and physician), neighborhood deprivation index (as a proxy measure for socioeconomic status, categorized into quintiles), insurance type (commercial, Medicare, or Medicaid), Charlson Comorbidity Index [23], and KPNC medical service area. The Patient Health Questionnaire 9-Item (PHQ-9) [24] and Generalized Anxiety Disorder 7-Item (GAD-7) [25] questionnaires, which identify symptoms for depression and anxiety, respectively, were extracted from the EHR if available within 30 days of the index visit. Each item in the PHQ-9 is scored 0 to 3, providing a 0 to 27 severity total score. The total score is used to determine cut points for depression: 0‐4 is normal, 5‐9 is mild, 10‐14 is moderate, 15‐19 is moderate severe, and 20 or more is severe. The GAD-7 is a 7-item generalized anxiety disorder scale. Each item is scored 0 to 3, providing a 0 to 21 severity total score. The total score is used to determine cut points for generalized anxiety disorder: 0‐4 is normal, 5‐9 is mild, 10‐14 is moderate, and 15 or more is severe.

Provider Survey

We conducted an anonymous web-based survey of a convenience sample of mental health providers throughout the KPNC region to understand the characteristics of providers recommending DTx tools and their recommendation behaviors. This survey was distributed 3 times in December 2022 to department managers and chiefs of psychiatry via an email that contained a Microsoft Forms link and a request to distribute it to clinical staff of approximately 2130 therapists and 375 psychiatrists. Participation was voluntary.

The survey included 26 multiple-choice questions and a free-text comment section (Multimedia Appendix 4). Specifically, we collected data on providers' attitudes toward DTx adoption, DTx utility and effectiveness, preference for the type of DTx, awareness of different DTx modules, patient criteria for DTx recommendation, patterns of recommendation to DTx (self-assessment regarding manner of recommendation and monthly volume), barriers to recommendation, and demographics (professional role, practice experience, service area, work hours, sex [female or male self-identified hereafter referred to as "female" or "male"], age, race or ethnicity, and work hours). The primary outcome was the percentage of KPNC providers who reported recommending DTx.

Ethical Considerations

This study received approval from the KPNC Institutional Review Board (1899252) and was determined exempt, with a waiver of consent for the EHR-based analyses and a waiver of written consent for the anonymized provider survey, which did not gather protected health information. No compensation was provided to the providers per health system policy.

Statistical Analysis

EHR-Based Characteristics of Patients Receiving Versus Not Receiving DTx Recommendations

Demographic variables, clinical characteristics, and depression and anxiety symptom burden scores (PHQ-9 and GAD-7) were compared between patients who were and were not recommended to DTx with a series of bivariate analyses. Among those who received a DTx recommendation, we also examined the relationships of mental health symptoms by sex. To examine depression and anxiety diagnoses and symptom burden associated with DTx recommendations while accounting for potential confounding factors, we performed a frequency match to create a 1:1 matched cohort of patients without a recommendation to those recommended to DTx. Patients were matched based on a 5-year age group, sex, and geographical service area within KPNC. We then conducted bivariate analyses comparing the PHQ-9 and GAD-7 using the cases and matched patients.

Provider Survey

We performed descriptive statistics (eg, frequencies and bivariate tests) to characterize provider characteristics and DTx recommendations based on multiple-choice responses. Comparisons were done based on whether respondents were "prescribers" (eg, psychiatrists and nurse practitioners and able to prescribe medications to patients) or "nonprescribers" (eg, therapists such as psychologists and marriage and family therapists) to examine potential differences by provider role.

For both the EHR-based provider recommendation and survey analyses, the Pearson or Mantel-Haenszel chi-square test or Fisher exact test was used to calculate *P* values for categorical variables, the 2-tailed *t* test was used to calculate *P* values for continuous variables when the assumptions for a parametric test assumption were met, and the Mann-Whitney *U* test was applied to calculate ordinal data when the assumptions for a nonparametric test assumption was met. All item responses were used. All analyses were performed using SAS (version 9.4) for Windows and SAS Studio (version 3.81; SAS Institute Inc).

Results

Characteristics of Patients Recommended Versus Not Recommended DTx

Of the 335,250 adults eligible during the study time period, 16% (53,546/335,250) were recommended a DTx by a provider (Table 1). Patients recommended versus not recommended to a DTx were more likely to be younger (age 18 - 44 years: 37,744/53,546, 70.5% vs 164,456/281,704, 58.4%; *P*<.001), Asian or Hispanic (20,360/53,546, 38% vs 88,827/281,704, 31.6%; *P*<.001), and female (38,338/53,546, 71.6% vs 193,063/281,704, 68.5%; *P*<.001). Additionally, there was a higher proportion of patients in the lowest (“0“) category of the Charlson Comorbidity Index (40,261/53,546, 75.2% vs 188,414,/281,704, 66.9%; *P*<.001) and a higher proportion of commercially insured patients (44,960/53,363, 84.3% vs 205,425/280,546, 73.2%; *P*<.001) among those who received recommendations versus those without a recommendation. Most DTx recommendations originated from therapy providers (48,916/53,546, 91.4%).

Table . Cohort characteristics[a,b].

Characteristic	Overall cohort (N=335,250), n (%)	No DTx[c] recommendation (n=281,704), n (%)	DTx recommendation (n=53,546), n (%)	*P* value
Age (years)				<.001
18 - 44	202,200 (60.3)	164,456 (58.4)	37,744 (70.5)	
45 - 64	92,245 (27.5)	79,668 (28.3)	12,577 (23.5)	
65 - 79	34,920 (10.4)	31,987 (11.4)	2933 (5.5)	
80+	5885 (1.8)	5593 (2)	292 (0.5)	
Race or ethnicity				<.001
Asian	39,615 (11.8)	32,039 (11.4)	7576 (14.1)	
Black	26,799 (8)	22,503 (8)	4296 (8)	
Hispanic	69,572 (20.8)	56,788 (20.2)	12,784 (23.9)	
White	172,215 (51.4)	147,744 (52.4)	24,471 (45.7)	
Other	27,049 (8.1)	22,630 (8)	4419 (8.3)	
Sex				<.001
Female	231,401 (69)	193,063 (68.5)	38,338 (71.6)	
Male	103,849 (31)	88,641 (31.5)	15,208 (28.4)	
Charlson Comorbidity Index				<.001
0	228,675 (68.2)	188,414 (66.9)	40,261 (75.2)	
1	59,561 (17.8)	50,990 (18.1)	8571 (16)	
2 - 3	29,133 (8.7)	25,874 (9.2)	3259 (6.1)	
≥4	17,881 (5.3)	16,426 (5.8)	1455 (2.7)	
Neighborhood Deprivation Index [d]				<.001
Quintile 1	67,080 (20)	57,034 (20.3)	10,046 (18.8)	
Quintile 2	66,850 (20)	56,281 (20)	10,569 (19.7)	
Quintile 3	67,106 (20)	56,260 (20)	10,846 (20.3)	
Quintile 4	67,114 (20)	56,097 (19.9)	11,017 (20.6)	
Quintile 5	66,866 (20)	55,826 (19.8)	11,040 (20.6)	
Provider type				<.001
Nurse	6331 (1.9)	6196 (2.2)	135 (0.3)	
Physician	91,733 (27.4)	88,598 (31.5)	3135 (5.9)	
Therapist	192,971 (57.6)	144,055 (51.1)	48,916 (91.4)	
Unknown	30,357 (9.1)	29,861 (10.6)	496 (0.9)	
Other	13,858 (4.1)	12,994 (4.6)	864 (1.6)	
Insurance type				<.001
Commercial	250,385 (75)	205,425 (73.2)	44,960 (84.3)	
Medicaid	32,186 (9.6)	27,820 (9.9)	4366 (8.2)	
Medicare	50,629 (15.2)	46,703 (16.6)	3926 (7.4)	
Other	709 (0.2)	598 (0.2)	111 (0.2)	
Mental health app				N/A[e]
Calm only	16,133 (4.8)	N/A	16,133 (30.1)	
Calm or myStrength	3328 (1)	N/A	3328 (6.2)	
Other app or any combined	31,342 (9.3)	N/A	31,342 (58.5)	

Characteristic	Overall cohort (N=335,250), n (%)	No DTx[c] recommendation (n=281,704), n (%)	DTx recommendation (n=53,546), n (%)	*P* value
myStrength only	2743 (0.8)	N/A	2743 (5.1)	
Missing	281,704 (84)	281,704 (100)	0 (0)	

[a]Provider referred members to DTx within 1 month of index mental health visit.

[b]DTx include Calm, myStrength, Headspace, SilverCloud, Thrive, or Whil.

[c]DTx: digital therapeutics.

[d]Neighborhood Deprivation Index quintile cut points for cohort.

[e]N/A: not applicable.

Matched Analyses—Mental Health Diagnoses and Symptoms of Patients Recommended Versus Not Recommended DTx

Given the significant demographic differences between patients recommended a DTx versus those not recommended, we performed a matched analysis to compare mental health diagnoses and symptoms between these 2 groups. Patients recommended to DTx versus not recommended had higher proportions of anxiety (22,247/53,545, 41.5% vs 19,039/53,545, 35.6%) and adjustment disorders (11,024/53,545, 20.6% vs 7592/53,545, 14.2%) and lower proportions of depression (13,151/53,545, 24.6% vs 16,517/53,545, 30.8%), bipolar (588/53,545, 1.1% vs 2685/53,545, 5%), or psychotic diagnoses (173/53,545, 0.3% vs 1239/53,545, 2.3%; all *P*<.001; Table 2). Additionally, more patients recommended to DTx had PHQ-9 depressive symptom scores in the mild (5-9: 10,643/37,282, 28.5% vs 6926/25,749, 26.9%), moderate (10-14: 10,597/37,282, 28.4% vs 7104/25,749, 27.6%), and moderate severe (15-19: 7513/37,282, 20.2% vs 5115/25,749, 19.9%) ranges compared to patients not recommended to DTx (*P*<.001 for all). The distribution of GAD-7 anxiety scores was similar between patients recommended and not recommended to DTx, albeit with the distribution of recommended patients having a slightly higher proportion of mild (5-9: 4852/23,251, 20.9% vs 2579/13,080, 19.7%) and lower proportion of severe (≥15: 9283/23,251, 39.9% vs 5391/13,080, 41.2%) GAD-7 scores (*P*=.003 for all).

Table . Mental health diagnoses and symptoms of patients recommended to digital therapeutics (DTx) and a matched cohort of patients not recommended to DTx[a,b].

Characteristic	Matched cohort (n=53,545), n (%)	Cases (n=53,545), n (%)	*P* value
Age (years)			>.99
18 - 44	37,743 (70.5)	37,743 (70.5)	
45 - 64	12,577 (23.5)	12,577 (23.5)	
65 - 79	2933 (5.5)	2933 (5.5)	
80+	292 (0.5)	292 (0.5)	
Race or ethnicity			.16
Asian	7577 (14.2)	7576 (14.1)	
Black	4465 (8.3)	4296 (8)	
Hispanic	12,512 (23.4)	12,783 (23.9)	
White	24,511 (45.8)	24,471 (45.7)	
Other	4480 (8.4)	4419 (8.3)	
Sex			>.99
Female	38,337 (71.6)	38,337 (71.6)	
Male	15,208 (28.4)	15,208 (28.4)	
Charlson Comorbidity Index			<.001
0	39,053 (72.9)	40,260 (75.2)	
1	9114 (17)	8571 (16)	
2 - 3	3575 (6.7)	3259 (6.1)	
≥4	1803 (3.4)	1455 (2.7)	
Neighborhood Deprivation Index [c]			<.001
Quintile 1	9337 (17.4)	10,045 (18.8)	
Quintile 2	9840 (18.4)	10,569 (19.7)	
Quintile 3	10,764 (20.1)	10,846 (20.3)	
Quintile 4	11,355 (21.2)	11,017 (20.6)	
Quintile 5	12,219 (22.8)	11,040 (20.6)	
Provider type			<.001
Nurse	1011 (1.9)	135 (0.3)	
Physician	13,224 (24.7)	3135 (5.9)	
Therapist	36,812 (68.7)	48,916 (91.4)	
Unknown	625 (1.2)	495 (0.9)	
Other	1873 (3.5)	864 (1.6)	
Insurance type[d]			<.001
Commercial	42,830 (80.3)	44,959 (84.3)	
Medicaid	5498 (10.3)	4366 (8.2)	
Medicare	4870 (9.1)	3926 (7.4)	
Other	124 (0.2)	111 (0.2)	
Mental health disorder diagnosis at visit			<.001
Anxiety	19,039 (35.6)	22,247 (41.5)	
Bipolar	2685 (5)	588 (1.1)	
Depression	16,517 (30.8)	13,151 (24.6)	
Other mood	7592 (14.2)	11,024 (20.6)	

Characteristic	Matched cohort (n=53,545), n (%)	Cases (n=53,545), n (%)	*P* value
Posttraumatic stress disorder	6441 (12)	6341 (11.8)	
Psychosis	1239 (2.3)	173 (0.3)	
Sleep	32 (0.1)	21 (0)	
Mental health symptom PHQ-9 [d,e]			<.001
Normal	3541 (13.8)	4063 (10.9)	
Mild	6926 (26.9)	10,643 (28.5)	
Moderate	7104 (27.6)	10,597 (28.4)	
Moderate severe	5115 (19.9)	7513 (20.2)	
Severe	3063 (11.9)	4466 (12)	
Mental health symptom GAD-7 [d,f]			.003
Normal	508 (3.9)	857 (3.7)	
Mild	2579 (19.7)	4852 (20.9)	
Moderate	4602 (35.2)	8259 (35.5)	
Severe	5391 (41.2)	9283 (39.9)	

[a]Provider referred members to DTx within 1 month of index mental health visit.

[b]DTx includes Calm, myStrength, Headspace, SilverCloud, Thrive, or Whil.

[c]Neighborhood Deprivation Index reports quintile cut points for our specific study population.

[d]Section does sum to a cohort of recommended patients because of missing item data.

[e]Patient Health Questionnaire 9-Item depression scale. Each item is scored 0 to 3, providing a 0 to 27 severity total score. Using the total score to represent the cut point for depression, 0 - 4 is normal, 5 - 9 is mild, 10 - 14 is moderate, 15 - 19 is moderate severe, and 20 or more is severe.

[f]Generalized Anxiety Disorder 7-Item scale. Each item is scored 0 to 3, providing a 0 to 21 severity total score. Using the total score to represent the cut point for generalized anxiety disorder, 0 - 4 is normal, 5 - 9 is mild, 10 - 14 is moderate, and 15 or more is severe.

Mental Health Symptom Burden by Sex in DTx-Recommended Patients

Since approximately two-thirds of patients recommended to DTx were female, we conducted ad hoc analyses examining symptom burden stratified by sex among those who received a recommendation to DTx. Female patients recommended to DTx had higher proportions of PHQ-9 depressive symptom scores in the moderate (7575/26,416, 28.7% vs 3022/10,867, 27.8%), moderate severe (5414/26,416, 20.5% vs 2099/10,867, 19.3%), and severe (3246/26,416, 12.3% vs 1220/10,867, 11.2%) range compared to male patients; and male patients had higher proportions of PHQ-9 scores in the normal (2804/26,416, 10.6% vs 1260/10,867, 11.6%) and mild (7377/26,416, 27.9% vs 3266/10,867, 30.1%) ranges (*P*=.003; Table 3). Female patients recommended to DTx had higher proportions of GAD-7 scores in the severe (6839/16,526, 41.4% vs 2444/6725, 36.3%) range compared to male patients; and male patients had more representation in the normal (516/16,526, 3.1% vs 341/6725, 5.1%) and mild (3296/16,526, 19.9% vs 1556/6725, 23.1%) ranges for GAD-7 (*P*<.001).

Table . Mental health symptoms by sex among patients recommended digital therapeutics (DTx)[a,b].

Characteristic	Overall cohort (N=53,546), n (%)	Female patients (n=38,337), n (%)	Male patients (n=15,208), n (%)	*P* value
Mental health symptom PHQ-9 [c, d]				.003
Normal	4064 (10.90)	2804 (10.6)	1260 (11.6)	
Mild	10,643 (28.55)	7377 (27.9)	3266 (30.1)	
Moderate	10,597 (28.42)	7575 (28.7)	3022 (27.8)	
Moderate severe	7513 (20.15)	5414 (20.5)	2099 (19.3)	
Severe	4466 (11.98)	3246 (12.3)	1220 (11.2)	
Mental health symptom GAD-7 [d, e]				<.001
Normal	857 (3.69)	516 (3.1)	341 (5.1)	
Mild	4852 (20.87)	3296 (19.9)	1556 (23.1)	
Moderate	8259 (35.52)	5875 (35.6)	2384 (35.4)	
Severe	9283 (39.93)	6839 (41.4)	2444 (36.3)	

[a]Member recommended DTx within 1 month of index mental health visit.

[b]DTx includes Calm, myStrength, Headspace, SilverCloud, Thrive, or Whil.

[c]Patient Health Questionnaire 9-Item depression scale. Each item is scored 0 to 3, providing a 0 to 27 severity total score. Using the total score to represent the cut point for depression, 0 - 4 is normal, 5 - 9 is mild, 10 - 14 is moderate, 15 - 19 is moderate severe, and 20 or more is severe.

[d]Columns do not sum to the respective cohort because of missing data.

[e]Generalized Anxiety Disorder 7-Item scale. Each item is scored 0 to 3, providing a 0 to 21 severity total score. Using the total score to represent the cut point for generalized anxiety disorder, 0 - 4 is normal, 5 - 9 is mild, 10 - 14 is moderate, and 15 or more is severe.

Characteristics of Providers Recommending DTx

A total of 211 individuals responded to the provider survey on DTx prescribing patterns for a participation rate of 8.4% (211/2505). Of those, 19.6% (40/204) identified as prescribers, 68.6% (140/204) as nonprescribers, and the remainder as having an administrative or managerial role in the clinic. Of respondents, 91.9% (192/209) were between the ages of 30 and 59 years, 54.1% (113/209) identified as White, 17.2% (36/209) as Asian, 12% (25/209) as Hispanic or Latino, and 6.2% (13/209) as Black. Nearly all respondents reported working at least half-time, and two-thirds of respondents reported working a full-time schedule.

Provider DTx Recommendation Behaviors

Overall, a higher proportion of nonprescribers (127/164, 77.4%) indicated they recommend DTx to patients compared to prescribers (22/39, 56.4%; *P*=.008). Of providers who recommend DTx, 83.6% (122/146) do so via an electronic secure message to patients, and only 12.3% (18/146) do so during a clinic appointment. Calm was the most commonly recommended DTx, followed by myStrength, Headspace, SilverCloud, Thrive, and Whil. Only 29.4% (45/153) of providers indicated using the suggested internal EHR–based smart text to recommend DTx. Prescribers were more likely than nonprescribers to report that they did not know how to make patient recommendations using EHR-based protocols (*P*=.01). Of providers recommending DTx to patients, 64.1% (98/153) report that they follow up with patients to inquire whether the resource was helpful. Only 38.4% (58/151) of providers recommend specific modules of DTx resources, and of those 58.6% (34/58) report following up with patients on these specific recommendations.

On a scale of 0 - 10 regarding the perceived effectiveness of DTx resources to impact symptoms (0=none and 10=extreme), both prescribers and nonprescribers rated cognitive behavioral therapy and mindfulness content between 6 and 7 or moderate to highly effective.

Discussion

Principal Findings

These are some of the first pragmatic results regarding provider recommendation patterns, characteristics, and attitudes toward mental health DTx. In this large, integrated health system with mental health DTx tools included as a benefit, a very modest percentage of patients with mental health disorders received recommendations to DTx tools. Those who did were more likely to be young, female, Asian or Hispanic, with less medical comorbidity, and to have commercial insurance compared to patients not recommended to DTx. Recommendations for DTx were more often made for anxiety or adjustment diagnoses and for patients with less severe mental health symptoms, although female patients recommended to DTx tools had a higher proportion of anxiety symptoms compared to male patients. A provider survey showed that DTx are recommended more frequently by nonprescribers and via electronic secure messages rather than during clinical appointments. Of providers recommending DTx, a modest majority follow up with the patient regarding DTx effectiveness, and few knew how to follow the recommended EHR protocols to do so.

Previous studies reporting demographic characteristics of patients using DTx have been from pilot trials or RCTs [7-9]. Using real-world data collected in the course of clinical care, we found that patient characteristics were similar to those of DTx trials. Overall, patients who received DTx recommendations were younger with less severe mental health symptoms, suggesting that providers are likely tailoring their recommendations to those they feel are most suitable. This may also reflect the initial emphasis of the health system DTx implementation focusing on patients with mild to moderate symptoms of depression or anxiety [20]. These results highlight a gap in the evidence base regarding the effectiveness of DTx for patients with mild to severe mental health symptoms or multiple comorbidities.

Asian or Hispanic patients were more likely to receive DTx recommendations. There are many potential drivers of this finding, including provider beliefs about patient DTx uptake, patient ability to interface with DTx, if providers felt the symptom severity reported by these patients was more appropriate for DTx, or other unrecognized factors. This finding merits further investigation. Of note, the significant association for insurance type with a DTx recommendation suggests slightly more patients with commercial insurance receiving recommendations. This may reflect provider-perceived barriers to care (eg, differences in disease severity or mobile phone access) for patients with other types of insurance coverage.

The finding that a higher proportion of female patients with severe mental health symptoms were referred to DTx compared to male patients could reflect a bias to implement more tools to treat symptoms in female patients, a reporting bias of symptoms by male or female patients, refusal of the DTx recommendation by male patients, or other sex-associated biases in care [26]. These findings warrant further research.

Provider survey results indicate that a higher proportion of nonprescribers recommend DTx and reported familiarity with EHR-based protocols to recommend DTx compared to prescribers. These results could reflect uneven training by provider type during the initial rollout or a lack of awareness or interest in the DTx tools by prescribers. Nonprescribers may appreciate having another treatment option to recommend beyond their typical strategies, given the therapeutic role DTx can fill and the silos of mental health care (ie, "therapy" vs "medication management" tasks). Interestingly, a minority of providers reported using the EHR-based protocols to recommend DTx to patients, with very few recommending specific modules or follow-up on module recommendations. Providers may view DTx tools as a passive, adjunctive resource rather than part of an active mental health treatment plan. Further, only a modest majority of providers reported asking patients if they found the DTx effective. A similar follow-up rate for a recommended or prescribed medical device, medication, or therapeutic program such as intensive outpatient or partial hospitalization would likely be regarded as very low. It is unclear why providers perceive DTx offerings in this way; it is possible that providers do not see DTx as having robust effectiveness data, which may impact their rate of follow-up. Similarly, providers may not view DTx resources as part of clinical care or within their scope of practice to monitor. This may reflect the early stage of growth and implementation of DTx in mental health and may evolve over time. If DTx are to be integrated into clinical mental health care, it may be more effective as a formalized part of the treatment plan and followed in the same way as other treatment aspects.

Limitations

This study was conducted in an integrated health care system, which limits generalizability to providers and patients in different systems of care. The survey participation rate was low, given constraints related to direct dissemination and reapproaching for response, though consistent with reported response rates in similar investigations [15]. While our integrated health system made multiple DTx available at no cost to patients during the study time period, such resources may not be as easily available to patients in other health care systems. Recommendations that were in person only or sent in messages without using the EHR protocols and smart text were not captured, nor were recommendations related to later visits, although our intent was to identify options presented earlier in the treatment course. The large sample size of EHR-based recommendation data makes it likely to detect associations that may have statistical significance but may not have meaningful clinical implications. The data we explored are observational and exploratory; future work warrants further examination into these initial findings along with an examination of patient-level outcomes associated with DTx.

Conclusions

The use of DTx is growing at a rapid pace, and health systems and patients see these as potentially valuable resources. However, the extent of provider recommendations to DTx is modest, as is provider knowledge about DTx and their follow-up with patients about DTx use, which suggests a largely passive DTx uptake by providers. Given the likely continued interest by patients, health care systems, and industry, future research on how to effectively implement these tools is critical.

Acknowledgments

This project was supported by The Permanente Medical Group Delivery Science and Applied Research initiative. The sponsors were not involved in the study design; collection, analysis, or interpretation of the data; writing of the manuscript; or decision to submit the paper for publication. The authors also gratefully acknowledge their survey respondents. The authors thank Doug Corley, MD, for his early consultation on study design. No generative artificial intelligence was used in the development of this study or writing of the manuscript.

Data Availability

The data sets used and analyzed during this study are not publicly available due to reasons of sensitivity but are available from the corresponding author on reasonable request and with appropriate data sharing agreements and institutional review board approval.

Authors' Contributions

SJR was responsible for funding acquisition and conceptualization. SJR, CIC, and TYL were responsible for developing the study design. SJR developed and fielded the survey. SJR, CIC, and TYL developed the statistical analysis plan. TYL prepared the data sets and conducted the statistical analyses. SJR wrote the initial draft of the paper. All authors participated in the review and revision process and approved the submission.

Conflicts of Interest

CIC has received support managed through their institution from the Industry PMR Consortium, a consortium of companies working together to conduct postmarketing studies required by the Food and Drug Administration that assess risks related to opioid analgesic use. The remaining authors declare no conflicts of interest.

Multimedia Appendix 1
Cohort of Kaiser Permanente Northern California psychiatry patients receiving digital therapeutics recommendation.
[PPTX File, 50 KB - mental_v11i1e56574_app1.pptx]

Multimedia Appendix 2
International Classification of Diseases codes for mental health conditions.
[DOCX File, 21 KB - mental_v11i1e56574_app2.docx]

Multimedia Appendix 3
Data sources and search terms used to capture digital health recommendations.
[DOCX File, 21 KB - mental_v11i1e56574_app3.docx]

Multimedia Appendix 4
Provider survey.
[PDF File, 1754 KB - mental_v11i1e56574_app4.pdf]

Checklist 1
STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) checklist.
[DOCX File, 41 KB - mental_v11i1e56574_app5.docx]

Checklist 2
CHERRIES (Checklist for Reporting Results of Internet E-Surveys).
[DOCX File, 27 KB - mental_v11i1e56574_app6.docx]

References

1. Digital transformation market size, share & trends analysis report by solution (analytics, mobility), by deployment, by service, by enterprise size, by end-use, by region, and segment forecasts, 2023 - 2030. Grand View Research. 2023. URL: https://www.grandviewresearch.com/industry-analysis/digital-transformation-market [accessed 2023-12-02]
2. 1H and Q2 2021 digital health funding and M&A report. Mercom Capital Group. 2023. URL: https://mercomcapital.com/product/1h-q2-2021-digital-health-healthcare-it-funding-ma-report [accessed 2024-08-19]
3. Zhong S, Yang X, Pan Z, et al. The usability, feasibility, acceptability, and efficacy of digital mental health services in the COVID-19 pandemic: scoping review, systematic review, and meta-analysis. JMIR Public Health Surveill 2023 Feb 13;9:e43730. [doi: 10.2196/43730] [Medline: 36634261]
4. Aziz M, Erbad A, Almourad MB, Altuwairiqi M, McAlaney J, Ali R. Did usage of mental health apps change during COVID-19? A comparative study based on an objective recording of usage data and demographics. Life (Basel) 2022 Aug 19;12(8):1266. [doi: 10.3390/life12081266] [Medline: 36013444]
5. Patel S, Akhtar A, Malins S, et al. The acceptability and usability of digital health interventions for adults with depression, anxiety, and somatoform disorders: qualitative systematic review and meta-synthesis. J Med Internet Res 2020 Jul 6;22(7):e16228. [doi: 10.2196/16228] [Medline: 32628116]

6. Satre DD, Meacham MC, Asarnow LD, Fisher WS, Fortuna LR, Iturralde E. Opportunities to integrate mobile app-based interventions into mental health and substance use disorder treatment services in the wake of COVID-19. Am J Health Promot 2021 Nov;35(8):1178-1183. [doi: 10.1177/08901171211055314] [Medline: 34652971]
7. Firth J, Torous J, Nicholas J, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry 2017 Oct;16(3):287-298. [doi: 10.1002/wps.20472] [Medline: 28941113]
8. Firth J, Torous J, Nicholas J, Carney R, Rosenbaum S, Sarris J. Can smartphone mental health interventions reduce symptoms of anxiety? A meta-analysis of randomized controlled trials. J Affect Disord 2017 Aug 15;218:15-22. [doi: 10.1016/j.jad.2017.04.046] [Medline: 28456072]
9. Tan ZYA, Wong SH, Cheng LJ, Lau ST. Effectiveness of mobile-based mindfulness interventions in improving mindfulness skills and psychological outcomes for adults: a systematic review and meta-regression. Mindfulness 2022 Oct;13(10):2379-2395. [doi: 10.1007/s12671-022-01962-z]
10. Huberty J, Green J, Glissmann C, Larkey L, Puzia M, Lee C. Efficacy of the mindfulness meditation mobile app “Calm” to reduce stress among college students: randomized controlled trial. JMIR Mhealth Uhealth 2019 Jun 25;7(6):e14273. [doi: 10.2196/14273] [Medline: 31237569]
11. Hirsch A, Luellen J, Holder JM, et al. Managing depressive symptoms in the workplace using a web-based self-care tool: a pilot randomized controlled trial. JMIR Res Protoc 2017 Apr 4;6(4):e51. [doi: 10.2196/resprot.7203] [Medline: 28377368]
12. Koh J, Tng GYQ, Hartanto A. Potential and pitfalls of mobile mental health apps in traditional treatment: an umbrella review. J Pers Med 2022 Aug 25;12(9):1376. [doi: 10.3390/jpm12091376] [Medline: 36143161]
13. Linardon J, Torous J, Firth J, Cuijpers P, Messer M, Fuller-Tyszkiewicz M. Current evidence on the efficacy of mental health smartphone apps for symptoms of depression and anxiety. a meta-analysis of 176 randomized controlled trials. World Psychiatry 2024 Feb;23(1):139-149. [doi: 10.1002/wps.21183] [Medline: 38214614]
14. Kong T, Scott MM, Li Y, Wichelman C. Physician attitudes towards—and adoption of—mobile health. Digit Health 2020 Feb;6:2055207620907187. [doi: 10.1177/2055207620907187] [Medline: 32128235]
15. Sezgin E, Özkan-Yildirim S, Yildirim S. Investigation of physicians’ awareness and use of mHealth apps: a mixed method study. Health Policy Technol 2017 Sep;6(3):251-267. [doi: 10.1016/j.hlpt.2017.07.007]
16. van Kessel R, Roman-Urrestarazu A, Anderson M, et al. Mapping factors that affect the uptake of digital therapeutics within health systems: scoping review. J Med Internet Res 2023 Jul 25;25:e48000. [doi: 10.2196/48000] [Medline: 37490322]
17. Borges do Nascimento IJ, Abdulazeem H, Vasanthan LT, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. NPJ Digit Med 2023 Sep 18;6(1):161. [doi: 10.1038/s41746-023-00899-4] [Medline: 37723240]
18. Dahlhausen F, Zinner M, Bieske L, Ehlers JP, Boehme P, Fehring L. Physicians’ attitudes toward prescribable mHealth apps and implications for adoption in Germany: mixed methods study. JMIR Mhealth Uhealth 2021 Nov 23;9(11):e33012. [doi: 10.2196/33012] [Medline: 34817385]
19. Gordon N. Similarity of adult Kaiser Permanente members to the adult population in Kaiser Permanente’s Northern California service area: comparisons based on the 2017/2018 cycle of the California Health Interview Survey. Kaiser Permanente Division of Research. 2020. URL: https://memberhealthsurvey.kaiser.org/Documents/compare_kp_ncal_chis2017-18.pdf [accessed 2024-08-27]
20. Mordecai D, Histon T, Neuwirth E, et al. How Kaiser Permanente created a mental health and wellness digital ecosystem. NEJM Catal 2021 Jan;2(1). [doi: 10.1056/CAT.20.0295]
21. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Epidemiology 2007 Nov;18(6):800-804. [doi: 10.1097/EDE.0b013e3181577654] [Medline: 18049194]
22. Eysenbach G. Improving the quality of web surveys: the Checklist for Reporting Results of Internet E-Surveys (CHERRIES). J Med Internet Res 2004 Sep 29;6(3):e34. [doi: 10.2196/jmir.6.3.e34] [Medline: 15471760]
23. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987;40(5):373-383. [doi: 10.1016/0021-9681(87)90171-8] [Medline: 3558716]
24. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613. [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
25. Löwe B, Decker O, Müller S, et al. Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. Med Care 2008 Mar;46(3):266-274. [doi: 10.1097/MLR.0b013e318160d093] [Medline: 18388841]
26. Pattyn E, Verhaeghe M, Bracke P. The gender gap in mental health service use. Soc Psychiatry Psychiatr Epidemiol 2015 Jul;50(7):1089-1095. [doi: 10.1007/s00127-015-1038-x] [Medline: 25788391]

Abbreviations

CHERRIES: Checklist for Reporting Results of Internet E-Surveys
DTx: digital therapeutics
EHR: electronic health record

GAD-7: Generalized Anxiety Disorder 7-Item
***ICD-10*:** *International Classification of Diseases, 10th Revision*
KPNC: Kaiser Permanente Northern California
PHQ-9: Patient Health Questionnaire 9-Item
RCT: randomized controlled trial
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by J Torous; submitted 20.01.24; peer-reviewed by JO Wrenn, M Niwa; revised version received 19.06.24; accepted 26.06.24; published 02.10.24.

Please cite as:
Ridout SJ, Ridout KK, Lin TY, Campbell CI
Clinical Use of Mental Health Digital Therapeutics in a Large Health Care Delivery System: Retrospective Patient Cohort Study and Provider Survey
JMIR Ment Health 2024;11:e56574
URL: https://mental.jmir.org/2024/1/e56574
doi:10.2196/56574

A Web-Based and Mobile Intervention Program Using a Spaced Education Approach for Workplace Mental Health Literacy: Cluster Randomized Controlled Trial

Lawrence T Lam[1,2,*], BSc, MAppPsy, MPH, PhD; Mary K P Lam[3,*], BA, MHIM, PhD

1

2

3

*all authors contributed equally

Corresponding Author:
Lawrence T Lam, BSc, MAppPsy, MPH, PhD

Abstract

Background: Workplace mental health is an important global health concern.

Objectives: This unblinded, phase-III, wait-listed cluster randomized controlled trial aimed to examine the effectiveness of a mobile health (mHealth) psychoeducation program using a spaced education approach on mental health literacy (MHL) in the workplace. The main interest of this paper was the immediate and 3-month medium-term effect of the program on the MHL of workers. The purposely built mHealth platform was also evaluated as a health-related app.

Methods: The mHealth platform was designed using the principle of spaced education as a psychoeducation intervention program, with various modules of web-based and mobile materials presented to the participant in a progressive manner. Short quizzes at the end of each module ensured adequate learning, and successful completion qualified the learner to progress to the next level. The trial recruited 456 employees of specific industries with high levels of work-related stress. Participants who were nested in different offices or units were allocated into the intervention and wait-listed control groups using a block randomization process, with the office or unit as the cluster. A separate sample of 70 individual raters were used for the evaluation of the mHealth platform. The Australian National MHL and Stigma Survey and the Mobile Apps Rating Scale were completed through a web-based self-reported survey to assess MHL and evaluate the app. The trial and follow-up data were analyzed by a generalized linear latent and mixed model with adjustments for the clustering effect of work sites and repeated measures.

Results: Of the 456 participants in the trial, 236 (51.8%) responded to the follow-up survey. Most MHL outcomes obtained significant results immediately after the intervention and across time. After adjusting for the clustering effect, the postintervention weighted mean scores were significantly higher in the intervention group than the control group for correct recognition of a mental health problem, help seeking, and stigmatization by 0.2 (SE 0.1; P=.003), 0.9 (SE 0.2; P<.001), and 1.8 (SE 0.4; P<.001), respectively. After adjusting for the clustering effect, significant differences across time were found in help-seeking intention (P=.01), stigmatization (P<.001), and social distancing (P<.001). The evaluation of the mHealth program resulted in average scores of the 4 major domains ranging from 3.8 to 4.2, with engagement having the lowest score.

Conclusions: The mHealth psychoeducation intervention program using this platform had immediate and 3-month medium-term effects of retaining and improving MHL. The platform was evaluated to have satisfactory performance in terms of functionality, aesthetics, information content, and utility in enhancing MHL. It is anticipated that ongoing development in digital health will provide great benefits in improving the mental health of the global population.

Trial Registration: Australian New Zealand Clinical Trials Registry (ANZCTR) ACTRN12619000464167; https://www.anzctr.org.au/Trial/Registration/TrialReview.aspx?id=377176

International Registered Report Identifier (IRRID): RR2-10.1186/s13063-019-3748-y

(JMIR Ment Health 2024;11:e51791) doi:10.2196/51791

KEYWORDS

mHealth; web-based intervention; mental health literacy; psychoeducation; randomized controlled trial; workplace; performance; worker; intervention; digital health; mental wellness; promote; well-being; mobile health; technology

Introduction

Mental ill health has long been recognized as an important global health problem [1]. The workplace has also been identified as an important venue for preventing mental health problems and for promoting mental wellness [2,3]. Preventive programs designed as a workplace strategy could provide benefits to workers in terms of early identification and intervention of mental health problems, as well as promoting mental well-being [4]. One specific approach is to enhance workers' mental health literacy (MHL). MHL was defined as "knowledge and beliefs about mental disorders which aid their recognition, management or prevention" by Jorm [5]. In this construct, there are different aspects, including the ability to recognize specific disorders, knowledge about mental health, attitudes toward help seeking and stigmatization, and social distancing from people with mental health problems.

In response to the urgent need for a well-designed workplace mental health intervention program in Hong Kong, a group of researchers developed the mobile health (mHealth) web-based and mobile Workplace Mental Health Literacy (WPMHL) project with funding support from the Hong Kong government [6]. The program consists of 2 main components: psychoeducation modules for the enhancement of workers' MHL and a work environment scan addressing more structural issues in the workplace [6].

The psychoeducation modules of the program incorporated elements of mental health first aid in the design but focused on the workplace environment, addressing common workplace mental health issues such as work-related stress and burnout [6]. The underpinning educational paradigm or approach of the proposed psychoeducation intervention program is called the spaced learning or spaced education approach. Based on the initial neuropsychological concept proposed by the internationally acclaimed neuroscientist Fields [7] and modified and advocated by Kelly [8], the spaced learning approach is a learning methodology for creating long-term memories. This concept is operationalized by presenting highly condensed learning materials repeatedly, based on a predesigned temporal pattern and a number of presentations, to allow the individual to encode the material into one's memory system. The repetition of the process will reinforce the encoding mechanism and, in turn, commit the materials into long-term memory [8]. Applying this empirically validated approach to education and training, the learning materials could be presented repeatedly in a temporally sequential manner to allow the learner to generate long-term memory. In the design of the psychoeducation component of the intervention program, the principle of spaced education was applied, with various modules of web-based and mobile materials presented to the participant in a progressive manner and some simple exercises presented at the end of each module. Successful completion of 1 module gained access to the next progressively until all modules were completed. Within each module, essential materials were presented repeatedly in a short duration with a temporal pattern of 3 times in a row for the participant to create a long-term memory. The spaced learning or spaced education approach has been used in different fields of training and education, such as continuous medical education [9,10]. Different versions of the main interface, as a website for PCs and as a downloadable mobile app for tablets and smartphones, are presented in Figure 1.

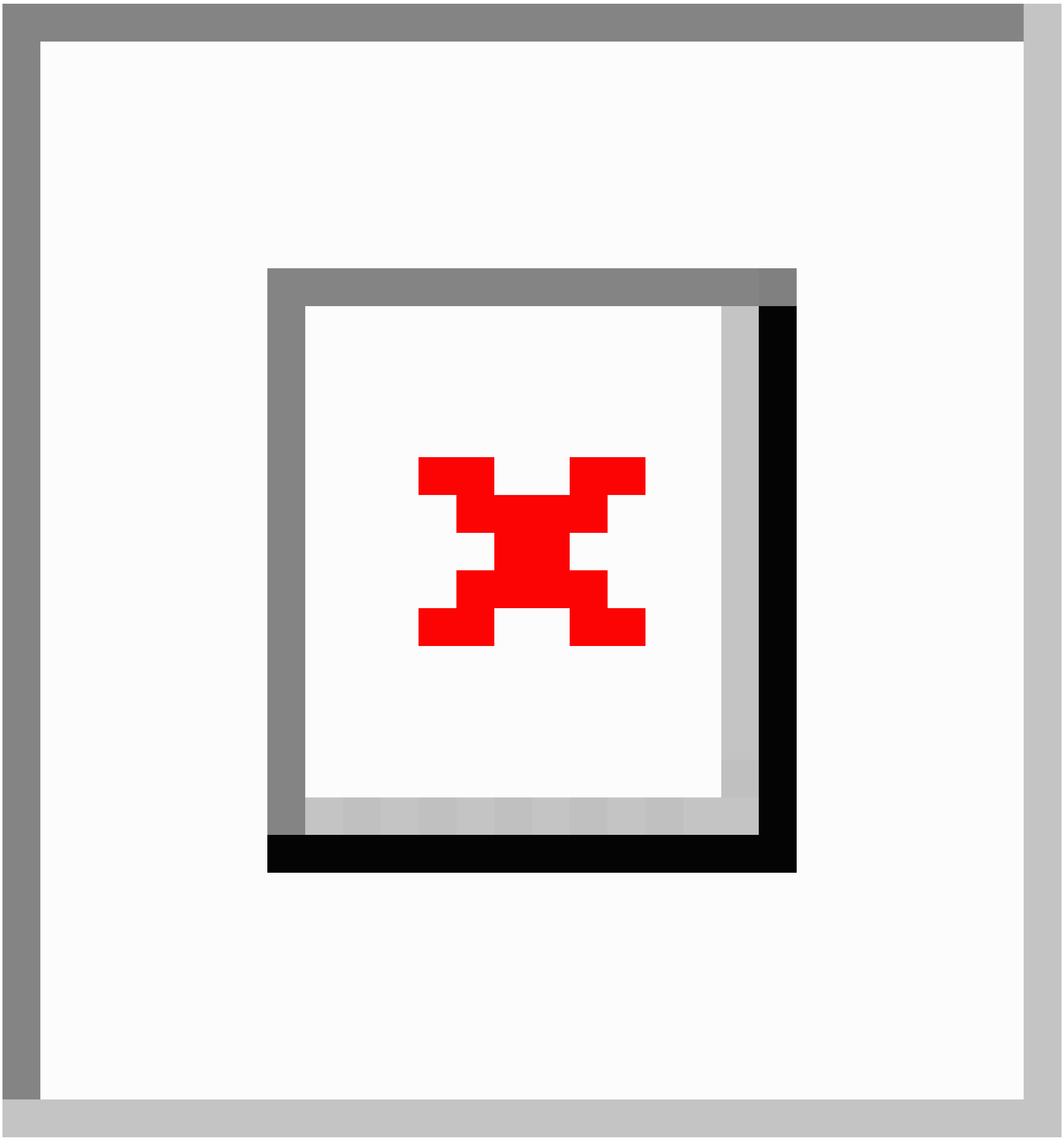
mHealth is defined by the World Health Organization (WHO) as technology "used for medical and public health practice supported by mobile devices, such as mobile phones, patient monitoring devices, Personal Digital Assistants (PDAs), and other wireless devices" [11]. With the rapid development of mobile technologies and further advancement in telecommunication platforms in recent years, mHealth has been adopted as one of the widely used approaches in preventive medicine across different diseases [12-16]. For example, mHealth was used for the promotion of a healthy lifestyle in the prevention of cardiovascular diseases and diabetes [13,16], sexually transmittable infections [9], and cancer [8]. In the prevention of mental health problems, the mHealth approach was also adopted for alcohol and other substance abuse, suicide prevention, and other problems [17-20].

In terms of the effectiveness of the mHealth approach on mental health problems, review studies found some evidence for a positive effect on the reduction of risk factors, such as a reduction in depressive symptoms and stress and an increase in coping, but not on the actual behavior [17,18]. This might be due to the small number of studies included in the reviews since mHealth is still a relatively novel approach. There is also a concern that a number of the studies under review had methodological issues or a lack of proper evaluation that might have contributed to the nonsignificant results obtained [21]. As such, it was recommended that more methodologically robust research should be implemented to further ascertain the efficacy of intervention programs using the mHealth approach [21].

To evaluate the efficacy of the mHealth web-based and mobile WPMHL intervention program, a wait-listed cluster randomized controlled trial (CRCT) was conducted, with MHL as the primary outcome and work-related stress and burnout as the secondary outcomes. The immediate effect of the intervention program was described in a previous report, suggesting a significant result on stress and burnout in favor of the intervention [22]. In terms of MHL, the intervention group scored significantly higher than the control group in the correct recognition of mental problems, help seeking, and stigmatization by 0.2 (SE 0.1; *P*=.003), 0.9 (SE 0.2; *P*<.001), and 1.8 (SE 0.4; *P*<.001), respectively [22]. These results provided evidence for the immediate effect of the psychoeducation component of the program in improving the mental health of the participants [22].

In this study, we aim to focus on both the immediate and 3-month medium-term effects of the intervention program on the MHL of workers. We also aim to report the results obtained for the evaluation of the mHealth psychoeducation platform, particularly engagement, functionality, aesthetics, and the information provided. It is anticipated that the psychoeducation modules will be efficacious in improving MHL immediately and that the effect will remain in the medium term.

Figure 1. Different versions of the main interface: as a website for PCs (left) and as a downloadable app for tablets and smartphones (right). MHFA: mental health first aid.



Methods

Study Design and Target Population

As aforementioned, the protocol for this phase-III, wait-listed CRCT targeting the workplace was reported previously [6]. In brief, 6 large corporations were invited to participate in the trial, and employees were recruited through their corresponding human resources departments. Voluntary participation of recruits was ensured without any influence from the companies' management. The recruited sample consisted of a wide range of work natures, ranging from manual labor to senior executives, since the business of these corporations covered a multitude of work types. The trial was conducted between March and December 2021 with a 3-month follow-up phase in both intervention and control arms (Checklist 1).

Recruitment of Participants and Randomization

The primary unit of randomization was the different offices or units of the participating corporations, and recruited workers were clustered in each office or unit. Potential participants were screened for their eligibility at the point of recruitment. Workers who had been exposed to any similar psychoeducation training were excluded from the trial. To randomly allocate the offices and units, the human resources departments of these corporations provided a list of participating offices and units with some basic staffing information, such as the number of staff in the offices and units and their positions. A qualified statistician, who was blinded to the process of recruitment and the subsequent operation and data collection, conducted the randomization using a block randomization process. Each cluster with participants nested in different offices or units was then allocated to either the intervention or wait-listed control groups. Once the cluster had been randomized, participants completed the baseline data collection using the built-in feature of the web-based platform. Postintervention data collection took place immediately after the completion of the psychoeducation modules and at the 3-month follow-up.

Intervention and the Platform

Details of the intervention program were described in the *Introduction* section. Regarding the mHealth platform, as aforementioned, it was specifically designed for this project by a software company. It was developed to be adaptable for multiple interfaces with web (PC) and app (tablet and smartphone) versions with identical content and functionalities. For the study, participants were provided with a unique log-in to a central server for accessing the platform upon random allocation. Participants who were allocated to the wait-listed control group could activate their log-in when the group changed from the control phase to the intervention phase. Participants could access the web-based platform directly through the server, or they could download the mobile app from the server to be used on their tablets or smartphones for both Android and iOS.

Study Outcome and Outcome Measures

For the outcome measures of this study, namely MHL and the evaluation of the mHealth platform, 2 instruments were used. For MHL, the Australian National MHL and Stigma Survey was used [23]. The instrument has been validated and widely used in many studies in different countries [24]. In this study, the full survey was not used, given that it is a lengthy instrument and comprises many different submodules, with each module adopting a validated scale that can be used separately. After considering the specific local context, some components of the instrument were selected. These included the correct identification of mental health problems, help seeking for a mental health problem, stigmatization, and social distancing. The responses of these scales were set in a positive direction such that a higher score represented a higher level of the construct. Therefore, a lower score on the stigmatization and social distancing scales reflected a lower level in both constructs. In terms of the evaluation of the mHealth platform, the Mobile App Rating Scale (MARS) was used to assess the quality of different aspects of the platform, including engagement, functionality, aesthetics, and the information provided [25]. The MARS is an instrument designed to assess the quality of mHealth apps developed for addressing health problems or improving health status [25]. As such, it includes an app-specific domain for assessing specific health literacy. It has been validated with a good internal consistency with a Cronbach α value of .90 and a moderately high interrater reliability with an intraclass correlation coefficient of 0.79 [25]. It has also been widely used for evaluating mHealth apps worldwide and translated into different languages [26,27]. Both the MARS and MHL assessments were self-reported, and data were collected through a web-based self-reported survey at different time points. Participants' information was also collected. This included demographics; employment status; and health behaviors, such as drinking, smoking, and physical activity.

Procedures

As per the protocol of a wait-listed CRCT, participants in the control arm were offered the same intervention upon the completion of the intervention program by the intervention arm. Hence, all participants received the same intervention treatment eventually. As a result, data collection on MHL was planned to be conducted at 3 different time points: at baseline, immediately after the completion of the program, and 3 months after the completion of the program. Participants were invited to respond to the data collection survey 3 months after the completion of the intervention program with multiple waves of invitations through emails. For the evaluation of the app, a small group of raters was recruited from a different population of workers who were not involved in the trial. They were invited to use the app and then evaluate it using the web-based questionnaire that included the MARS. Training was provided to these raters prior to their exposure to the platform. The raters were invited to rate the platform from a user's perspective in order to mimic the user's experience.

Sample Size and Data Analysis

Based on the assumed effect size of an approximate 0.5 SD difference in the MHL scores between the intervention and control arms, 80% power of the study, a type I error rate of 5%, and an interclass correlation of 0.01, it was estimated that about 400 workers were required. It was also assumed that 10% of participants would drop out of the project. Data were analyzed using the statistical software program Stata BE18.0 (StataCorp LLC). Data obtained from the MARS were analyzed descriptively, with means and SDs or SEs calculated. As suggested by the authors of the scale, the rating was designed as a 5-point Likert scale with 1=inadequate, 2=poor, 3=acceptable, 4=good, and 5=excellent [25]. The mean score of each aspect of the platform was interpreted following the suggested levels of quality and specific health literacy. For MHL, the analysis focused on the efficacy of the intervention program and the changes in outcome measures across time. For investigating the efficacy of the intervention program, comparisons of the mean scores of the outcome measures between the intervention and control groups after the intervention program, adjusting for the clustering effect, and the baseline assessment of the outcome measures were conducted. Given that participants were nested in different clusters and the outcome variables were measured repeatedly, these factors were taken into consideration for the analyses. As a result, a generalized linear latent and mixed model was applied to test the time effect while adjusting for the clustering effect. To handle any loss to follow-up, the main analyses were conducted according to the intention-to-treat principle, and missing data in any variables were imputed using the multiple imputation approach with an assumption of missing at random for all missing values. Since all respondents should have received the intervention at follow-up, including those in the control group, the focus of the comparison was mainly on the change of MHL scores across time with adjustment for the clustering effect. A type I error rate of 5% was adopted for the testing of hypotheses.

Ethical Considerations

Participation in the trial was voluntary, and no compensation was provided. Potential participants were provided with information on the trial before participation through the corporations' human resource departments. Willing participants were enrolled in the trial through direct, personal contact with the research team. Informed consent was implied when the participant logged in to the web-based platform. All participants were free to opt out at any time. Confidentiality and privacy were ensured with minimal personal information collected for enrollment purposes, and this information was stored separately on a password-protected database from the trial data. All data collected from the trial were deidentified and stored on a double-layered, password-protected database. The study obtained human ethics approval from the Human Research Ethics Committee of the Tung Wah College (approval REC2018020). Trial registration was also completed with the Australian New Zealand Clinical Trials Registry (ANZCTR; registration ACTRN12619000464167).

Results

Primary Outcome of MHL

The results obtained from the wait-listed CRCT, including the demographic and other health-related information of the full sample, had been reported previously [22]. Interested readers should refer to the results in the published paper. In brief, the mean age of the 456 participants was 40.7 (SD 9.8) years, with 215 (47.2%) being male and the majority (n=363, 79.6%) having attained an education level of university or above. Slightly more than half (n=271, 59.4%) were married, and nearly all (n=454, 99.6%) worked full-time. In terms of their health, the majority (n=344, 76.6%) did not exercise regularly; however, very few were reported to be regular smokers (n=17, 3.6%) and drinkers (n=8, 1.8%). A total of 229 (50.2%) participants were randomized to receive the intervention program first, and the rest (n=227, 49.8%) were wait-listed controls providing data for the analyses for the immediate effect. As reported previously, comparisons of the demographics and health-related variables at baseline between groups found no significant differences at all (all P>.05). The results obtained on the immediate effect of the psychoeducation program on MHL are summarized in Table 1. As shown, significant differences between groups were observed in all domains except social distancing. After adjusting for the clustering effect and the potential confounding factors, the postintervention weighted mean scores were significantly higher in the intervention group than the control group for correct recognition of a mental health problem, help seeking, and stigmatization by 0.2 (SE 0.1; P=.003), 0.9 (SE 0.2; P<.001), and 1.8 (SE 0.4; P<.001), respectively.

Of the 456 participants, 236 (51.8%) responded to the follow-up survey after multiple waves of invitations and reminders. Comparisons of the demographic variables between respondents and nonrespondents of the follow-up survey indicated no significant differences in basic demographics (all P<.05). The demographics and health-related variables of the respondents are summarized in Table 2.

Table . Different aspects of mental health literacy (MHL) assessed immediately after the intervention by groups and the results on comparisons between groups (N=456).

MHL domains	Intervention (n=229), mean (SE)	Control (n=227), mean (SE)	*P* value[a]
Correct recognition	3.4 (0.1)	3.2 (0.1)	.003
Help seeking	12.9 (0.3)	11.9 (0.2)	<.001
Stigmatization	26.3 (0.5)	24.5 (0.6)	<.001
Social distancing	11.9 (0.5)	12.3 (0.6)	.16

[a]Adjusted for the clustering effect, age, education level, and baseline assessment of the outcome measure.

Table . Demographics and health-related variables of respondents who responded at baseline, postintervention, and 3-month follow-up (n=236).

Characteristics		Value[a]
Demographics		
	Age (years), mean (SD)	42.8 (9.7)
	Male sex, n (%)	107 (45.3)
	Education level (university or above), n (%)	184 (78)
	Marital status (married), n (%)	155 (65.7)
	Full-time employment (yes), n (%)	236 (100)
Health-related variables, n (%)		
	Regular exercise (yes)	168 (71.2)
	Smoker (yes)	6 (2.5)
	Drinker (moderate or heavy)	4 (1.7)

[a]Adjusted for the clustering effect.

The results of the changes in the MHL outcome measures and the comparisons are presented in Table 3. As shown, of the 4 main MHL outcomes, all but 1 yielded a significant time effect. After adjusting for the clustering effect, significant differences across time were found in help-seeking intention (*P*=.01), stigmatization (*P*<.001), and social distancing (*P*<.001). As suggested from the mean values, on the whole, there was an increase in help-seeking intention scores and a reduction in stigmatization and social distancing scores over time.

Table . Different aspects of mental health literacy (MHL) at baseline, postintervention, and 3-month follow-up and the results on the time effect (n=236).

MHL domains	Baseline, mean (SD)	Postintervention, mean (SD)	3-month follow-up, mean (SD)	*P* value
Correct recognition	3.2 (0.6)	3.4 (0.7)	3.3 (0.7)	.20
Help seeking	12.4 (1.9)	12.7 (2.3)	12.7 (2.1)	.01
Stigmatization	24.7 (3.6)	25.8 (4.1)	25.8 (3.9)	<.001
Social distancing	12.5 (3.2)	12.1 (3.3)	11.9 (3.5)	<.001

Evaluation of the mHealth Platform

For the evaluation of the mHealth Platform, a total of 70 individuals were recruited. Of these, there were 26 (37%) male individuals with an average age of 34.2 (SD 1.6) years. None of these participants had been exposed to any similar psychoeducation materials or mHealth platforms prior to the evaluation exercise. The results of the evaluation are summarized in Table 4. As shown, the average scores of the 4 major domains ranged from 3.8 to 4.2, with engagement having the lowest score. For the perceived impact on MHL, all items were scored higher than 4.0, suggesting a greater tendency of respondents to endorse the positive impact of the content on their MHL.

Table . Different domains of the Mobile Application Rating Scale (MARS) and the perceived impact on mental health literacy (n=70).

Variable		Value, mean (SD)
MARS domains		
	Engagement	3.8 (0.6)
	Functionality	4.2 (0.5)
	Aesthetics	4.0 (0.5)
	Information	4.2 (0.5)
	Overall	4.0 (0.4)
Perceived impact on mental health literacy		
	Awareness	4.4 (0.6)
	Knowledge	4.4 (0.5)
	Attitude	4.4 (0.6)
	Intention to change	4.3 (0.6)
	Help seeking	4.4 (0.6)
	Behavior change	4.3 (0.6)

Discussions

Principal Findings

This study was a continuation of a wait-listed CRCT of the effect of a psychoeducation intervention program using an mHealth spaced education approach on workplace MHL. The main results of the immediate efficacy of the intervention program had been reported previously [22]. This study aimed to further report the results obtained on the evaluation of the mHealth platform and the medium-term effect of the intervention program on the MHL of participants. Based on the MARS scores of different aspects of the platform, the results have demonstrated that the mobile platform is well acceptable, with an overall score of 4.0 (SD 0.4) and individual domain scores ranging between 3.8 and 4.2. This represents a generally good reception of the platform by the respondents in terms of the functionality, aesthetics, and the information provided. The result for the engagement domain was slightly below the level of good, in accordance with the classification of the original author of the MARS [25]. These results are compatible with the results obtained from a study rating 50 mental health and well-being mobile apps using the MARS by the original authors [25]. The average domain scores of these 50 apps ranged between 2.7 to 4.0 [25]. In comparison, the mHealth platform has shown better performance than these reported results.

The immediate effect of the psychoeducation intervention program on MHL was interpreted and discussed previously [22]. The significant and positive results indicated the efficacy of the intervention program in improving workers' MHL. There were significant improvements in 3 of the 4 main domains of MHL, including correct recognition, help seeking, and stigmatization, after the intervention. In terms of the medium-term sustainability of these gained benefits, the results obtained from the follow-up phase of the study provided some evidence. As shown, these effects were sustained 3 months after the intervention, with further improvement in social distancing, as reflected in the average scores. These results suggest the retention of the gained benefits in the enhancement of MHL in these participants. These results are consistent with those in the literature. A recent systematic review and meta-analytical study on the long-term effects of interventions for MHL in children and adolescents reported similar results [28]. In pooling data from 25 studies with an average follow-up period of 5 months, it was found that the improvement in MHL was sustained, particularly for stigmatization (*d*=0.30, 95% CI 0.24-0.36) and social distancing (*d*=0.16, 95% CI 0.03-0.29) [28]. When combining the results obtained from the evaluation exercise using the MARS, the results provide stronger evidence that the mHealth psychoeducation intervention program is effective in improving and retaining the gain in the MHL of workers in the workplace.

The *Mental Health Action Plan 2013-2020* produced by the WHO has affirmed the importance of the workplace as a venue for mental health education and advancement [2]. The results reported previously by the authors, in conjunction with those obtained from this study, provide further evidence that a well-developed and executed psychoeducation intervention program in the workplace can not only reduce burnout and stress but also enhance MHL. Moreover, the effect of MHL enhancement can be sustained and, to a certain extent, further improved over time. The results of this study have also demonstrated that a well-designed and engaging mHealth platform with good functionality and aesthetics would also be a vehicle for enhancing MHL. As aforementioned, the mHealth approach for disease treatment, management, and health promotion has been used in many different health areas [11,12]. However, it is still considered a rather recent area of development according to some scholars in the field [15]. It is anticipated that the COVID-19 pandemic could be a catalyst in stimulating and motivating health professionals to develop and adopt digital health more readily and rapidly [29]. With the latest development in the field of data science and artificial intelligence, it is also expected that greater health advancements could be achieved through personalized information provision for education and promotion [30].

The strengths and limitations of the wait-listed CRCT were discussed in the previous report and will not be reiterated [22]. In this study, the focuses were the evaluation of the mHealth platform and the medium-term effect of the intervention program. One of the strengths of the study is the use of standardized and validated assessment instruments for outcome variables. The MARS and MHL measures are widely used instruments with ample evidence for their reliability and validity. As a result, the measurement and interpretation biases can be minimized. Another strength of the study is the samples recruited for the evaluation study and the CRCT. The random sample of the CRCT consisted of employees of different large-sized industries, thus covering a wide range of work nature and seniority. Hence, it could be considered a representative sample. For the evaluation study, the sample also consisted of participants from different companies and working backgrounds, thus preserving some degree of representation of the target working population. A few limitations have been identified in this study. The design of the mHealth platform for the psychoeducation intervention program is based on the spaced education concept. However, in the evaluation of the platform, the use of the MARS could only focus on the design aspects of the app but not the effect of the approach on participants' learning. As a result, the effect of the spaced education approach adopted in this platform could not be evaluated. Another drawback of the study on the medium-term effect of the program is the low follow-up rate, with slightly more than half (236/456, 51.8%) of the original trial sample fully completing all assessments. This may incur a response bias in the follow-up study, although it has been demonstrated that there are no significant differences in the demographic variables between the follow-up and non–follow-up groups. To evaluate the learning effect of the spaced education approach, it is suggested that a randomized controlled trial with the use of an appropriate assessment instrument should be conducted. For minimizing the loss to follow-up, as suggested by some trialists, more frequent contact and better communication could help in retaining trial participants [31].

Conclusion

In conclusion, the web-based and mobile spaced education psychoeducation intervention program (WPMHL) has immediate and 3-month medium-term effects of retaining and improving MHL. The platform performed satisfactorily in terms of functionality, aesthetics, information content, and utility in enhancing MHL. It is anticipated that ongoing development in digital health will provide great benefits in improving the mental health of the global population.

Acknowledgments

The authors would like to acknowledge the Health and Medical Research Fund, the Food and Health Bureau, the Government of the Hong Kong Special Administrative Region for providing financial support for the study (grant #02181028). We would also like to acknowledge the kind assistance provided by colleagues, particularly Mr. Vincent Wong, of the Mental Health Association Hong Kong and Macau (MAHHKM), in providing administrative support.

Authors' Contributions

LTL and MKPL designed the study, and LTL obtained the funding. LTL also designed the statistical analysis plan and will direct the data analyses. LTL masterminded the design of the psychoeducation intervention program. The data collection questionnaire was developed by LTL and MKPL, with the mental health literacy scale translated and validated by LTL with the permission of the original author. All authors read and approved the final manuscript. The allocation of authorship is in accordance with the requirements of the International Committee of Medical Journal Editors (ICMJE).

Conflicts of Interest

None declared.

Checklist 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File, 3972 KB - mental_v11i1e51791_app1.pdf]

References

1. Whiteford HA, Degenhardt L, Rehm J, et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. Lancet 2013 Nov 9;382(9904):1575-1586. [doi: 10.1016/S0140-6736(13)61611-6] [Medline: 23993280]
2. World Health Organization. Mental Health Action Plan 2013-2020: WHO Press; 2013 URL: https://www.who.int/publications/i/item/9789241506021 [accessed 2024-04-10]
3. Gabriel P, Liimatainen MR. Mental health in the workplace: introduction: executive summary: International Labour Organization; 2000 URL: https://www.ilo.org/skills/pubs/WCMS_108221/lang--en/index.htm [accessed 2024-04-10]
4. Joyce S, Modini M, Christensen H, et al. Workplace interventions for common mental disorders: a systematic meta-review. Psychol Med 2016 Mar;46(4):683-697. [doi: 10.1017/S0033291715002408] [Medline: 26620157]

5. Jorm AF. Mental health literacy: public knowledge and beliefs about mental disorders. Br J Psychiatry 2000 Nov;177:396-401. [doi: 10.1192/bjp.177.5.396] [Medline: 11059991]
6. Lam LT, Wong P, Lam MK. Protocol for a phase III wait-listed cluster randomised controlled trial of an intervention for mental well-being through enhancing mental health literacy and improving work friendliness in Hong Kong. Trials 2019 Dec 4;20(1):672. [doi: 10.1186/s13063-019-3748-y] [Medline: 31801622]
7. Fields RD. Making memories stick. Sci Am 2005 Feb;292(2):75-81. [Medline: 15715394]
8. Kelley P, Whatson T. Making long-term memories in minutes: a spaced learning pattern from memory research in education. Front Hum Neurosci 2013 Sep 25;7:589. [doi: 10.3389/fnhum.2013.00589] [Medline: 24093012]
9. Kerfoot BP, Baker H. An online spaced-education game for global continuing medical education: a randomized trial. Ann Surg 2012 Jul;256(1):33-38. [doi: 10.1097/SLA.0b013e31825b3912] [Medline: 22664558]
10. Robinson T, Janssen A, Kirk J, et al. New approaches to continuing medical education: a QStream (spaced education) program for research translation in ovarian cancer. J Cancer Educ 2017 Sep;32(3):476-482. [doi: 10.1007/s13187-015-0944-7] [Medline: 26574041]
11. Tuberculosis: global task force on digital health. World Health Organization. 2015 Jun 3. URL: https://www.who.int/news-room/questions-and-answers/item/tuberculosis-global-task-force-on-digital-health [accessed 2024-04-10]
12. Carolan S, Harris PR, Cavanagh K. Improving employee well-being and effectiveness: systematic review and meta-analysis of web-based psychological interventions delivered in the workplace. J Med Internet Res 2017 Jul 26;19(7):e271. [doi: 10.2196/jmir.7583] [Medline: 28747293]
13. Sequi-Dominguez I, Alvarez-Bueno C, Martinez-Vizcaino V, Fernandez-Rodriguez R, del Saz Lara A, Cavero-Redondo I. Effectiveness of mobile health interventions promoting physical activity and lifestyle interventions to reduce cardiovascular risk among individuals with metabolic syndrome: systematic review and meta-analysis. J Med Internet Res 2020 Aug 31;22(8):e17790. [doi: 10.2196/17790] [Medline: 32865503]
14. Ruco A, Dossa F, Tinmouth J, et al. Social media and mHealth technology for cancer screening: systematic review and meta-analysis. J Med Internet Res 2021 Jul 30;23(7):e26759. [doi: 10.2196/26759] [Medline: 34328423]
15. Nwaozuru U, Obiezu-Umeh C, Shato T, et al. Mobile health interventions for HIV/STI prevention among youth in low- and middle-income countries (LMICs): a systematic review of studies reporting implementation outcomes. Implement Sci Commun 2021 Nov 6;2(1):126. [doi: 10.1186/s43058-021-00230-w] [Medline: 34742357]
16. Buss VH, Leesong S, Barr M, Varnfield M, Harris M. Primary prevention of cardiovascular disease and type 2 diabetes mellitus using mobile health technology: systematic review of the literature. J Med Internet Res 2020 Oct 29;22(10):e21159. [doi: 10.2196/21159] [Medline: 33118936]
17. Melia R, Francis K, Hickey E, et al. Mobile health technology interventions for suicide prevention: systematic review. JMIR Mhealth Uhealth 2020 Jan 15;8(1):e12516. [doi: 10.2196/12516] [Medline: 31939744]
18. Milne-Ives M, Lam C, De Cock C, Van Velthoven MH, Meinert E. Mobile apps for health behavior change in physical activity, diet, drug and alcohol use, and mental health: systematic review. JMIR Mhealth Uhealth 2020 Mar 18;8(3):e17046. [doi: 10.2196/17046] [Medline: 32186518]
19. Berrouiguet S, Baca-García E, Brandt S, Walter M, Courtet P. Fundamentals for future mobile-health (mHealth): a systematic review of mobile phone and web-based text messaging in mental health. J Med Internet Res 2016 Jun 10;18(6):e135. [doi: 10.2196/jmir.5066] [Medline: 27287668]
20. Punukollu M, Marques M. Use of mobile apps and technologies in child and adolescent mental health: a systematic review. Evid Based Ment Health 2019 Nov;22(4):161-166. [doi: 10.1136/ebmental-2019-300093] [Medline: 31358537]
21. Grist R, Porter J, Stallard P. Mental health mobile apps for preadolescents and adolescents: a systematic review. J Med Internet Res 2017 May 25;19(5):e176. [doi: 10.2196/jmir.7332] [Medline: 28546138]
22. Lam LT, Lam MK, Reddy P, Wong P. Efficacy of a workplace intervention program with web-based online and offline modalities for improving workers' mental health. Front Psychiatry 2022 May 31;13:888157. [doi: 10.3389/fpsyt.2022.888157] [Medline: 35711597]
23. Jorm AF, Korten AE, Jacomb PA, Christensen H, Rodgers B, Pollitt P. “Mental health literacy”: a survey of the public's ability to recognise mental disorders and their beliefs about the effectiveness of treatment. Med J Aust 1997 Feb 17;166(4):182-186. [doi: 10.5694/j.1326-5377.1997.tb140071.x] [Medline: 9066546]
24. Gulliver A, Griffiths KM, Christensen H, Brewer JL. A systematic review of help-seeking interventions for depression, anxiety and general psychological distress. BMC Psychiatry 2012 Jul 16;12:81. [doi: 10.1186/1471-244X-12-81] [Medline: 22799879]
25. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile App Rating Scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 2015 Mar 11;3(1):e27. [doi: 10.2196/mhealth.3422] [Medline: 25760773]
26. Bardus M, Awada N, Ghandour LA, et al. The Arabic version of the Mobile App Rating Scale: development and validation study. JMIR Mhealth Uhealth 2020 Mar 3;8(3):e16956. [doi: 10.2196/16956] [Medline: 32130183]
27. Messner EM, Terhorst Y, Barke A, et al. The German version of the Mobile App Rating Scale (MARS-G): development and validation study. JMIR Mhealth Uhealth 2020 Mar 27;8(3):e14479. [doi: 10.2196/14479] [Medline: 32217504]

28. Freţian AM, Graf P, Kirchhoff S, et al. The long-term effectiveness of interventions addressing mental health literacy and stigma of mental illness in children and adolescents: systematic review and meta-analysis. Int J Public Health 2021 Dec 15;66:1604072. [doi: 10.3389/ijph.2021.1604072] [Medline: 34975363]
29. Budd J, Miller BS, Manning EM, et al. Digital technologies in the public-health response to COVID-19. Nat Med 2020 Aug;26(8):1183-1192. [doi: 10.1038/s41591-020-1011-4] [Medline: 32770165]
30. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. J Med Internet Res 2023 Feb 24;25:e40789. [doi: 10.2196/40789] [Medline: 36826990]
31. Madden K, Scott T, McKay P, et al. Predicting and preventing loss to follow-up of adult trauma patients in randomized controlled trials: an example from the FLOW trial. J Bone Joint Surg Am 2017 Jul 5;99(13):1086-1092. [doi: 10.2106/JBJS.16.00900] [Medline: 28678121]

Abbreviations

ANZCTR: Australian New Zealand Clinical Trials Registry
CRCT: cluster randomized controlled trial
MARS: Mobile Apps Rating Scale
mHealth: mobile health
MHL: mental health literacy
WHO: World Health Organization
WPMHL: Workplace Mental Health Literacy

Edited by J Torous; submitted 11.08.23; peer-reviewed by AAL Sawafi, R Eckhoff; revised version received 23.02.24; accepted 15.03.24; published 23.04.24.

Please cite as:
Lam LT, Lam MKP
A Web-Based and Mobile Intervention Program Using a Spaced Education Approach for Workplace Mental Health Literacy: Cluster Randomized Controlled Trial
JMIR Ment Health 2024;11:e51791
URL: https://mental.jmir.org/2024/1/e51791
doi:10.2196/51791

Original Paper

Implementation of an Electronic Mental Health Platform for Youth and Young Adults in a School Context Across Alberta, Canada: Thematic Analysis of the Perspectives of Stakeholders

Gina Dimitropoulos[1,2,3,4], MSW, PhD, RSW, RMFT; Emilie M Bassi[1], BA, MSc; Katherine S Bright[5,6], BSc, BN, MN, PhD; Jason Gondziola[7], BA, MBA; Jessica Bradley[7], RN, MN; Melanie Fersovitch[7], BN; Leanne Stamp[7], BEd; Haley M LaMonica[8], PhD, ABPP-CN; Frank Iorfino[8], BSc, MBMSc, PhD; Tanya Gaskell[7], BA, MBA; Sara Tomlinson[7], MSc, CHE; David Wyatt Johnson[9,10], MD

[1]Faculty of Social Work, University of Calgary, Calgary, AB, Canada
[2]Calgary Eating Disorders Program, Alberta Health Services, Calgary, AB, Canada
[3]Mathison Centre for Mental Health Research and Education, University of Calgary, Calgary, AB, Canada
[4]Alberta Children's Hospital Research Institute, University of Calgary, Calgary, AB, Canada
[5]School of Nursing and Midwifery, Faculty of Health, Community, and Education, Mount Royal University, Calgary, AB, Canada
[6]Heroes in Mind, Advocacy, and Research Consortium (HiMARC), Faculty of Rehabilitation Medicine, College of Health Sciences, University of Alberta, Edmonton, AB, Canada
[7]Provincial Addiction and Mental Health, Alberta Health Services, Calgary, AB, Canada
[8]Brain and Mind Centre, University of Sydney, Sydney, Australia
[9]Departments of Pediatrics, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
[10]Maternal Newborn Child and Youth Strategic Clinical Network, Alberta Health Services, Calgary, AB, Canada

Corresponding Author:
Gina Dimitropoulos, MSW, PhD, RSW, RMFT
Faculty of Social Work
University of Calgary
2500 University Drive NW
Calgary, AB, T2N 1N4
Canada
Phone: 1 403 220 7332
Email: gdimit@ucalgary.ca

Abstract

Background: Youth, aged 15 to 24 years, are more likely to experience mental health (MH) or substance use issues than other age groups. This is a critical period for intervention because MH disorders, if left unattended, may become chronic and serious and negatively affect many aspects of a young person’s life. Even among those who are treated, poor outcomes will still occur for a percentage of youth. Electronic MH (eMH) tools have been implemented in traditional MH settings to reach youth requiring assistance with MH and substance use issues. However, the utility of eMH tools in school settings has yet to be investigated.

Objective: The objective of this study was to gain an understanding of the perspectives of key school staff stakeholders regarding barriers and facilitators to the implementation of the Innowell eMH platform in secondary schools across the province of Alberta, Canada.

Methods: Guided by a qualitative descriptive approach, focus groups were conducted to elicit stakeholder perspectives on the perceived implementation challenges and opportunities of embedding the Innowell eMH platform in secondary school MH services. In total, 8 focus groups were conducted with 52 key school staff stakeholders.

Results: Themes related to barriers and facilitators to youth and school MH care professional (MHCP) capacity in implementing and using eMH tools were identified. With respect to youth capacity barriers, the following themes were inductively generated: (1) concerns about some students not being suitable for eMH services, (2) minors requiring consent from parents or caregivers to use eMH services as well as confidentiality and privacy concerns, and (3) limited access to technology and internet service among youth. A second theme related to school MHCP barriers to implementation, which included (1) feeling stretched with high caseloads and change fatigue, (2) concerns with risk and liability, and (3) unmasking MH issues in the face of limited resources. In contrast to the barriers to youth and MHCP capacity, many facilitators to implementation were discussed. Youth capacity facilitators included (1) the potential for youth to be empowered using eMH tools, (2) the platform fostering therapeutic relationships with school personnel, and (3) enhancing access to needed services and resources. MHCP capacity facilitators to implementation were (1) system transformation through flexibility and problem-solving, (2) opportunities for collaboration with youth and MHCPs and across different systems, and (3) an opportunity for the continuity of services.

Conclusions: Our findings highlight nuanced school MHCP perspectives that demonstrate critical youth and MHCP capacity concerns, with consideration for organizational factors that may impede or enhance the implementation processes for embedding eMH in a school context. The barriers and facilitators to implementation provide future researchers and decision makers with challenges and opportunities that could be addressed in the preimplementation phase.

(JMIR Ment Health 2024;11:e49099) doi:10.2196/49099

KEYWORDS

electronic mental health; eMH; digital mental health; youth and young adult mental health; secondary schools; implementation science; qualitative descriptive methods; mental health platform; mental health; mobile phone

Introduction

Background

Globally, mental health (MH) issues are on the rise among young people, particularly since the onset of the COVID-19 pandemic [1,2]. A recent meta-analysis showed that of all people with MH disorders, 62.5% had onset before the age of 25 years [3]. Furthermore, 1 in 10 youths is estimated to experience an MH disorder in their lifetime, including depression, anxiety, and eating disorders [4]. Approximately 1 in 3 youths is estimated to experience an MH disorder in their lifetime [5]. Since the onset of the COVID-19 pandemic, the rates of depression and anxiety are estimated to have doubled among youth compared with the prepandemic period [6]. A recent systematic review found evidence of decreases in access to preventive MH supports coupled with increases in attempted suicide, self-harm, and suicidal ideation among adolescents during the COVID-19 pandemic [2]. The increase is linked to a disruption to daily routines and family connections [4], a reduction in the availability of social support [1], and a lack of necessary coping skills to navigate challenges and stressors experienced by youth [4]. Furthermore, public health restrictions and the increasing rates of social isolation have resulted in higher rates of internet and gaming problems among youth [7,8].

Despite the concerns about the rise in MH problems among youth, many barriers to help-seeking behaviors have been noted, including stigma [9], low MH literacy, and a reluctance to seek MH services [10]. There are many systemic barriers that further interfere with youth accessing needed services when they seek support, including long waitlists and limited availability of evidence-based treatments that are youth friendly and developmentally appropriate [11]. In a recent systematic review, long wait times for MH services were linked to a deterioration in MH outcomes [12]. As a result of the pandemic, web-based care options, including the application of electronic MH (eMH) tools, were widely and rapidly implemented, demonstrating their utility in reaching youth affected by MH and substance misuse and their ability to transcend geographic distance during public health restrictions [13].

The global spike in poor youth MH calls for early detection and intervention of MH disorders using traditional and nontraditional MH services, including eMH services [9,14]. eMH is defined as the "provision of guided mental health care where consumers navigate a rapid and more effective system experience of service entry, skilled assessment, and multidisciplinary and coordinated care, as well as ongoing outcome-based monitoring" [15]. A proliferation of eMH services has been observed using smartphone apps and electronic tools (e-tools) [16,17]. eMH services may supplement various aspects of MH service delivery, including disseminating resources and information, performing ongoing assessments, tracking changes to symptoms and behaviors, delivering psychosocial therapies, and providing peer support [16]. Randomized controlled trials have consistently demonstrated a reduction in some MH symptoms, including mild to moderate depression and anxiety [18-20]. These randomized controlled trials have especially focused on delivering treatment via smartphone interventions (those that provided in-app feedback, and those used to enhance or support face-to-face interventions) [18]. However, research on the implementation of eMH tools outside of routine clinical settings is lacking.

Research suggests the importance of the school setting as a place of first contact for youth who seek MH support [21-23], positioning teachers as the first point of contact to see MH "red flags" [24]. The use of eMH technology provides an opportunity to transform services, including improving the early detection of MH issues, better care coordination and care plans, and referrals to necessary services for youth [25]. However, there is limited research on the implementation and use of eMH tools in school systems.

Existing studies have highlighted general barriers and facilitators to the implementation of digital MH tools among both youth and MH care professionals (MHCPs). Stigma remains a daunting barrier to using eMH services, deterring individuals from seeking help owing to societal and internal biases [26,27]. The quality of the therapeutic relationship, fostered by well-designed interfaces and personalized content, serves as a facilitator, particularly in remote settings [26,27]. Systemic influences (eg, access and infrastructure) can both hinder and facilitate engagement [26]. In addition, one's level of MH literacy significantly impacts one's decision to engage, with greater literacy leading to the increased acceptance and use of eMH resources [26]. In summary, these intertwined factors create a complex landscape for eMH and digital MH interventions. The available literature exploring the barriers and facilitators to implementing eMH interventions in the school setting is notably scarce, underscoring the need for a deeper exploration of this important topic. To our knowledge, this study is the first to explore the barriers and facilitators to implementing eMH interventions in the school setting.

The Innowell eMH platform is a configurable web-based tool that seeks to provide a more personalized approach using measurement-based care to complement existing MH services [25]. The platform empowers youth and young adults to actively participate in their care plan by using measurement-based MH care, including a comprehensive questionnaire consisting of psychometric tools exploring 20 biopsychosocial domains (eg, depression, anxiety, and social connectedness). Examples of health priorities and health cards on the platform's dashboard are provided in Figure 1 [28] and Figure 2 [29]. After completing baseline measures, platform users are immediately provided access to vetted web-based resources, including apps, e-tools, and crisis line options to use between sessions, and cues to discuss care options with their MHCPs. The platform can assist "with the assessment, feedback, management, and monitoring of their mental ill health and maintenance of well-being by collecting personal and health information from a young person, their clinician(s), and supportive others" [25]. The platform was designed using participatory design in Australia. The participatory design processes included participants with lived experience, health professionals, and service staff from diverse service populations to optimize the usability of the platform design [25,30]. It is important to note that the "platform does not provide stand-alone medical or health advice, risk assessment, clinical diagnosis, or treatment. Instead, it guides and supports (but does not direct) young people and their providers to decide what may be suitable care options" [25].

Figure 1. Participant and clinician electronic mental health Innowell platform dashboard health priorities [28].

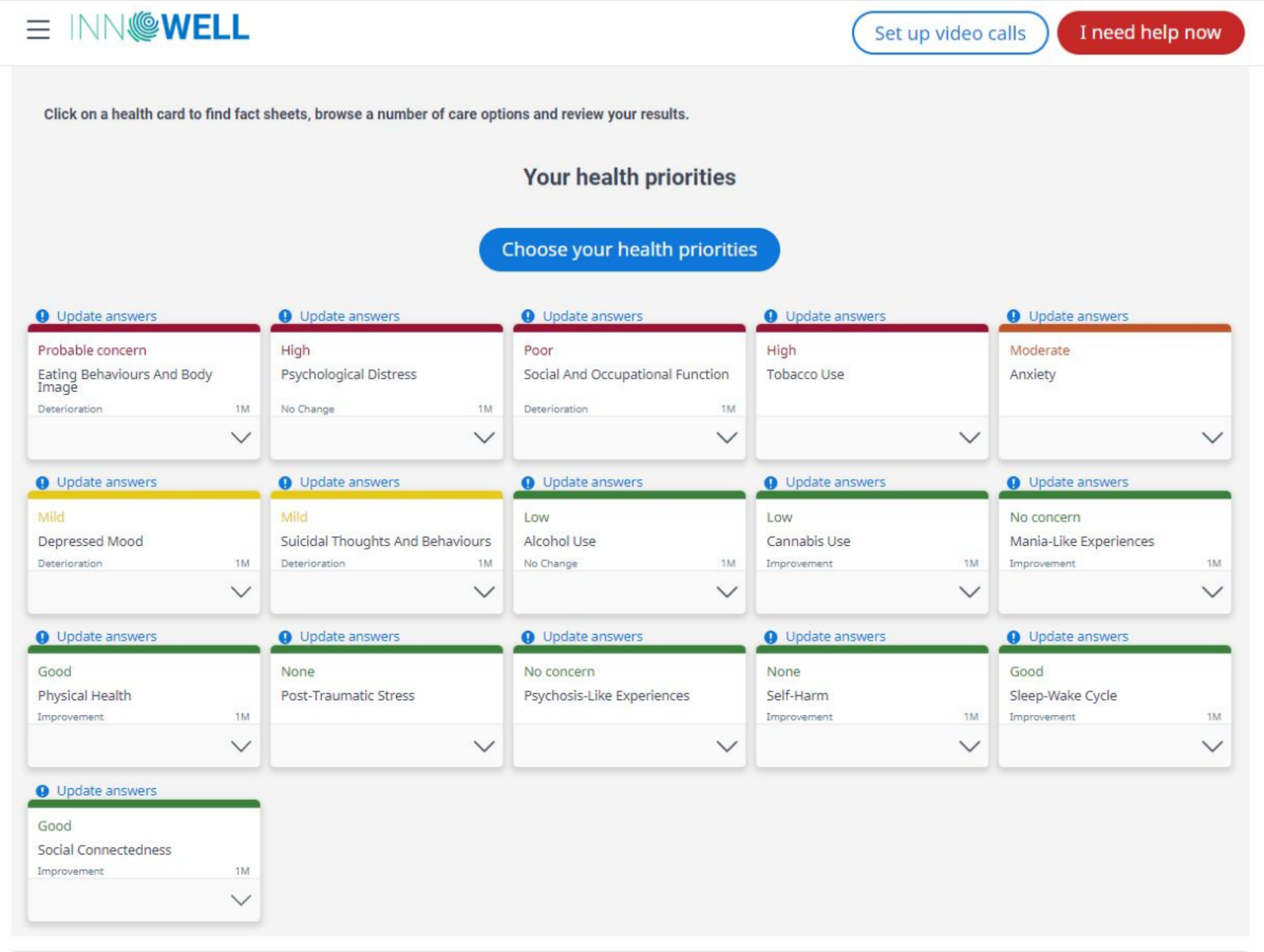
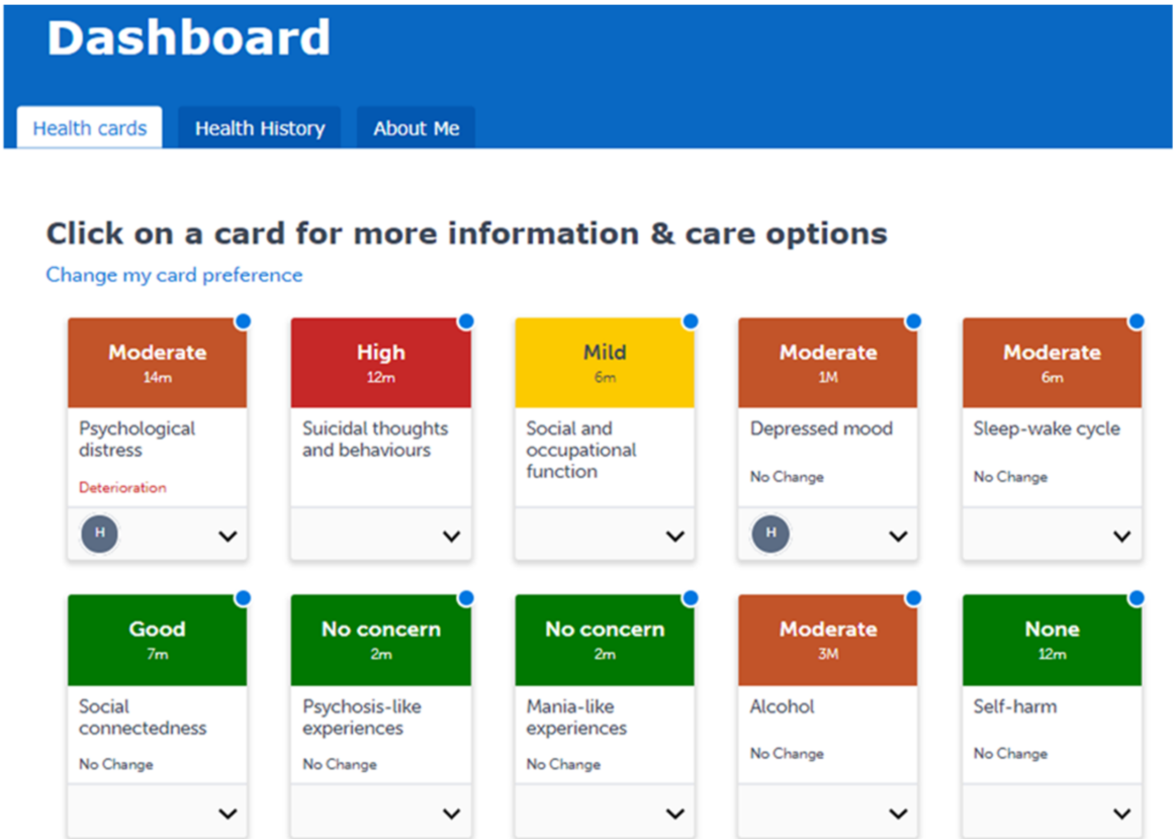


Figure 2. Participant and clinician electronic mental health Innowell platform dashboard health cards [29].



Objectives

Hence, it is critical to evaluate the implementation of an eMH tool such as the Innowell eMH platform in secondary schools to assess its acceptability, usability, and efficacy in school contexts. Furthermore, there is a need to examine how eMH tools can be customized for school-based organizations where youth often spend most of their time interacting with educators, guidance counselors, and school MHCPs. This study seeks to elicit the perspectives of key stakeholders within the school divisions regarding their perceptions of barriers and facilitators to the potential implementation of an eMH platform in schools across the province of Alberta, Canada.

Methods

This study used a qualitative thematic analysis of school stakeholder focus groups (FGs) to better understand their perceived barriers and facilitators to the implementation of an eMH platform for youth and young adults in Alberta.

Ethical Considerations

This research study received human participant research ethics approval from the University of Calgary Research Ethics Board (project name "Pre-Implementation and Implementation Phase of E-Mental Health for Youth and Young Adults in Alberta"; REB20-1137). Informed consent was completed on the web before the commencement of the FGs. The study participants' data were deidentified to ensure privacy and confidentiality. Participants were not compensated for their participation.

Qualitative Descriptive Methodology

A qualitative descriptive methodology was used to describe barriers and facilitators to the implementation of the Innowell eMH platform through the exploration of participants' perceptions and experiences [31]. This methodology is well suited to health research and is descriptive in nature [31,32], capturing events or conditions from the perspective of individuals and in their everyday language [33]. We sought to obtain a rich description of key stakeholder perspectives rather than researcher-provided interpretations of their preconceived views about the value of an eMH platform. We used an FG approach to obtain data from a group of key stakeholders in schools who may hold diverse concerns about the utility of the platform. In contrast to individual interviews, FGs offer participants the opportunity to share a range of thoughts and to iteratively build on ideas and considerations for integrating an eMH platform in school settings across different geographic regions and diverse student populations [34]. The sample consisted of 52 school district key stakeholders representing 8 communities across Alberta. The web-based FGs were conducted from February 1, 2021 to May 31, 2022.

A semistructured interview guide was created with input from the research and implementation teams on the project. The questions included a focus on thoughts and feelings about the platform and the implementation processes. The interview guide was structured into 3 separate but related sections and was constructed to suit the study objectives, research questions, and the theoretical domains framework (TDF). The TDF is an integrative framework that includes 14 theoretical domains derived from 33 validated health and social psychology theories and 128 constructs that have the potential to drive and explain health-related behavior change [35]. The 14 theoretical domains are knowledge; skills; social or professional role and identity; beliefs about capabilities; optimism; beliefs about consequences; reinforcement; intentions; goals; memory, attention, and decision processes; environmental context and resources; social influences; emotions; and behavioral regulation [36]. The interview guide was designed to explore which TDF domains were relevant for the implementation of the Innowell eMH platform.

FGs were conducted to evaluate the barriers and facilitators that may affect the implementation of the Innowell eMH platform in existing school services. The interview guide covered unique factors about the service setting and organization, unique community contexts that can influence implementation, questions about the unique youth population that the school serves, and the barriers and facilitators that might influence implementation.

The study setting for recruitment included kindergarten to grade 12 schools situated in rural and urban communities across Alberta. This study used a targeted email and web-based meeting recruitment, whereby practice leads from the implementation team working with each community shared recruitment materials with senior decision makers and clinical supervisors at schools taking part in the eMH study. Key stakeholders in this area were recruited, including teachers, administrators, psychologists, school counselors, community connectors, and social workers. The site clinical lead for each community identified potential participants who would be involved in using the platform from a range of different roles, including managers, leaders, and frontline clinicians. Three members of the research team moderated all FGs.

There were 3 inclusion criteria for participants of this study. Each participant had to be employed in the school district as an MHCP, decision maker, educator, or administrator; be proficient in written and spoken English; and use web-based technologies on a laptop computer, desktop computer, tablet, or smartphone.

The FGs were conducted by the research team and included an implementation practice lead, 2 facilitators, and a notetaker who was often a youth research partner. The FGs lasted approximately 90 minutes. Of the 52 participants, 3 (6%) to 10 (19%) participated in each FG. All interviews were audio recorded and transcribed verbatim by a professional transcriber. The transcribed interviews were checked for accuracy by a member of the research team.

A total of 8 FGs were conducted with the 52 key stakeholders from 11 school divisions. The FGs collectively contained representation from 21 schools, including high schools (grades 10-12; n=10, 48%), an elementary school (kindergarten to grade 6; n=1, 5%), combined elementary to high school (kindergarten to grade 12; n=3, 14%), combined junior high to high school (grades 7-12; n=3, 14%), as well as outreach and web-based schools (n=4, 19%). Participants included teachers, administrators, psychologists, school counselors, community connectors, and social workers. Community connectors in the school division are members of the community who support and inform individuals about how to access support groups, services, and information that might help improve their MH and well-being [37]. Sample size was justified at 52 key stakeholders when the research team determined that the sample provided "information power," a concept used in qualitative methodology to denote that the sample held substantial information [38]. There was representation from public, Catholic, Francophone, and web-based and alternative outreach schools for students (also referred to as clients by some of the participants) who benefit from nontraditional learning methods. The FGs were composed of diverse school district MHCPs, from managers and administrators to teachers.

Analytic Plan: Thematic Analysis

Using a combined inductive-deductive approach, the research team used the 6 stages of thematic analysis outlined by Braun and Clarke [39]. Thematic analysis is a method of identifying and reporting on themes that emerge through the data [39]. The six-stage analytical process consists of (1) familiarization with the data, (2) coding, (3) generating themes, (4) reviewing themes, (5) defining and naming themes, and (6) producing the report [39].

As part of the project's youth engagement strategy, a team of coders, including youth and young adults with varying research experience, reviewed the transcribed FGs independently to increase familiarization with the data. After familiarization, the coders conducted a preliminary descriptive coding of their assigned transcripts. The research team generated these codes independently using Microsoft Word and Taguette qualitative coding software. The coders then exchanged FG transcripts and conducted secondary coding. Each transcript was coded by a minimum of 2 coders, which ensured reliability because the codes depended on 2 different individuals achieving and deciding on the same code outcome [40]. The research team then worked collaboratively to group the codes into broader themes through discussion and consensus. Finally, illustrative quotes for each theme were identified. The core research team met regularly to share their impressions of patterns across and within the FG transcripts.

The team closely followed guidelines for publishing qualitative data [41]; for instance, the research team recorded decisions about the coding and the decision-making processes for establishing the themes. The coders worked independently to analyze the data and met regularly to discuss the codes established and assigned. Through consensus, the coding group members formulated their themes based on patterns across the FGs with different school settings. As a form of member checking, our team presented our findings to the broader research team consisting of researchers from different academic institutions to ensure that the data resonated with their experiences and understanding of the academic literature. The practice of reflexivity was used through memoing and group discussions, where members of the research team reflected on their biases and their positionality and questioned their own assumptions throughout the direction of the research process. The practice of keeping written memos was used to record the process and document the rationale for how the data were coded [42].

Results

Participant Demographics

Most of the participants (38/52, 73%) identified as women. Many of the participants were counselors or therapists (12/52, 23%) or social workers (11/52, 21%). Of the 52 participants, 19 (37%) were aged between 40 and 49 years, and 32 (62%) had been practicing their profession for ≥11 years. The majority of the participants were born in Canada (45/52, 87%) and held a graduate degree (22/52, 42%). A complete list of descriptive information about the participants is presented in Table 1.

Table 1. Descriptive quantitative survey information of school stakeholders (N=52).

Variable	Values, n (%)
Gender	
Woman	38 (73)
Man	7 (14)
Not reported	7 (13)
Profession	
Counselor or therapist	12 (23)
Social worker	11 (21)
Teacher	7 (13)
Administrator	6 (12)
Psychologist	6 (12)
Community Connector	2 (4)
Not reported	8 (15)
Age group (y)	
20-29	3 (6)
30-39	15 (29)
40-49	19 (37)
50-59	8 (15)
60-69	2 (4)
Not reported	5 (10)
Member of a visible minority group	
Yes	5 (10)
No	39 (75)
Not reported	8 (15)
Length of time practicing (y)	
<1-5	6 (12)
6-10	9 (17)
≥11	32 (62)
Not reported	5 (10)
Length of time at organization (y)	
<1	6 (12)
1-5	10 (19)
6-10	9 (17)
≥11	22 (42)
Not reported	5 (10)
Employment status	
Full time	35 (67)
Part time	7 (13)
Other	5 (10)
Not reported	5 (10)
Ethnicity (n=54)[a]	
Indigenous to North America	5 (9)
Other North American origins	7 (13)

Variable	Values, n (%)
European origins	28 (52)
Latin, Central, and South American origins	1 (2)
Not reported	11 (21)
Born in Canada	
Yes	45 (87)
No	1 (2)
Not reported	6 (12)
Level of education	
Some university, college, or trades school	1 (2)
Completed college or trades school	5 (10)
Bachelor's degree from a university	18 (35)
Graduate school	22 (42)
Not reported	6 (12)

[a]Numbers do not equal 52 because participants were allowed to select multiple responses.

Barriers and Facilitators

In each of the following subsections, we highlight youth and MHCP capacity barriers and facilitators to the implementation of eMH tools in school settings. Youth capacity to engage in eMH services refers to their ability to effectively use digital tools and platforms, such as eMH, for addressing MH concerns. This involves not only their technical skills but also their understanding of the potential benefits and limitations of eMH services. Clinician capacity to engage in eMH services refers to their proficiency in using digital tools and platforms, such as MH assessment tools, to provide MH services to their patients. This also encompasses their ability to navigate the technical aspects of an eMH platform as well as their competence in delivering quality MH care while adhering to the regulations of their profession's governing body as well as their health care organization's ethics and regulatory guidelines in the eMH realm.

First, youth capacity barriers were identified, which included the suitability of measurement-based care for all youth with different types of MH disorders that may impact their engagement in eMH services; the role of caregivers, confidentiality, and perceived consent concerns among youth; and inadequate device and internet access. We then established a second theme relating to MHCP capacity barriers to implementation, which included MHCPs feeling stretched with high caseloads and change fatigue, concerns about liability and risk considerations, and the potential to unmask MH issues in the face of service and resource constraints.

By contrast, the final 2 themes consisted of many facilitators to implementation. The third theme—youth capacity facilitators—included the potential for an active and empowered role in care, the potential to foster and enhance therapeutic relationships, and the importance of improving access to services and resources. The fourth theme covered MHCP capacity facilitators to implementation, including the unique flexibility and natural problem-solving skills of school staff that can contribute to system transformation, the potential for collaboration with MHCPs across the continuum and with different systems, and an opportunity to strengthen the continuity of services. Collectively, these themes highlight the potential of the school setting to create opportunities for system transformation through the implementation of eMH tools.

Barriers to the Implementation of eMH Tools in School Settings

Youth Capacity Barriers

Individual Characteristics and Implications for the Engagement of Measurement-Based Care

Participants identified numerous factors as potential barriers to youth using the measurement-based care protocol integrated within the Innowell eMH platform. Given the large and diverse youth population served by secondary schools in Alberta, participants expressed concern that some youth with certain types of MH issues and sociodemographic characteristics may be unable to use the Innowell eMH platform. Participants perceived that measurement-based care may not be suitable for youth with the following conditions: severe MH diagnoses (eg, moderate to severe depression or anorexia nervosa), disability diagnoses (eg, attention-deficit/hyperactivity disorder), low cognitive capacities, those experiencing suicidal and self-harm behaviors, and low literacy levels and new or migrating Canadians not fluent in English. Participants were concerned that measurement-based care, which includes the 20-domain assessment features on the Innowell eMH platform, might be lengthy, difficult to understand, and overwhelming:

> *I do wonder about say students just with their literacy levels, like their comprehension. And especially those who maybe present a specific neurotypical, they present that way but actually maybe their comprehension is really, really weak. And so, onboarding and then going through the diagnostic*

> *tools and the self-assessment tools. Like I do wonder about the implications of that.* [FG21-P02]

This quote demonstrates many participants' reflections about the implications of onboarding youth who may be unable to easily navigate certain aspects of the platform on their own.

Participants also described a concern that the extensive measurement-based care assessment protocol on the Innowell eMH platform might exacerbate symptoms for some of the youth experiencing complex MH issues:

> *I do worry that part of this platform, although [it] gives them the tools and the apps and all of those things, and reminds them to speak to the therapist or go see their person. We're also asking a lot of kids who are struggling...Right? So, they're not feeling very mentally well and then we're asking them to do other things to take care of their own wellness. And I worry—are they capable of that? I guess we'll see.* [FG17-P03]

This quote demonstrates a perception that some youth may find it difficult to navigate the Innowell eMH platform when they are struggling with MH issues and wellness.

Role of Caregivers, Confidentiality, and Caregiver Consent

Other potential barriers identified by participants are problems that may arise in obtaining consent from caregivers and protecting the confidentiality of youth who will access the platform but do not wish to disclose their MH concerns to family members:

> *And if a lot of the issues surrounds the parent relationship, then that becomes a problem and that has definitely been the past barrier to them getting services sometimes is they don't want their parents having to sign consents.* [FG22-P03]

Although not a pervasive issue, participants raised concerns about some youth living in unsafe family environments where it would not be acceptable to disclose an MH issue owing to stigma and fear of reprisals from family members. These factors were described as significant barriers if consent is required from caregivers for the access and use of the Innowell eMH platform. In the absence of clear confidentiality assurances, it may prevent young people from engaging in the apps and e-tools available through the platform and in MH services more broadly.

Information-sharing requirements within the school district with caregivers was also described as a potential barrier to implementation. Participants were concerned about who *owns* the assessment information completed by youth and having to navigate information-sharing requests or requirements from caregivers or other MHCPs:

> *I think it would need to be very clear what, if any, of the information would be getting shared with their parents, like abundantly clear. I get questions on this all the time when I meet the kids for the first time is, "What exactly is confidentiality?" And many, many questions surrounding that.* [FG10-P03]

This participant builds on this concern, stating that many young people come to their first session reluctantly, which could spread to minimal uptake of the Innowell eMH platform if information-sharing protocols about consent and parental involvement are unclear. Participants shared that being unclear about confidentiality with young people would be a barrier not only to using the platform but also to young people accessing MH services and resources more generally.

Limited Access to Technology and Internet Service Among Youth

A significant barrier described in the school context is a lack of access to technology for youth to use the Innowell eMH platform. With the shift to web-based school and MH services during the COVID-19 pandemic, many MHCPs learned of the access-related challenges with remote and web-based schooling for some youth, especially those residing in remote or rural geographic regions. Participants learned that many youth lack access to technology or devices that would be necessary to use the platform; and many youth also have unstable, or completely lack, Wi-Fi, data, or an internet connection needed to access the platform:

> *[O]ne of the things that has been a challenge over—specifically with the course of the pandemic, is it really has brought to light how many students—because we are in a rural, northern remote area where Wi-Fi access and device access, cell phone access, all that stuff can be really spotty if not almost impossible for some of our families which means some of the students don't have access to this kind of thing outside of typical school hours.* [FG22-P03]

Participants were concerned that introducing an eMH platform, the use of which requires access to devices and a stable Wi-Fi connection, would be an unsuitable initiative for many of their youth population, especially those who cannot afford Wi-Fi access, thus not addressing existing barriers to care for youth considered disadvantaged and vulnerable.

MHCP Capacity Barriers

Feeling "Really Stretched" With High Caseloads and Change Fatigue

Participants described the school system as often the first point of contact for "a revolving door" (FG17-P02) of students experiencing diverse challenges, from interpersonal problems to the onset of MH disorders. This creates difficulty for them to envision the integration of an eMH platform that might worsen their already high caseloads. Participants described feeling "really stretched" (FG5-P08) and supporting students in the school district as "mentally taxing" (FG21-P02). A participant highlighted these concerns:

> *[L]ike my vested interest in this point-in-time is it not just a huge, huge source of time and energy suck for them because obviously, before this project, they were busy with their full-time work anyways. Now this is like another thing on their plate. So, that would be my sort of trepidation with this is how it will add more to their plate and potentially increase their stress.* [FG21-P02]

Participants were concerned about implementing a new eMH platform that would be time consuming and potentially overwhelming to navigate, given the many conflicting demands on their schedules in an ever-changing work environment. Participants also described a reluctance to undergo the necessary organizational changes needed to integrate eMH into their routine practice:

> *[W]e're always given new tools constantly. "Try this, these are different assessments and mental health promotions has this going on."* [FG17-P02]

Participants questioned how a new tool could be integrated into the existing organizational flow and day-to-day tasks for school staff. This includes the allocation of resources to integrate a new eMH tool, the limited contact time with youth, how this tool factors into waitlist management, and how to integrate this into current data management practices and software. Participants described other data management tools that they have been grappling with in the school system and a concern about how additional eMH tools may create work duplication, especially if there is a lack of interoperability.

Furthermore, the impact of the COVID-19 pandemic on the school system resulted in higher workloads for many school MH staff. Some of the participants noted experiencing a high volume of cases with students with severe MH acuity and a reduction in the availability of community-based MH services; for example, a participant shared how the COVID-19 pandemic has impacted the availability of other community-based MHCPs:

> *[L]ack of resources. We lost our child psychiatrist in the community. The contract wasn't renewed. So, that puts an extra burden on the schools for services. That affects both school jurisdictions as well as with the current COVID situation again, it's the lack of service providers being able to access the school environment.* [FG22-P04]

Thus, when grappling with external factors, such as the unexpected consequences of the COVID-19 pandemic and the constant upgrading of organizational tools, participants were apprehensive about additional training requirements for another tool to integrate into their daily school routines.

MHCP Concerns About Liability

One of the key concerns among participants was the possibility of being liable for the *suicidal thoughts and behaviors* notification embedded in the Innowell eMH platform. Many of the participants shared that they were concerned that a slow response time to a notification that a student is reporting high levels of suicidal thoughts and behaviors may inadvertently prevent timely intervention and therefore contribute to a youth's suicide. This lag time in responding to a notification of acute crisis could be seen as negligence and thus render the MHCP liable. Participants described a concern that youth might erroneously assume that school professionals are immediately made aware that they are in distress or experiencing an emotional crisis after triggering the suicidal thoughts and behaviors notification:

> *[B]ut the consistent challenge that happens over and over again is when the school closes at 3:30 PM, our mental health closes at 4:30 PM, what do these parents—so now we know that there's ideation and we're at high risk. What do we do in [the community] with these students?* [FG22-P02]

Thus, participants were concerned about who would be responsible for responding to youth outside of school hours and during school closures, including weekends and summer holidays.

Unmasking MH Issues in the Face of Limited Services

Finally, many of the participants expressed concern that they may not have the capacity to respond to the needs of the students within their school district. Some of the participants worried that the increased assessments conducted using measurement-based care embedded within the Innowell eMH platform would unmask the degree to which MH problems impact their student population. Increased assessment would inevitably lead to increased identification of students requiring MH care within student MH services in the school environment. However, participants expressed concern that they would not be able to access needed resources, supports, and specialty services for students presenting with MH concerns owing to long wait times and the absence of evidence-based treatments:

> *So, to me it's no different than a lot of rural communities, just a lack of support services to address mental health needs and the negative stigma around mental health and its needs as well.* [FG22-P04]

Participants were concerned about the potential for this to create a crisis in the school district where MHCPs would be inundated with students with MH problems without proper services and resources to refer them to in the community:

> *Like I don't want to say that as "Be careful, let's not [open the floodgates]." I think we just need to be prepared for that and how are we going to support everybody in the meantime or through that?* [FG5-P07]

The knowledge of the shortage of primary care services and specialty clinics within the school and across the MH community created significant reservations among participants in considering using the Innowell eMH platform.

Facilitators to Implementation of eMH Tools in School Settings

Youth Capacity Facilitators

Creates Potential for an Empowered and Active Role in Care

Participants argued that the most important facilitator to implementation is the belief that the Innowell eMH platform may empower youth to take an active role in their care journey. Being able to access the platform independently, completing assessments that are relevant to them, and accessing apps and e-tools that reflect their needs are key facilitators to implementation. Relatedly, participants acknowledged that young people are far more comfortable using technology, and therefore their readiness to adopt the Innowell eMH platform may be high. Many young people are already using technology to complete and upload homework assignments, schedule appointments, and, in some cases, attend appointments and classes. Thus, participants in our study asserted that there was strong potential for young people to use the platform as an extension of these existing practices.

Participants further suggested that integrating an eMH platform into the school system from primary to secondary schools would allow students to incrementally develop self-management skills and scaffold the information and resources they need for wellness and MH care. Using the Innowell eMH platform within the school context would also allow young people to monitor MH changes over time:

> *[A] grade 9 student could carry through this platform for the next 4 years and actually manage their own care.* [FG5-P04]

Collectively, participants viewed the many advantages of eMH technology for their students, including youth empowerment and the ability to acquire self-management skills, resources, and psychoeducation information:

> *But also, empowering youth. I love the idea that they can access this and it's at their fingertips and they can start to really see their growth and that's exciting to me.* [FG5-P01]

Participants acknowledged youth being able to see their growth over time and access apps and e-tools while in the school system as a salient facilitator to them agreeing to adopt and implement eMH technology in their school. Thus, capitalizing on the young people's receptivity to technology, embracing their desire to self-manage, and acknowledging this tool as an opportunity to empower youth are all potential implementation enablers.

Fosters Therapeutic Relationships

Many of the participants highlighted the important ways in which the Innowell eMH platform can foster a stronger therapeutic relationship between themselves and youth. Participants described several ways in which this can occur, including the potential for youth to develop greater MH literacy, which would lead to awareness and the use of terminology about MH and well-being, leading to a shared language with their MHCP and increased collaboration to explore what care is needed:

> *[I]t can be a great tool by the looks of it for youth to identify areas. We might know how they're feeling but not quite able to categorize. "Oh, maybe I'm struggling due to grief," and being able to recognize, name, and have the language for that, I think, can also be an empowering tool.* [FG5-P02]

Many of the participants further highlighted that the repeated use of the measures to assess MH issues might be an effective way for youth to alert their MHCP on how they are coping between sessions or how their MH symptoms are changing over time. Participants described the ease of reviewing assessment information that could be updated between sessions and identifying trends that students may not be verbally sharing with their MHCP:

> *I can see the benefit of it in terms of that—the narrative in between sessions. I often have encouraged the youth that I work with to use the accompanying email to sort of do exactly this piece...there's a narrative in between things that are happening between sessions, I can't answer all the time, but you can just send me little notes to say, "This is for next session." And just sort of in that hopes of creating some sort of form of accountability.* [FG8-P01]

Participants also anticipated the measurement-based care of the Innowell eMH platform providing information on students' progress at different time points during the academic year. In a school setting, the platform can facilitate the sharing of information and the strengthening of the therapeutic relationship over many years. The accessibility of the baseline assessments and records of MH changes over time helps inform the students and their MHCP while also increasing the continuity of care.

Enhances Access to Services and Resources

The final youth capacity facilitator identified by participants is the opportunity to enhance access to information, resources, wellness apps and e-tools, and web-based support by using the Innowell eMH platform. Closely interconnected to the previous facilitator, participants suggested that youth can learn about, explore, and access resources, services, apps, and e-tools more easily by using the platform:

> *I feel like this could be beneficial to them because they can kind of access areas that is of interest or is of concern to them and they can get resources right there and then. That they can click on those apps, those different phone numbers or websites are things that I saw when I was clicking on some of those. And I think that that will be great for them.* [FG10-P01]

Participants recognized that long wait times for specialty services, limited resources, and barriers to accessing MH services, especially in rural areas, might be temporarily addressed by students accessing resources via the Innowell eMH platform.

MHCP Capacity Facilitators

System Transformation Through Flexibility and Problem-Solving

Participants view the school environment as creating the conditions for them to be early adopters of the Innowell eMH platform because educators are flexible, resilient, and solution focused when faced with challenges:

> *Yeah. I think our program is unique just in the flexibility that we have...We have the flexibility that, you know, you meet a student, the next day they're in tears, they can stop by. Or you have a high-risk who could check in in a couple days as opposed to having to wait for 2 weeks.* [FG14-P03]

Participants suggested that by using the Innowell eMH platform, youth would be monitored more carefully, which is different from what might typically occur at the MH clinics in the community. Many of the participants also brainstormed solutions to the barriers listed in the preceding section; for example, a participant brainstormed solutions for youth who may not have access to devices that allow them to use the Innowell eMH platform and apps:

> *[E]ven if they don't have the media devices, there's nothing wrong with using the Chromebook [in the school] and sitting in the next room and working on it.* [FG5-P05]

Despite the many barriers delineated by the school personnel regarding the implementation of the Innowell eMH platform, participants frequently brainstormed solutions to some of their unique challenges while participating in the FGs. This speaks to the flexibility and natural problem-solving ability of school personnel to contribute to creating system transformation.

Collaboration With MHCPs Across the Continuum and With Different Systems

By developing a comprehensive baseline assessment of students using measurement-based care on the Innowell eMH platform, participants believed that they would develop a better understanding of the care required to meet their students' unique needs. The information gleaned from the measurement-based care assessment protocol was further perceived as supporting a stepped and staged care approach. This was highlighted by a participant who described how the information could be used:

> *[F]or like, a clinician's dashboard—to be able to see those alerts right away would help us really prioritize where we need to target.* [FG14-P01]

Participants further highlighted that sharing the results of the assessments and ongoing monitoring of MH symptoms with other MHCPs would allow them to share the mutual understanding and language that would enhance collaborations with other MHCPs:

> *I agree, collaboration and communication is so important. I think being able to get out of our silos and work as a team of professionals will only serve to benefit our young people in need of supports.* [FG9-P02]

Participants were eager to highlight the potential to remove silos from the MH care system by embracing an eMH platform that could be used across MH care systems. Collectively, many of the MHCPs recognized the potential for these factors to foster a therapeutic relationship with youth, especially if there is increased collaboration among MH teams.

Opportunity for the Continuity of Services

Finally, participants shared that the Innowell eMH platform may be used to facilitate the continuity of care across services in the community. Participants highlighted the potential for all members of the care team to share, communicate, and access similar information on the platform about a shared youth client and thus promote the continuity of care. This can help in streamlining the service process by enhancing the continuity of care among services:

> *'Cause we want a wraparound service for the kids and I know there's been lots of privacy issues and that's kind of been better actually for some clinicians—the more open sharing.* [FG9-P04]

Some of the participants suggested that MHCPs may embrace the implementation of the Innowell eMH platform as an opportunity to create change within the health care system:

> *We have a beautiful opportunity here at the middle to drive what happens at the top and at the bottom from a client perspective, our students and our families that'll be impacted and our community and then our leadership. Our superintendents, our mayors, all of those people. And we right now are the driving middle force that can actually change what's happening. And we need to utilize this opportunity well.* [FG5-P04]

Participants of this study strongly advocated that school staff hold a unique role within the community: they establish relationships with a majority of the youth and young adults in every community for many years, while also maintaining relationships with leaders of the school districts and with municipal and provincial levels of government as well, thus providing them with the opportunity to build stronger relationships and drive MH care change by embracing this implementation project.

Discussion

Principal Findings

This study aimed to explore perceived barriers and facilitators relating to the implementation of the Innowell eMH platform in secondary schools in Alberta. Using a descriptive qualitative methodology, we held FGs with key stakeholders in school divisions, including administrators, teachers, management staff, school counselors, psychologists, and community connectors (N=52). Our research shows interconnected barriers and facilitators to implementation as it relates to youth and school MHCP capacities, with system-level considerations. We conclude the discussion with a summary of recommendations for addressing implementation in school settings (Textbox 1).

Textbox 1. Recommendations from the qualitative focus groups with school stakeholders.

Recommendation and description

Clear policies and processes for consent (with regard to accessing mental health [MH] services)

- School-based leaders and decision makers establish policies and processes regarding consenting mature minors and obtaining and navigating parental consent
- Development or tailoring of existing policies and processes in the local context and culture around the consenting process

Web-based training environment

- Create interactive learning activities to enhance educators’ knowledge regarding the application of electronic MH (eMH) tools with diverse students

Level of education, knowledge, and skills

- Train school MH care professional (MHCPs) on how to support youth to identify when and how to share MH issues with caregivers, if appropriate
- Establish communities of practice as an approach to provide post training education and supervisory support to ensure that school personnel can apply their knowledge and skills of measurement-based care and eMH tools

Level of support and supervision

- Decision makers within school settings identify how to support staff to receive adequate training and supervision to learn to use, and embrace the implementation of, eMH tools and apps
- Ongoing mentorship, supervision, and support is needed to integrate eMH tools into the school settings

Structure of the school system and contexts of practice

- Integrated eMH tools fit into established workflows and processes, and work duplication is removed where possible to maximize implementation efforts
- The process of adaption and adoption requires attention to the cultural and contextual components of assessment, formulation, and intervention, including the ways school personnel recognize, explain, and manage distress

Existing socioeconomic barriers to access

- Considering socioeconomic status and access in the communities of implementation is a key pillar of equity that should be addressed in the implementation of eMH tools and measurement-based care
- Emphasize the inclusion and integration of local culture beliefs, practices, language, social norms, family, community, and social network for better understanding of help-seeking behaviors

Address liability concerns and ensure crisis response protocol

- Liability concerns among stakeholders should be heard, integrated, and rapidly addressed through training and clinical supervision to increase willingness to use eMH tools in school settings
- Ensure that MHCPs have the competencies to effectively respond to a student’s disclosure of suicidal thoughts and behaviors via the Innowell eMH platform
- School administrators and decision makers must establish risk mitigation protocols and procedures to assure school MHCPs that clear pathways are determined and easily implemented to rapidly respond to students experiencing suicidal thoughts and behaviors
- Clinical supervision and administrative support must be made available to school MHCPs requiring assistance with students’ disclosures of suicidal thoughts and behaviors and need of acute care

Youth focus

- Keeping youth at the center of eMH implementation strategies could inspire and enliven uptake among MHCPs

Youth engagement

- Use eMH tools to enhance and improve the way that youth and MHCPs interact with each other and the way that MH teams from different systems communicate
- Youth should be included in discussions about how to implement eMH in schools

The first theme highlights concern about the extensive assessments embedded within the Innowell eMH platform and youth capacity (eg, attention span and literacy skills) to complete assessment measures. In general, there is concern about whether the literacy level of the Innowell eMH platform matches that of its intended users in the community. In line with our findings, MH status and demographic variables are among the primary user capacity factors described in the literature that can affect eMH use [26,43]. A systematic review by Borghouts et al [43] found 59 studies that reported that the MH status of the user can play a significant role in the engagement and uptake of eMH tools. Although the severity of MH symptoms can be a barrier to some patients using eMH tools [44,45], this may also depend on the type of eMH tool; for example, individuals with more severe MH symptoms may be more likely to use eMH tools inclusive of assessment measures [43,46]. To achieve this, researchers have co-designed a game that trains clinicians to identify how and when to adopt eMH tools with their clients [47]. A similar type of learning activity can be used to enhance educators' knowledge about the application of eMH tools with diverse students. MHCPs would also benefit from evaluating digital tools through the adoption of adapted rating scales and digital navigators [48].

The results of our study also suggest that school personnel share a concern about how to obtain parental consent and maintain student confidentiality when considering using eMH tools with youth. In keeping with existing research, this finding shows that one of the most significant barriers to young people accessing school MH services is a lack of clarity about confidentiality, especially from their caregivers [10,49]. Young people may avoid accessing school-based MH services and resources if they anticipate negative consequences from their family members [50,51] or if they are required to obtain parental consent [52,53]. This concern also extends to digital MH tools: Cavazos-Rehg et al [51] found that two-thirds of underage youths displaying symptoms of eating disorders were unwilling to obtain parental consent to access a mobile MH intervention. Researchers and clinicians alike strongly advocate that MHCPs use parental consent waivers, reassure young people of their privacy and autonomy, and address adolescent stigma concerns [51] to increase the use of eMH tools. Given the pervasive concerns about confidentiality and the requirement of parental consent to use the Innowell eMH platform in schools, we strongly recommend that school-based leaders and decision makers establish clear policies and processes about consenting mature minors and navigating parental consent. School staff would benefit from informed consent policies, inclusive of teachers, social workers, and administrators that have been approved across school districts in the province or country. We also recommend training school MH personnel on how to support young people to identify when and how to share MH issues with family, if appropriate.

Related to youth capacity barriers, participants were concerned that introducing an eMH platform, the use of which requires access to devices and a stable internet connection, would be unsuitable for many of their youth population, especially for those residing in remote and rural communities. This finding reaffirms the concern about how to integrate eMH technology, given the existing barriers to service access noted in multiple studies [43,54]. Strudwick et al [54] reviewed a total of 31 mobile apps and 114 web-based applications and resources that had the potential to support the MH needs of the broader Canadian population during the COVID-19 pandemic. Key barriers of concern tended to be access, cost, and poor connectivity [54]. Socioeconomic status and access in the communities of implementation are considered key pillars of equity that should be addressed to support the success of future implementation strategies and ensure equitable access of this opportunity [54].

This study points to specific issues and concerns about the lack of available time to build capacity and integrate the Innowell eMH platform into practice. Time constraints, burnout, and change fatigue were also identified as significant barriers to implementation. In alignment with our findings, although high-quality person-centered care is a priority for MH services [55], there are issues pertinent to school MHCPs, such as limited time, competing demands, high caseloads [56], high degrees of burnout [57], and insufficient training and administrative support [58], all of which can create barriers to providing optimal care and adopting a new eMH platform in a system. Furthermore, LaMonica et al [28] argue that if digital solutions are to be successfully used by MH professionals, decision makers must reduce the administrative burden and responsibilities placed on individuals to adopt eMH technology. Importantly, if eMH tools are introduced in an MH service setting, implementation strategies must consider what could be removed or combined to avoid increasing workloads of school MHCPs. We recommend that decision makers within school settings identify how to support staff to receive adequate training and supervision to learn to use, and embrace their curiosity about the implementation of, eMH tools and apps.

From an MHCP capacity perspective, participants expressed concern about how a new eMH platform could be integrated into the existing organizational flow and day-to-day tasks for school MHCPs. If an eMH platform is operationalized into current vision, mission, priorities, and work plans, this is reported to enable implementation and delivery [59]. Operationalization factors that increase implementation are reported to include workflow processes; leadership, including workplace culture and management; and systems, including the organization of people and resources to meet the needs of the community [59]. Greenhalgh et al [60] draw attention to the importance of integrating technological advances in MH care into the work processes and existing tools and resources used by MHCPs. When eMH tools do not fit into traditional workflows and processes, there is a risk of low engagement and poorly sustained implementation once trials end [17,60,61]. Thus, eMH initiatives must fit into the standard workflows of the health system setting [61,62] to improve youth MH outcomes [63]. Work duplication could be removed by reducing administrative burden on professionals by ensuring the interoperability of MH tools with existing management systems (removing the need to enter the same data across multiple systems) [28]. This finding is particularly meaningful for school district professionals who already use existing data management processes, as well as MH tools and resources, and are concerned about the integration of eMH tools into the established organizational workflow. Our findings support the recommendation that eMH tools must fit into established workflows and processes and work duplication removed where possible to maximize implementation efforts.

A major implementation barrier concern among participants is the potential to be held liable for a youth's suicidal thoughts and behaviors notification alerted through the platform. Although liability concerns have not been sufficiently discussed in the literature, Scott et al [64] examined telehealth policy implications and suggested that some risk should be anticipated and expected in the implementation of eMH tools. Thus, liability concerns among stakeholders should be heard and considered and rapidly addressed through training and clinical supervision to increase willingness to use eMH tools in school settings. Our research sheds light on key liability concerns among school staff that should be considered and urgently addressed during the preimplementation phase to ensure that school MHCPs have the competencies to effectively respond to a student's disclosure of suicidal thoughts and behaviors via the Innowell eMH platform. This points to a key improvement in quality care by providing young people the opportunity to assess and detect suicidality and, furthermore, empowering the young person to seek support and care pathways for suicidality [65]. Equally important, school administrators and decision makers must establish risk mitigation protocols and procedures to assure school MHCPs that clear pathways are determined and easily implemented to rapidly respond to students experiencing suicidal thoughts and behaviors. A systematic review of MH training for secondary teachers shows that most training interventions have been carried out through facilitated course trainings and workshops, such as MH first aid training, peer support, and suicide prevention and postvention, to name a few [66]. This review showed an improvement in MH knowledge and attitudes among teachers, and the interventions reviewed should be considered in training and preparing schools to implement eMH tools, especially when suicide assessments and alert systems are included in the eMH options. Finally, clinical supervision and administrative support must be made available to school MHCPs requiring assistance with students' disclosures of suicidal thoughts and behaviors and need of acute care. An area of future research is to review the legal and ethical considerations of telehealth services [64] across different systems to facilitate the successful implementation of eMH tools.

Many of the participants also expressed concern that they may not have the capacity to respond to the needs of the students within their school district owing to limited resources in their school and the broader community. These concerns regarding the availability and accessibility of resources are reported to be challenges with accessing MH services across the globe [67]. Research has shown that school MHCPs tend to have variable or insufficient training coupled with a lack of support from administration, affecting their ability to respond to the diverse needs of students in the school setting [58]. Without the resources available to meet the needs of students who are identified as requiring more MH support, there is great concern that MH symptoms and risk of suicide will worsen [68]. Our findings suggest that early assessment and intervention, including the capacity to respond to the needs of young people and potentially refer them on to additional resources, are critical to the improvement of long-term health and social outcomes [69]. Introducing an eMH platform in a school setting may inevitably lead to the identification of youth who urgently need MH care, validating this concern. Alternatively, the Innowell eMH platform may also identify youth who are able to self-manage through the platform's apps and electronic resources, preserving current MH services for those individuals in greatest need. Furthermore, gaining an understanding of the number of youth requiring MH services provides a starting point for advocating for increased publicly funded MH services. This advocacy can be done by capitalizing on the surge of interest in digital MH tools through building awareness regarding ways to modernize access to MH services, highlighting evidence of the effectiveness of digital MH tools, and advocating for financial investment [14].

When discussing facilitators to the implementation of eMH tools, participants described the capacity among young people to use technology to access services. We learned that participants view young people as having a strong desire to direct and manage their own care. The ability of young people to take an active role in their own care journey and access resources and information at times that work for them, whether independently or with their MHCP, all highlight the importance of youth capacity as a facilitator to the implementation of the Innowell eMH platform. In line with this finding, Iorfino et al [70] suggest that health systems will see an increased push toward youth owning and managing their own health data to the benefit of both youth and MHCPs. Many studies have addressed the importance of eMH tools as empowering their users by affording more control, choice [71], and shared decision-making [72]. By contrast, MHCPs who are early adopters of digital tools tend to be those who believe that the initiatives will be beneficial to their clients [73]. The scoping review by Hawke et al [74] demonstrated that youth prefer to provide feedback on the care they receive because they have a strong desire to be involved in decision-making. Playing an active role in their care enables young people to cope better, increases their sense of empowerment, and strengthens their connections with health professionals [75]. As MHCPs expressed enthusiasm around the implementation of the Innowell eMH platform, particularly when considering its potential benefits for their students, keeping youth at the center of eMH implementation strategies could inspire and enliven uptake among school MHCPs.

The findings of our study suggest that school MHCPs may use measurement-based care through an eMH platform to identify and monitor how students' MH concerns are progressing and where new MH problems may be emerging. In fact, our study demonstrated that MHCPs view the measurement-based care protocol embedded in the Innowell eMH platform as providing critical information about a young person's MH status, a common language for talking about MH concerns, and when changes in MH symptoms are a signal for adaptations in the intensity of services required. This common language as well as the ability to communicate, understand, and interpret their problems and strengths, subjective symptoms, and preferences for care with their MHCP could be established in their collaboration with other MHCPs as well [76]. The changes or lack of changes in the MH symptoms could open collaboration through thoughtful conversations between youth and their MHCP about the direction for care moving forward [77]. To inspire uptake among MHCPs, we support stressing the importance of eMH tools enhancing and improving the way that youth and MHCPs interact with each other and the way that MH teams from different systems communicate, where possible [78].

This study further highlighted organizational implications to the implementation of eMH tools, recognizing the potential for increased access to community-based services and resources and available apps and e-tools. In a scoping review of MH service-level factors for access and engagement for young people, Anderson et al [79] note that choices around resources, increased information, flexible treatment delivery, and person-centered care contributed to young people being engaged in MH services. These benefits also extend to MHCPs and service delivery by helping to better connect young people to the services that they need and providing options to explore when their MHCP is unavailable. Study participants highlighted the potential for all members of the school MH team to share and access similar information about a student to ensure the continuity of care. Furthermore, participants also perceived school MHCPs being involved in the implementation of eMH tools as a turning point for MH services in the community. The unique relationship of school districts with municipal and provincial levels of government was seen as an implementation facilitator that could support significant transformation of youth MH services. Davenport et al [15] demonstrate how primary MH services can be "flipped" using digital health tools. Their work highlights the potential of digital MH services when assessment, triage, and care pathways are created to ensure that young people are matched to the care they need [15]. More recently, the Innowell eMH platform has been used to compare the needs of clients among services and across geographic regions, which can be used to understand the needs of a heterogeneous population and plan services accordingly [80]. This highlights the potential of eMH tools to ensure that youth are accessing the right care at the right time and promoting the continuity of care to avoid the fragmentation of services [81]. Our findings point to the potential of the school setting to support "flipping" youth MH service experiences and outcomes. The school setting is of particular importance, given the unique community-based relationships formed with young people and the potential to participate in deciding the direction of youth MH strategies with local and provincial governments, especially through the implementation of new eMH tools.

Limitations and Strengths

This study has a few limitations and strengths. Dynamics varied and differed among the FGs, with some groups having more representation from diverse stakeholders than others despite our best efforts to encourage diversity. An open discussion was embraced with each group; therefore, some groups spoke more organically about systemic and organizational implementation challenges than others. A potential limitation is that the clinical lead was often present in the FGs, which may have prevented some people from being forthright about the barriers and enablers specific to their community context as well as their views of the implementation of eMH tools and organizational culture. FGs may also result in certain types of socially acceptable opinions emerging and certain participants dominating. Therefore, this may not be the collective voice we expected. Some strengths of this study that we would like to highlight are the intentional involvement of youth research partners in the analysis of the FG data. This study also recruited from diverse school settings with respect to size, approach, and location (both urban and rural areas in the province). An opportunity for future research would be to target more diverse individuals and ask more targeted questions about organizational and systemic barriers and facilitators experienced by MHCPs and youth, respectively. Future studies could also use multiple methods to increase validity, such as observations and interviews. In addition, this gives way to include young people with lived experiences as participants, which was done as part of this study, and these results will be published in a subsequent manuscript.

Conclusions

This study sought to explore school MHCPs' perspectives relating to the implementation of the Innowell eMH platform. Schools are a critical setting to implement eMH tools for youth. Our findings highlight the nuanced perspectives among MHCPs with regard to implementation. Their insights demonstrate critical youth and MHCP concerns, with considerations for organizational-level factors that may impede or enhance the implementation processes for embedding eMH tools in the school context. The identified barriers and facilitators to implementation in a school setting provide future researchers and decision makers with expected (and unexpected) challenges that could be addressed in the preimplementation phase.

Acknowledgments

The authors are grateful for the contributions from their research assistants Brea McLaughlin, Simron Sidhu, Katelyn Greer, Karina Pintson, and Sarah Daniel. The authors acknowledge the contributions of the team implementing the electronic mental health tools in coordinating and delivering the focus group presentations on the Innowell electronic mental health platform and providing manuscript feedback. The authors could not have conducted this research without the support of the various stakeholders and participants of this study. Finally, the authors would like to acknowledge the Alberta Children's Hospital Foundation and the Partnership for Research and Innovation in the Health System for funding and supporting the project.

Data Availability

The data generated and analyzed during this study are restricted because the transcripts in entirety could potentially identify participants and are confidential.

Authors' Contributions

GD, EMB, and KSB wrote the initial draft of the manuscript. All authors participated in the refinement of the draft, critically reviewed it, and provided feedback on the final version submitted for publication in accordance with the International Committee of Medical Journal Editors criteria.

Conflicts of Interest

HML is a Section Editor for JMIR Aging (at the time of this publication). All other authors declare no other conflicts of interest.

References

1. Racine N, McArthur BA, Cooke JE, Eirich R, Zhu J, Madigan S. Global prevalence of depressive and anxiety symptoms in children and adolescents during COVID-19: a meta-analysis. JAMA Pediatr 2021 Nov 01;175(11):1142-1150 [FREE Full text] [doi: 10.1001/jamapediatrics.2021.2482] [Medline: 34369987]
2. Madigan S, Korczak DJ, Vaillancourt T, Racine N, Hopkins WG, Pador P, et al. Comparison of paediatric emergency department visits for attempted suicide, self-harm, and suicidal ideation before and during the COVID-19 pandemic: a systematic review and meta-analysis. Lancet Psychiatry 2023 May;10(5):342-351 [FREE Full text] [doi: 10.1016/S2215-0366(23)00036-6] [Medline: 36907199]
3. Solmi M, Radua J, Olivola M, Croce E, Soardo L, Salazar de Pablo G, et al. Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. Mol Psychiatry 2022 Jan;27(1):281-295 [FREE Full text] [doi: 10.1038/s41380-021-01161-7] [Medline: 34079068]
4. Jones EA, Mitra AK, Bhuiyan AR. Impact of COVID-19 on mental health in adolescents: a systematic review. Int J Environ Res Public Health 2021 Mar 03;18(5):2470 [FREE Full text] [doi: 10.3390/ijerph18052470] [Medline: 33802278]
5. Merikangas KR, Nakamura EF, Kessler RC. Epidemiology of mental disorders in children and adolescents. Dialogues Clin Neurosci 2009;11(1):7-20 [FREE Full text] [doi: 10.31887/DCNS.2009.11.1/krmerikangas] [Medline: 19432384]
6. Orgilés M, Morales A, Delvecchio E, Francisco R, Mazzeschi C, Pedro M, et al. Coping behaviors and psychological disturbances in youth affected by the COVID-19 health crisis. Front Psychol 2021 Mar 22;12:565657 [FREE Full text] [doi: 10.3389/fpsyg.2021.565657] [Medline: 33828499]
7. Higuchi S, Mihara S, Kitayuguchi T, Miyakoshi H, Ooi M, Maezono M, et al. Prolonged use of internet and gaming among treatment seekers arising out of social restrictions related to COVID-19 pandemic. Psychiatry Clin Neurosci 2020 Nov;74(11):607-608 [FREE Full text] [doi: 10.1111/pcn.13127] [Medline: 32767600]
8. Zhu S, Zhuang Y, Lee P, Li JC, Wong PW. Leisure and problem gaming behaviors among children and adolescents during school closures caused by COVID-19 in Hong Kong: quantitative cross-sectional survey study. JMIR Serious Games 2021 May 07;9(2):e26808 [FREE Full text] [doi: 10.2196/26808] [Medline: 33960954]
9. Sheikhan NY, Henderson JL, Halsall T, Daley M, Brownell S, Shah J, et al. Stigma as a barrier to early intervention among youth seeking mental health services in Ontario, Canada: a qualitative study. BMC Health Serv Res 2023 Jan 26;23(1):86 [FREE Full text] [doi: 10.1186/s12913-023-09075-6] [Medline: 36703119]
10. Radez J, Reardon T, Creswell C, Lawrence PJ, Evdoka-Burton G, Waite P. Why do children and adolescents (not) seek and access professional help for their mental health problems? a systematic review of quantitative and qualitative studies. Eur Child Adolesc Psychiatry 2021 Feb;30(2):183-211 [FREE Full text] [doi: 10.1007/s00787-019-01469-4] [Medline: 31965309]
11. Kourgiantakis T, Markoulakis R, Lee E, Hussain A, Lau C, Ashcroft R, et al. Access to mental health and addiction services for youth and their families in Ontario: perspectives of parents, youth, and service providers. Int J Ment Health Syst 2023 Mar 14;17(1):4 [FREE Full text] [doi: 10.1186/s13033-023-00572-z] [Medline: 36918893]
12. Punton G, Dodd AL, McNeill A. 'You're on the waiting list': an interpretive phenomenological analysis of young adults' experiences of waiting lists within mental health services in the UK. PLoS One 2022 Mar 18;17(3):e0265542 [FREE Full text] [doi: 10.1371/journal.pone.0265542] [Medline: 35303040]
13. Nicholas J, Bell IH, Thompson A, Valentine L, Simsir P, Sheppard H, et al. Implementation lessons from the transition to telehealth during COVID-19: a survey of clinicians and young people from youth mental health services. Psychiatry Res 2021 May;299:113848 [FREE Full text] [doi: 10.1016/j.psychres.2021.113848] [Medline: 33725578]
14. Moroz N, Moroz I, D'Angelo MS. Mental health services in Canada: barriers and cost-effective solutions to increase access. Healthc Manage Forum 2020 Nov;33(6):282-287. [doi: 10.1177/0840470420933911] [Medline: 32613867]
15. Davenport TA, Cheng VW, Iorfino F, Hamilton B, Castaldi E, Burton A, et al. Flip the clinic: a digital health approach to youth mental health service delivery during the COVID-19 pandemic and beyond. JMIR Ment Health 2020 Dec 15;7(12):e24578 [FREE Full text] [doi: 10.2196/24578] [Medline: 33206051]
16. Lal S, Dell'Elce J, Tucci N, Fuhrer R, Tamblyn R, Malla A. Preferences of young adults with first-episode psychosis for receiving specialized mental health services using technology: a survey study. JMIR Ment Health 2015 May 20;2(2):e18 [FREE Full text] [doi: 10.2196/mental.4400] [Medline: 26543922]

17. Mohr DC, Weingardt KR, Reddy M, Schueller SM. Three problems with current digital mental health research . . . and three things we can do about them. Psychiatr Serv 2017 May 01;68(5):427-429 [FREE Full text] [doi: 10.1176/appi.ps.201600541] [Medline: 28412890]
18. Firth J, Torous J, Nicholas J, Carney R, Pratap A, Rosenbaum S, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry 2017 Oct;16(3):287-298 [FREE Full text] [doi: 10.1002/wps.20472] [Medline: 28941113]
19. Firth J, Torous J, Nicholas J, Carney R, Rosenbaum S, Sarris J. Can smartphone mental health interventions reduce symptoms of anxiety? a meta-analysis of randomized controlled trials. J Affect Disord 2017 Aug 15;218:15-22 [FREE Full text] [doi: 10.1016/j.jad.2017.04.046] [Medline: 28456072]
20. Kruzan KP, Fitzsimmons-Craft EE, Dobias M, Schleider JL, Pratap A. Developing, deploying, and evaluating digital mental health interventions in spaces of online help- and information-seeking. Procedia Comput Sci 2022;206:6-22 [FREE Full text] [doi: 10.1016/j.procs.2022.09.081] [Medline: 37063642]
21. Georgiades K, Duncan L, Wang L, Comeau J, Boyle MH, 2014 Ontario Child Health Study Team. Six-month prevalence of mental disorders and service contacts among children and youth in Ontario: evidence from the 2014 Ontario child health study. Can J Psychiatry 2019 Apr;64(4):246-255 [FREE Full text] [doi: 10.1177/0706743719830024] [Medline: 30978138]
22. Rickwood D, Deane FP, Wilson CJ, Ciarrochi J. Young people’s help-seeking for mental health problems. AeJAMH 2014 Dec 17;4(3):218-251 [FREE Full text] [doi: 10.5172/jamh.4.3.218]
23. Rowling L, Whitman CV, Biewener M. International survey of principals concerning emotional and mental health and well-being. International Association of Child and Adolescent Mental Health and Schools (Intercamhs) and International Confederation of Principals (ICP). 2009. URL: https://intercamhs.edc.org/files/2009%20International%20Principals%20Survey%20-%20ICP%20and%20Intercamhs%20-%20Major%20Findings%20Report.pdf [accessed 2023-12-29]
24. Dimitropoulos G, Cullen E, Cullen O, Pawluk C, McLuckie A, Patten S, et al. “Teachers often see the red flags first”: perceptions of school staff regarding their roles in supporting students with mental health concerns. School Ment Health 2021 Sep 23;14(2):402-415 [FREE Full text] [doi: 10.1007/s12310-021-09475-1]
25. Iorfino F, Scott EM, Carpenter JS, Cross SP, Hermens DF, Killedar M, et al. Clinical stage transitions in persons aged 12 to 25 years presenting to early intervention mental health services with anxiety, mood, and psychotic disorders. JAMA Psychiatry 2019 Nov 01;76(11):1167-1175 [FREE Full text] [doi: 10.1001/jamapsychiatry.2019.2360] [Medline: 31461129]
26. Meurk C, Leung J, Hall W, Head BW, Whiteford H. Establishing and governing e-mental health care in Australia: a systematic review of challenges and a call for policy-focussed research. J Med Internet Res 2016 Jan 13;18(1):e10 [FREE Full text] [doi: 10.2196/jmir.4827] [Medline: 26764181]
27. Roberts M, Jones J, Garcia L, Techau A. Adolescents' perceptions of barriers and facilitators to engaging in mental health treatment: a qualitative meta-synthesis. J Child Adolesc Psychiatr Nurs 2022 May;35(2):113-125. [doi: 10.1111/jcap.12354] [Medline: 34609034]
28. LaMonica HM, Iorfino F, Lee GY, Piper S, Occhipinti JA, Davenport TA, et al. Informing the future of integrated digital and clinical mental health care: synthesis of the outcomes from project synergy. JMIR Ment Health 2022 Mar 09;9(3):e33060 [FREE Full text] [doi: 10.2196/33060] [Medline: 34974414]
29. Davenport TA, LaMonica HM, Whittle L, English A, Iorfino F, Cross S, et al. Validation of the InnoWell platform: protocol for a clinical trial. JMIR Res Protoc 2019 May 31;8(5):e13955 [FREE Full text] [doi: 10.2196/13955] [Medline: 31152524]
30. Ospina-Pinillos L, Davenport T, Iorfino F, Tickell A, Cross S, Scott EM, et al. Using new and innovative technologies to assess clinical stage in early intervention youth mental health services: evaluation study. J Med Internet Res 2018 Sep 10;20(9):e259 [FREE Full text] [doi: 10.2196/jmir.9966] [Medline: 30201602]
31. Kim H, Sefcik JS, Bradway C. Characteristics of qualitative descriptive studies: a systematic review. Res Nurs Health 2017 Feb;40(1):23-42 [FREE Full text] [doi: 10.1002/nur.21768] [Medline: 27686751]
32. Colorafi KJ, Evans B. Qualitative descriptive methods in health science research. HERD 2016 Jul;9(4):16-25 [FREE Full text] [doi: 10.1177/1937586715614171] [Medline: 26791375]
33. Siedlecki SL. Understanding descriptive research designs and methods. Clin Nurse Spec 2020;34(1):8-12. [doi: 10.1097/NUR.0000000000000493] [Medline: 31789957]
34. Kaplowitz MD, Hoehn JP. Do focus groups and individual interviews reveal the same information for natural resource valuation? Ecol Econ 2001 Feb;36(2):237-247 [FREE Full text] [doi: 10.1016/s0921-8009(00)00226-3]
35. Cane J, O'Connor D, Michie S. Validation of the theoretical domains framework for use in behaviour change and implementation research. Implement Sci 2012 Apr 24;7:37. [doi: 10.1186/1748-5908-7-37] [Medline: 22530986]
36. Atkins L, Francis J, Islam R, O'Connor D, Patey A, Ivers N, et al. A guide to using the Theoretical Domains Framework of behaviour change to investigate implementation problems. Implement Sci 2017 Jun 21;12(1):77 [FREE Full text] [doi: 10.1186/s13012-017-0605-9] [Medline: 28637486]
37. Wallace C, Farmer J, McCosker A. Boundary spanning practices of community connectors for engaging 'hardly reached' people in health services. Soc Sci Med 2019 Jul;232:366-373. [doi: 10.1016/j.socscimed.2019.05.034] [Medline: 31132544]
38. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]

39. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101 [FREE Full text] [doi: 10.1191/1478088706qp063oa]
40. Terry G, Hayfield N, Clarke V, Braun V. Thematic analysis. In: Willig C, Rogers WS, editors. The SAGE Handbook of Qualitative Research in Psychology. Thousand Oaks, CA: SAGE Publications Ltd; 2017.
41. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
42. Gibbs GR. Chapter 4: thematic coding and categorizing. In: Analyzing Qualitative Data. Thousand Oaks, CA: SAGE Publications, Inc; 2007.
43. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
44. Doherty G, Coyle D, Sharry J. Engagement with online mental health interventions: an exploratory clinical study of a treatment for depression. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2012 Presented at: CHI '12: CHI Conference on Human Factors in Computing Systems; May 5-10, 2012; Austin, TX. [doi: 10.1145/2207676.2208602]
45. Forchuk C, Reiss J, Eichstedt J, Singh D, Collins K, Rudnick A, et al. The youth-mental health engagement network: an exploratory pilot study of a smartphone and computer-based personal health record for youth experiencing depressive symptoms. Int J Ment Health 2016;45(3):205-222 [FREE Full text] [doi: 10.1080/00207411.2016.1204823]
46. Crisp DA, Griffiths KM. Participating in online mental health interventions: who is most likely to sign up and why? Depress Res Treat 2014;2014:790457 [FREE Full text] [doi: 10.1155/2014/790457] [Medline: 24804089]
47. Bierbooms J, Feijt MA, IJsselsteijn WA, Bongers IM. Design of a game-based training environment to enhance mental health care professionals' skills in using e-mental health: multiple methods user requirements analysis. JMIR Serious Games 2022 Jul 27;10(3):e34700 [FREE Full text] [doi: 10.2196/34700] [Medline: 35896032]
48. Roberts AE, Davenport TA, Wong T, Moon HW, Hickie IB, LaMonica HM. Evaluating the quality and safety of health-related apps and e-tools: adapting the mobile app rating scale and developing a quality assurance protocol. Internet Interv 2021 Mar 17;24:100379 [FREE Full text] [doi: 10.1016/j.invent.2021.100379] [Medline: 33777705]
49. Kay CM, Morgan DL, Tripp JH, Davies C, Sykes S. To what extent are school drop-in clinics meeting pupils' self-identified health concerns? Health Educ J 2006;65(3):236-251 [FREE Full text] [doi: 10.1177/0017896906067775]
50. Chandra A, Minkovitz CS. Factors that influence mental health stigma among 8th grade adolescents. J Youth Adolesc 2006 Jul 25;36(6):763-774 [FREE Full text] [doi: 10.1007/s10964-006-9091-0]
51. Cavazos-Rehg P, Min C, Fitzsimmons-Craft EE, Savoy B, Kaiser N, Riordan R, et al. Parental consent: a potential barrier for underage teens' participation in an mHealth mental health intervention. Internet Interv 2020 Sep;21:100328 [FREE Full text] [doi: 10.1016/j.invent.2020.100328] [Medline: 32528858]
52. Evans SW. Mental health services in schools: utilization, effectiveness, and consent. Clin Psychol Rev 1999 Mar;19(2):165-178. [doi: 10.1016/s0272-7358(98)00069-5] [Medline: 10078418]
53. Planey AM, Smith SM, Moore S, Walker TD. Barriers and facilitators to mental health help-seeking among African American youth and their families: a systematic review study. Child Youth Serv Rev 2019 Jun;101:190-200 [FREE Full text] [doi: 10.1016/j.childyouth.2019.04.001]
54. Strudwick G, Sockalingam S, Kassam I, Sequeira L, Bonato S, Youssef A, et al. Digital interventions to support population mental health in Canada during the COVID-19 pandemic: rapid review. JMIR Ment Health 2021 Mar 02;8(3):e26550 [FREE Full text] [doi: 10.2196/26550] [Medline: 33650985]
55. Fukui S, Salyers MP, Morse G, Rollins AL. Factors that affect quality of care among mental health providers: focusing on job stress and resources. Psychiatr Rehabil J 2021 Sep;44(3):204-211 [FREE Full text] [doi: 10.1037/prj0000469] [Medline: 33539113]
56. Atkinson C, Squires G, Bragg J, Muscutt J, Wasilewski D. Facilitators and barriers to the provision of therapeutic interventions by school psychologists. Sch Psychol Int 2014 Aug;35(4):384-397 [FREE Full text] [doi: 10.1177/0143034313485849] [Medline: 26412911]
57. Kim N, Lambie GW. Burnout and implications for professional school counselors. Prof Couns 2018;8(3):277-294. [doi: 10.15241/nk.8.3.277]
58. Suldo SM, Friedrich A, Michalowski J. Personal and systems‐level factors that limit and facilitate school psychologists' involvement in school‐based mental health services. Psychol Sch 2010 Apr;47(4):354-373 [FREE Full text] [doi: 10.1002/pits.20475]
59. Vis C, Mol M, Kleiboer A, Bührmann L, Finch T, Smit J, et al. Improving implementation of eMental health for mood disorders in routine practice: systematic review of barriers and facilitating factors. JMIR Ment Health 2018 Mar 16;5(1):e20 [FREE Full text] [doi: 10.2196/mental.9769] [Medline: 29549072]
60. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. J Med Internet Res 2017 Nov 01;19(11):e367 [FREE Full text] [doi: 10.2196/jmir.8775] [Medline: 29092808]

61. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
62. Joshi G, Sharma G. Burnout: a risk factor amongst mental health professionals during COVID-19. Asian J Psychiatr 2020 Dec;54:102300 [FREE Full text] [doi: 10.1016/j.ajp.2020.102300] [Medline: 32683251]
63. Hetrick SE, Bailey AP, Smith KE, Malla A, Mathias S, Singh SP, et al. Integrated (one-stop shop) youth health care: best available evidence and future directions. Med J Aust 2017 Nov 20;207(10):S5-18. [doi: 10.5694/mja17.00694] [Medline: 29129182]
64. Scott RE, Jennett PA, Hebert M, Rush B. Telehealth and e-health policy considerations for Alberta. Faculty of Medicine, University of Calgary Research and Training Program. 2004 Aug. URL: https://prism.ucalgary.ca/server/api/core/bitstreams/6906529d-4483-4284-9214-6512b5c657e4/content [accessed 2024-01-04]
65. Iorfino F, Davenport TA, Ospina-Pinillos L, Hermens DF, Cross S, Burns J, et al. Using new and emerging technologies to identify and respond to suicidality among help-seeking young people: a cross-sectional study. J Med Internet Res 2017 Jul 12;19(7):e247 [FREE Full text] [doi: 10.2196/jmir.7897] [Medline: 28701290]
66. Anderson M, Werner-Seidler A, King C, Gayed A, Harvey SB, O'Dea B. Mental health training programs for secondary school teachers: a systematic review. Sch Ment Health 2018 Oct 3;11(3):489-508 [FREE Full text] [doi: 10.1007/s12310-018-9291-2]
67. Ćosić K, Popović S, Šarlija M, Kesedžić I. Impact of human disasters and COVID-19 pandemic on mental health: potential of digital psychiatry. Psychiatr Danub 2020;32(1):25-31 [FREE Full text] [doi: 10.24869/psyd.2020.25] [Medline: 32303026]
68. Chevance A, Gourion D, Hoertel N, Llorca PM, Thomas P, Bocher R, et al. [Ensuring mental health care during the SARS-CoV-2 epidemic in France: a narrative review]. Encephale 2020 Jun;46(3S):S3-13 [FREE Full text] [doi: 10.1016/j.encep.2020.03.001] [Medline: 32312567]
69. Golberstein E, Wen H, Miller BF. Coronavirus disease 2019 (COVID-19) and mental health for children and adolescents. JAMA Pediatr 2020 Sep 01;174(9):819-820. [doi: 10.1001/jamapediatrics.2020.1456] [Medline: 32286618]
70. Iorfino F, Piper SE, Prodan A, LaMonica HM, Davenport TA, Lee GY, et al. Using digital technologies to facilitate care coordination between youth mental health services: a guide for implementation. Front Health Serv 2021 Nov 18;1:745456 [FREE Full text] [doi: 10.3389/frhs.2021.745456] [Medline: 36926493]
71. Bucci S, Berry N, Morris R, Berry K, Haddock G, Lewis S, et al. "They are not hard-to-reach clients. We have just got hard-to-reach services." Staff views of digital health tools in specialist mental health services. Front Psychiatry 2019 May 10;10:344 [FREE Full text] [doi: 10.3389/fpsyt.2019.00344] [Medline: 31133906]
72. Berry N, Lobban F, Bucci S. A qualitative exploration of service user views about using digital health interventions for self-management in severe mental health problems. BMC Psychiatry 2019 Jan 21;19(1):35 [FREE Full text] [doi: 10.1186/s12888-018-1979-1] [Medline: 30665384]
73. Waneka R, Spetz J. Hospital information technology systems' impact on nurses and nursing care. J Nurs Adm 2010 Dec;40(12):509-514. [doi: 10.1097/NNA.0b013e3181fc1a1c] [Medline: 21084885]
74. Hawke LD, Mehra K, Settipani C, Relihan J, Darnay K, Chaim G, et al. What makes mental health and substance use services youth friendly? a scoping review of literature. BMC Health Serv Res 2019 Apr 27;19(1):257 [FREE Full text] [doi: 10.1186/s12913-019-4066-5] [Medline: 31029109]
75. Muir K, Powell A, McDermott S. 'They don't treat you like a virus': youth-friendly lessons from the Australian National Youth Mental Health Foundation. Health Soc Care Community 2012 Mar;20(2):181-189. [doi: 10.1111/j.1365-2524.2011.01029.x] [Medline: 21929697]
76. Brämberg EB, Sandman L. Communication through in-person interpreters: a qualitative study of home care providers' and social workers' views. J Clin Nurs 2013 Jan;22(1-2):159-167. [doi: 10.1111/j.1365-2702.2012.04312.x] [Medline: 23170919]
77. Rush AJ, Carmody TJ, Ibrahim HM, Trivedi MH, Biggs MM, Shores-Wilson K, et al. Comparison of self-report and clinician ratings on two inventories of depressive symptomatology. Psychiatr Serv 2006 Jun;57(6):829-837. [doi: 10.1176/ps.2006.57.6.829] [Medline: 16754760]
78. Osma J, Sprenger M, Mettler T. Introduction of e-mental health in national health systems – a health professionals' perspective. Health Policy Technol 2017 Dec;6(4):436-445 [FREE Full text] [doi: 10.1016/j.hlpt.2017.07.001]
79. Anderson JK, Howarth E, Vainre M, Jones PB, Humphrey A. A scoping literature review of service-level barriers for access and engagement with mental health services for children and young people. Child Youth Serv Rev 2017 Jun;77:164-176 [FREE Full text] [doi: 10.1016/j.childyouth.2017.04.017]
80. Capon W, Hickie IB, McKenna S, Varidel M, Richards M, LaMonica HM, et al. Characterising variability in youth mental health service populations: a detailed and scalable approach using digital technology. Australas Psychiatry 2023 Jun;31(3):295-301 [FREE Full text] [doi: 10.1177/10398562231167681] [Medline: 37035873]
81. Almeda N, García-Alonso C, Salvador-Carulla L. Mental health planning at a very early stage of the COVID-19 crisis: a systematic review of online international strategies and recommendations. BMC Psychiatry 2021 Jan 15;21(1):43 [FREE Full text] [doi: 10.1186/s12888-020-03015-y] [Medline: 33451305]

Abbreviations

eMH: electronic mental health
e-tools: electronic tools
FG: focus group
MH: mental health
MHCP: mental health care professional
TDF: theoretical domains framework

Edited by J Torous; submitted 31.05.23; peer-reviewed by M Gasmi , N Esfandiari, S McKenna, S Zhu; comments to author 18.10.23; revised version received 06.11.23; accepted 01.12.23; published 17.01.24.

Please cite as:
Dimitropoulos G, Bassi EM, Bright KS, Gondziola J, Bradley J, Fersovitch M, Stamp L, LaMonica HM, Iorfino F, Gaskell T, Tomlinson S, Johnson DW
Implementation of an Electronic Mental Health Platform for Youth and Young Adults in a School Context Across Alberta, Canada: Thematic Analysis of the Perspectives of Stakeholders
JMIR Ment Health 2024;11:e49099
URL: https://mental.jmir.org/2024/1/e49099
doi:10.2196/49099
PMID:38231558

Original Paper

Reasons for Acceptance or Rejection of Online Record Access Among Patients Affected by a Severe Mental Illness: Mixed Methods Study

Julian Schwarz[1,2,3], MD; Eva Meier-Diedrich[1,3], MSc; Katharina Neumann[1], MSc; Martin Heinze[1,2,3], MD; Yvonne Eisenmann[2,3], PhD; Samuel Thoma[1,3], MD, PhD

[1]Department of Psychiatry and Psychotherapy, Center for Mental Health, Immanuel Hospital Rüdersdorf, Brandenburg Medical School Theodor Fontane, Rüdersdorf, Germany

[2]Center for Health Service Research Brandenburg, Brandenburg Medical School Theodor Fontane, Rüdersdorf, Germany

[3]Faculty of Health Sciences Brandenburg, Brandenburg Medical School Theodor Fontane, Neuruppin, Germany

Corresponding Author:
Julian Schwarz, MD
Department of Psychiatry and Psychotherapy
Center for Mental Health, Immanuel Hospital Rüdersdorf
Brandenburg Medical School Theodor Fontane
Seebad 82/83
Rüdersdorf, 15562
Germany
Phone: 49 33638 83 5
Email: julian.schwarz@mhb-fontane.de

Abstract

Background: Over the past few years, online record access (ORA) has been established through secure patient portals in various countries, allowing patients to access their health data, including clinical notes ("open notes"). Previous research indicates that ORA in mental health, particularly among patients with severe mental illness (SMI), has been rarely offered. Little is known about the expectations and motivations of patients with SMI when reading what their clinicians share via ORA.

Objective: The aim of this study is to explore the reasons why patients with SMI consider or reject ORA and whether sociodemographic characteristics may influence patient decisions.

Methods: ORA was offered to randomly selected patients at 3 university outpatient clinics in Brandenburg, Germany, which exclusively treat patients with SMI. Within the framework of a mixed methods evaluation, qualitative interviews were conducted with patients who chose to participate in ORA and those who declined, aiming to explore the underlying reasons for their decisions. The interviews were transcribed and analyzed using thematic analysis. Sociodemographic characteristics of patients were examined using descriptive statistics to identify predictors of acceptance or rejection of ORA.

Results: Out of 103 included patients, 58% (n=60) wished to read their clinical notes. The reasons varied, ranging from a desire to engage more actively in their treatment to critically monitoring it and using the accessible data for third-party purposes. Conversely, 42% (n=43) chose not to use ORA, voicing concerns about possibly harming the trustful relationship with their clinicians as well as potential personal distress or uncertainty arising from reading the notes. Practical barriers such as a lack of digital literacy or suspected difficult-to-understand medical language were also named as contributing factors. Correlation analysis revealed that the majority of patients with depressive disorder desired to read the clinical notes ($P<.001$), while individuals with psychotic disorders showed a higher tendency to decline ORA ($P<.05$). No significant group differences were observed for other patient groups or characteristics.

Conclusions: The adoption of ORA is influenced by a wide range of motivational factors, while patients also present a similar variety of reasons for declining its use. The results emphasize the urgent need for knowledge and patient education regarding factors that may hinder the decision to use ORA, including its practical usage, its application possibilities, and concerns related to data privacy. Further research is needed to explore approaches for adequately preparing individuals with SMI to transition from their inherent interest to active engagement with ORA.

Trial Registration: German Clinical Trial Register DRKS00030188; https://drks.de/search/en/trial/DRKS00030188

(JMIR Ment Health 2024;11:e51126) doi:10.2196/51126

KEYWORDS

open notes; patient-clinician relations; electronic health record; clinical notes; visit notes; patient participation; online record access; mental illness; patient portal; mental health; qualitative interview; patient education

Introduction

In recent years, several countries have established secure patient portals to enable online record access (ORA), allowing patients to view their health data, including clinical notes ("open notes") from their health care providers [1]. The United States, Canada, and Scandinavian countries, particularly Estonia, Sweden, and Norway, have been at the forefront, providing access to a significant number of patients across multiple regions [2,3]. More recently, the United Kingdom introduced the NHS app, offering access to primary care provider medical records since just last year [4]. In Germany, offering ORA was made mandatory for statutory health insurance providers by 2021, although the inclusion of open notes remains uncertain [5,6].

Research conducted in general health settings indicates clear benefits of patient access to clinical notes, including improved treatment satisfaction, transparency, patient engagement, patient-clinician communication, and health literacy [7-9]. Additionally, ORA enhances medication adherence and security in patients, helps patients to identify and correct treatment errors [10], increases a sense of control over the treatment, and reduces anxieties regarding the treatment process [11]. Health care providers generally view ORA as a valuable tool to promote patient engagement, even though it may be connected to an additional workload related to documentation and communication [12].

Studies conducted in mental health generally yield similar results to those in general health settings, but they also highlight distinct ethical and practical challenges [13,14]. For example, these challenges encompass navigating disagreements between patients and health care professionals (HCPs) regarding clinical notes, as well as discussing exceptions to or limitations of access for patient groups with specific diagnoses or acute conditions, such as severe mental illnesses (SMI) [15,16]. SMI is commonly characterized by conditions including (1) nonorganic psychoses, bipolar disorders, personality disorders, or severe chronic depression; (2) a prolonged psychiatric history involving multiple hospitalizations or outpatient treatments; or (3) moderate impairment in work and leisure activities alongside mild impairment in basic needs [17]. Clinicians often hold reservations about offering ORA to patients affected by SMI. Their concerns primarily revolve around the apprehension that ORA might contribute to self-harming or violent behaviors, especially in patients with whom establishing a trusting relationship is challenging [14]. Patient surveys, however, have demonstrated that individuals with SMI using ORA also exhibit an improved understanding of medication prescriptions, higher medication adherence, and a greater sense of control over their medication [18]. Nevertheless, the adoption of ORA among patients with SMI remains lower compared to those receiving treatment for somatic conditions [18,19].

Apart from assumed positive effects on the patient-clinician relationship and concerns regarding data security, little is known about the motivations and expectations of patients with SMI toward ORA [14,20]. Examining these aspects is crucial for addressing the barriers that hinder the widespread acceptance of ORA among patients with SMI. Therefore, this study aims to thoroughly explore the reasons why patients with SMI consider or reject ORA and whether sociodemographic characteristics may influence patient decisions. More specifically, the following research questions will be addressed (1) which factors influence the decision of patients with SMI, who are offered access to their clinical notes, to either embrace or reject this option? (2) Does the decision for or against ORA in the context of SMI relate to any patient characteristics?

Methods

Design

This study is part of the PEPPPSY project (Piloting and Evaluation of a Participatory Patient-Accessible Electronic Health Record in Psychiatry and Somatics; 2021-2023) that focuses on piloting and evaluating a participatory patient record in psychiatry and somatic medicine [21,22]. It aims to examine the development, implementation, processes, and outcomes of the corresponding patient portal, also known as PEPPPSY, from the perspectives of patients and HCPs. Based on a concurrent mixed methods design, a qualitative methodology was used to comprehensively explore the reasons for and against ORA as well as the potential benefits and challenges [23]. This was complemented by a quantitative analysis to study the possible association between the decision for or against ORA and patient characteristics. Qualitative and quantitative data were analyzed separately. The concurrent design was chosen in order to obtain a comprehensive view of the research question [23], as it simultaneously provides in-depth, qualitative insights into the reasons for acceptance or rejection of ORA and broad, quantitatively measurable data on the patient characteristics in relation to this approval or rejection.

Ethical Considerations

The ethics approval from the Ethics Committee of the Brandenburg Medical School (MHB) was obtained (E-01-20210727), and the study was registered with the German Clinical Trial Register (DRKS00030188).

PEPPPSY App

The patient portal PEPPPSY was developed as part of a research collaboration between the Norwegian University of Science and Technology (NTNU) and the MHB. It emerged from an iterative process of participatory design, development, application, and evaluation [21,22]. PEPPPSY's primary function is to provide patients with secure, 2-factor authenticated access to their physician's notes, but it also includes other information such as laboratory results and a list of prescribed medications. In the current second phase of the pilot, PEPPPSY is being expanded to serve a broader patient population, offering additional features such as a messaging function to allow communication between patients and clinicians concerning the open notes.

Study Setting

The study was conducted at 3 psychiatric outpatient university clinics of psychiatry and psychotherapy at MHB, Immanuel Klinik Rüdersdorf. Located in the metropolitan region Berlin/Brandenburg, the clinics are responsible for providing mental health care services to approximately 255,000 inhabitants in the catchment area. These outpatient clinics offer specialized care for patients who require a comprehensive, multidisciplinary approach due to the nature, severity, or duration of their conditions. The eligibility for receiving treatment in these psychiatric outpatient clinics is based on specific diagnoses, including SMI and other diagnoses, as determined by the insurance providers and the hospital association.

Recruitment and Sampling

From January to June 2023, eligible patients were randomly selected from the 3 study centers by the PEPPPSY research team consecutively. Participants had to meet the following inclusion criteria: age 18 years or older, diagnosed with SMI and confirmed by an external report, and currently receiving treatment in an outpatient psychiatric clinic. Exclusion criteria were previous ORA use related to mental health issues; acute psychiatric conditions or symptoms such as disorientation, severe delusions, hallucinations, katatonia, or agitation that may be associated with significant impairment of cognitive and social functioning; or acute self-harm or harm to others. Eligible patients were informed about the study and written informed consent was obtained. The former included detailed information about ORA, such as how it works, the health information it provides, and offers to participate in (1) this qualitative study and (2) the intervention part of the study, that is, to try out ORA for a period of several months in the study setting described.

Data Collection

The interviews that form the basis of the data in this study were conducted immediately after informed consent was obtained, which included a decision about whether or not participants wished to receive ORA. Sociodemographic data were collected, followed by face-to-face interviews. Information on the diagnoses of the study participants according to the International Classification of Diseases (ICD) was taken from the patient's medical records. The interviews were performed at the aforesaid outpatient clinics by the authors and psychiatrists, ST and JS, who were not the outpatient treatment providers of the participating patients. The interviewers conducted the interviews out of genuine interest in understanding why ORA is accepted or rejected. For the interviews, a semistructured approach based on an interview guide was chosen in order to ensure the comparability of the interviews. This interview guide (see Multimedia Appendix 1) was developed deductively with the participation of all researchers on the basis of a desktop study (or literary research) on the topic of acceptance of ORA among patients with psychiatric disorders using Google Scholar and PubMed [24]. The interview guide was then tested in 2 pilot sessions with patients within the authors' institution. Since no changes to the guide were necessary, the sample interviews could be included in the analysis.

The interviews explored each patient's reasons for acceptance or rejection of ORA usage and had a mean duration of 11.3 (SD 4.5) minutes. In addition to the interviews, the researchers also took field notes, which were later included in the data analysis. Data collection continued until thematic saturation of categories was reached, which occurred when no new themes emerged from the transcripts. The saturation of categories was defined as the point at which no new codes appeared and the meaning of the category and subcategories were established [25].

Data Analysis

The interviews were transcribed verbatim and pseudonymized (JS and ST) and analyzed with thematic analysis (JS, ST, and KN) using the MAXQDA Software (Verbi Software Ltd). Thematic analysis is a flexible approach that was used to inductively ("bottom up") analyze data gathered from semistructured interviews [23]. Data analysis was initially conducted by 2 researchers on each interview individually and verified for consensus; a third person joined in when coders could not reach consensus. The analysis proceeded in six steps (1) familiarizing oneself or becoming familiar with the data, (2) generating initial codes, (3) generating initial themes, (4) reviewing themes, (5) defining and adequately naming themes related to the research questions, and (6) formulating key concepts. After all the themes were generated for each of the interviews, they were divided into 2 groups, reasons for acceptance or rejection of ORA. Subsequently, these themes were clustered within these groups and overarching categories and subcategories were formed. This was a recursive process where different categories were repeatedly tested for coherence and differentiability from the other categories and subcategories. In the final step, 2 researchers jointly selected the most relevant and succinct quotes from the subjects for each of the categories and subcategories. For the group of acceptance of ORA, 4 categories and 13 subcategories were formed. For the group of rejection of ORA, 5 categories and 13 subcategories were formed. For quality assurance purposes, the Consolidated Criteria for Reporting Qualitative research (COREQ) checklist was used (see Multimedia Appendix 2) [26].

The sociodemographic data of participants were analyzed according to their group affiliation (acceptance vs rejection ORA). Descriptive statistics were used to examine possible differences in sociodemographic characteristics between groups using R software (R Core Team) [27], which is available license-free. These between-group differences and their significance were assessed using chi-square test and *t* test [28]. No data were excluded from the data analysis.

Results

Sociodemographic Data

Out of the 124 eligible patients, 83.1% (n=103) agreed to participate in the study about reasons to use or not to use ORA. Sociodemographics are summarized in Table 1. The respondents had an average age of 46.1 (ranging from 19 to 86) years. The majority of the participants were women (n=64, 62%) with an average age of 47.2 years. Respondents who were men had an average age of 44.2 years. Among the approached patients, 58% (n=60) expressed a desire to use ORA, while 42% (n=43) declined ORA. When differentiating for gender, 56% (n=36) of the respondents who were women agreed to participate, while 44% (n=28) declined. Among respondents who were men, 62% (n=24) agreed to participate, while 38% (n=15) declined. The willingness to participate was highest among younger respondents (aged 18 to 39 years) and among patients aged 50 to 59 years. In terms of diagnosis, a high willingness to participate was observed among individuals with affective disorders (ICD 10, F3 [mood (affective) disorders]) at 91% (n=48; *P*<.001). The lowest agreement was found among individuals with schizophrenia, schizotypal, and delusional disorders (ICD 10, F2 [schizophrenia, schizotypal, and delusional disorders]) at 35% (n=6; *P*=.01).

Table 1. Sociodemographic characteristics of the study sample (N=103).

Characteristics	All (N=103)	Do you want online record access?		*P* value
		Yes (n=60)	No (n=43)	
Gender, n (%)				.68
Women	64 (62.1)	36 (56.2)	28 (43.8)	N/A[a]
Men	39 (37.9)	24 (61.5)	15 (38.5)	N/A
Age (years), mean (SD)	46.06 (16.9)	45.1 (16.44)	47.4 (17.66)	.51[b]
Age (years), n (%)				
18-29	20 (19.4)	12 (60.0)	8 (40.0)	>.99
30-39	23 (22.3)	15 (65.2)	8 (34.8)	.34
40-49	15 (14.6)	7 (46.7)	8 (53.3)	.25
50-59	20 (19.4)	13 (65.0)	7 (35.0)	.61
≥60	25 (24.3)	13 (52.0)	12 (48.0)	.35
Diagnosis, n (%)				
All	173 (100.0)	116 (67.1)	57 (32.9)	N/A
F1[c]	20 (11.6)	11 (55.0)	9 (45.0)	.32
F2[d]	17 (9.8)	6 (35.3)	11 (64.7)	.01
F3[e]	53 (30.6)	48 (90.6)	5 (9.4)	<.001
F4[f]	36 (20.8)	21 (58.3)	15 (41.7)	.23
F6[g]	15 (8.7)	8 (53.3)	7 (46.7)	.25
Others[h]	32 (18.5)	22 (68.8)	10 (31.2)	.68
Number of diagnosis, mean (SD)	1.88 (0.87)	1.95 (0.81)	1.79 (0.94)	.37[b]

[a]N/A: not applicable.

[b]*P* values were calculated using *t* test, while all other values were calculated based on chi-square test.

[c]F1: Mental and behavioral disorders due to psychoactive substance use.

[d]F2: Schizophrenia, schizotypal, and delusional disorders.

[e]F3: Mood (affective) disorders.

[f]F4: Neurotic, stress-related, and somatoform disorders.

[g]F6: Disorders of adult personality and behavior.

[h]Others: All mental disorders in the International Classification of Diseases-F chapter beyond those previously listed were subsumed under this category.

Reasons for Acceptance and Rejection of Participation in ORA

The categories and subcategories for the respective reasons provided by the respondents are summarized in Table 2. Subcategories and quotes in the following text are represented in italics.

Table 2. Categories and subcategories of stated reasons for acceptance and declination of ORA (N=103).

Categories and subcategories	Values[a], n (%)
Reasons for acceptance of ORA[b]	
Wish to engage in treatment	
Improved self-understanding and self-knowledge	14 (13.6)
Interest in the external perspective provided by clinicians	12 (11.7)
Continual contact and exchange	10 (9.7)
Incentive for increased engagement in treatment	6 (5.8)
Understanding the treatment process	
Reminder of content discussed in therapy sessions	27 (26.2)
Ability to track the progress of treatment over time	5 (4.9)
Interest in medical translation of own symptoms	3 (2.9)
Critically assessing clinicians	
Gaining more transparency into the perspective of health care providers on patients	6 (5.8)
Needing to verify the correctness of the notes in order to be able to trust the clinician	10 (9.7)
Avoiding and correcting misunderstandings	13 (12.6)
Sharing personal health data with third parties	
Improving communication about the illness with significant others	7 (6.8)
Ability to share own health data with public institutions	3 (2.9)
Having access to own health data	2 (1.9)
Reasons for rejection of ORA	
Feeling well supported in face-to-face interactions	
Sufficient oral “transmission” of notes	15 (14.6)
Desire to address problems and inquiries more effectively in direct conversation	10 (9.7)
Adequate satisfaction with in-person appointments	7 (6.8)
Self-uncertainty	
Feeling emotionally burdened by reading the notes	16 (15.5)
Fear of excessive confrontation with one’s own condition	3 (2.9)
Adequate satisfaction with one’s own perspective	7 (6.8)
Uncertainty in the relationship with the clinician	
Control weakens trust	2 (1.9)
Trust does not require control	19 (18.4)
Concern about technical demands for clinicians	1 (1.0)
Concerns about the misuse of health data by third parties	
Worries about data security	5 (4.9)
Concerns of unwanted control by family members	2 (1.9)
Practical barriers	
Difficulties in dealing with technology	13 (12.6)
Difficulties in reading and understanding the notes	7 (6.8)

[a]The number of patients (n, %) who mentioned each theme is indicated in parentheses.

[b]ORA: online record access.

Reasons for Acceptance

Overview

The stated reasons for accepting participation in ORA can be grouped into 4 main categories with a total of 13 subcategories.

Wish to Engage in Treatment

The respondents associated their agreement to participate with the wish and motivation to become more actively engaged in their treatment. They hoped that by using the portal, they could gain a better understanding of themselves and their condition, often driven by their interest in the external perspectives of their clinicians.

> *I am interested in knowing what they actually [think] about me here, because it's about me, my health. Maybe I can understand everything [about why I'm feeling unwell] better.* [Patient 14]

In the responses, this interest was often connected to a wish for ongoing communication and interaction.

> *I find it quite practical because it helps me stay in touch with my doctor and keep track of documentation. This way, I can tell the doctor when I'm not feeling well.* [Patient 55]

Overall, the participants viewed their participation as an incentive to become more engaged in their treatment.

Overview of the Treatment Process

Gaining a comprehensive understanding of the treatment process is another category that emerged from the participants' responses. While closely related to their willingness to engage in treatment, it primarily focused on the desire to have an overview of the treatment process. Many participants appreciated the ability to read their open notes through ORA, as it served as a reminder of therapy sessions and allowed them to track the chronological progression of their treatment.

> *I would like to have an overview of how my condition has changed over time, whether things have improved or worsened. Otherwise, you just live in the moment and with the things I tell you in this moment. But having it documented from appointment to appointment, and knowing that things might have gotten better without me realizing it, I would like to have that on paper.* [Patient 89]

Additionally, embracing ORA was motivated by an interest in understanding how their own symptoms are translated in a medical context. This includes the use of specialized terminology to describe their symptoms and the subsequent treatment recommendations.

> *I would like to know how you medically process what I tell you during the treatment and what implications it has for the diagnosis. I'm just sharing things from my life, but what does it mean for the illness and what needs to be done now? I would like to see that.* [Patient 84]

Critically Assessing Clinicians

Participants expressed their acceptance of ORA not only as a means to engage more actively in their treatment but also as an opportunity to critically assess the perspectives and approaches of clinicians. Participants valued the chance to gain more transparency about how clinicians view their patients by reading clinical notes and being able to provide feedback via comments.

> *To find out what therapists think about me behind my back and whether they even notice you in the hospital setting.* [Patient 72]

Participants expressed concerns that clinicians might not accurately understand or document patients' individual needs, leading to doubts about whether they can be trusted. This led to a need to verify the correctness of the notes in order to be able to trust their clinician.

> *I have so much mistrust towards doctors, especially regarding my psychosis and the forced medication, and how things I've said and done have been twisted. I can see it happening with my grandma too, how she's being treated. That's why I just want to see what you actually write down.* [Patient 91]

From a more constructive perspective, many participants saw the possibility of viewing open notes within ORA as a way of preventing and rectifying misunderstandings that may arise during conversations.

Sharing Personal Health Data With Third Parties

This category describes aspects that are less focused on the treatment itself and its documentation, but rather on the use of this documentation with third parties. Participants expressed the hope that sharing the clinical documentation with significant others (eg, family members and friends) and other health care providers would improve the exchange of information about their illness.

> *For instance, my wife would like to talk to someone about how to handle my condition. That was my first thought, that she could also read what you write. I can't remember and convey everything. This way, she could participate without me burdening her with it.* [Patient 27]

Furthermore, participants viewed the ability to share their own health data with public institutions such as health insurance companies or the police through ORA as a positive aspect.

> *I recently had an issue with the health insurance company where they just declared me as healthy. They requested my medical records, but nobody was available at the psychiatric outpatient clinic. I could simply print out my documentation. That would be just great.* [Patient 36]

Furthermore, the basic opportunity to have access to one's own health data was also mentioned as a reason for accepting ORA.

Reasons for Rejection

Overview

The reasons for declining participation in ORA can be grouped into 5 categories with 13 subcategories.

Feeling Well-Supported in Face-To-Face Interactions

This category includes topics in which participants decline ORA because they already feel adequately cared for through the current mode of contact. For instance, they perceived their in-office appointments as sufficient for their needs.

> *I am feeling satisfied with conversations. (...) I'm not someone who spends a lot of time on their phone. I prefer being outdoors. (...) If I don't understand something, I can simply ask for clarification. Looking at the notes would only add more information to my already busy mind.* [Patient 32]

This feeling of being adequately cared for through in-office appointments was repeatedly associated with the desire of the respondents to address problems and inquiries through direct conversation. Moreover, they expressed a preference for discussing their own notes through verbal communication.

> *We have already discussed it [the topic of today's session]. If there's anything or if I want to know more, I can always ask.* [Patient 12]

Self-Uncertainty

Another central theme in the patients' statements was the concern or fear of becoming unsettled by reading the notes. While several respondents mentioned being sufficiently satisfied with their own perspective on themselves, there was often an underlying fear of being burdened by reading the notes.

> *I don't need to read that; I'm already experiencing all this crap myself. I don't need to see it in black and white too.* [Patient 78]

In this context, some respondents specifically expressed fear of too much confrontation with their own condition, and some of the participants wished to leave that within the scope of the therapeutic conversation and not reactivate it through reading clinical notes.

> *I wouldn't be up for that. Because, well, I unload all this stuff on you here that makes me sick, and afterwards, I actually feel better. But if I were to read through all that I've told you again, it would really bring me down all over again.* [Patient 67]

Uncertainty in the Relationship to the Clinician

The respondents expressed concerns about not only their own self-uncertainty but also about feeling uncertain toward the health care provider and the therapeutic relationship when it comes to using ORA. Specifically, they highlighted that allowing patients to review their notes could potentially undermine the trust and rapport they have established with their clinicians.

> *I am a doctor myself and I know that it harms the doctor-patient relationship when patients read what doctors write about them. It is very important to me that I trust you without constantly reviewing what you document.* [Patient 92]

Contrary to the wish to critically assess the clinician as a justification for the use of ORA (as mentioned above), the respondents emphasized that satisfaction with treatment and a trusting relationship do not require such control.

> *I trust you that everything is accurate, right? How you write it down. I'm really satisfied with the treatment, I've even had my pension extended, and all the services I need are being provided, so everything you document and how you communicate it to others must be correct, right? Others might want to know sometimes, but I also feel that what I tell you is being understood, so I don't need to read anything extra.* [Patient 73]

Here, concerns were raised about the potential increase in workload for clinicians, which could potentially strain the therapeutic relationship due to the perceived additional workload.

Concerns About the Misuse of Health Data by Third Parties

Some respondents explained their rejection based on concerns about data security, specifically regarding the potentially insecure storage of documentation for instance on mobile devices, which could result in unauthorized access by third parties. Unlike the proponents of ORA, those who expressed opposition to it also raised concerns about unwanted control by care partners through unauthorized access to the patient portal.

> *I have a curious girlfriend who doesn't necessarily need to read along. (...) It's not for my family members. I would feel too controlled by my partner. She already opens my mail and goes through my bank statements.* [Patient 52]

Practical Barriers

Finally, some respondents mentioned technical and practical challenges as reasons for their rejection. Specifically, participants over the age of 49 years highlighted the difficulty of dealing with the technology required for ORA, such as smartphones, browser apps, and 2-factor authentication. Additionally, some expressed feeling overwhelmed by the comprehension of the notes, as they encountered challenges due to the use of medical terminology and their own difficulties in reading caused by issues with concentration.

> *I have really bad concentration problems, so that I can't understand anything anymore and can't fully engage in anything, [I] would only understand half of it, especially when reading. Additionally, I don't have internet access, and I don't understand how to set it up.* [Patient 97]

Discussion

Principal Findings

In summary, the reasons provided by the interviewed patients with SMI for their decision to use or not use ORA are diverse.

Among those in favor, motivations range from a desire for increased engagement in treatment to critical evaluation of clinicians and using accessible health data for sharing with third parties. In contrast, those who opposed ORA perceived their therapeutic relationship as already well-established and feared that it might be jeopardized by the use of ORA. Finally, practical barriers, mostly related to digital literacy, were cited as reasons for their opposition.

Acceptance of ORA

The reasons for approval are briefly discussed as they largely align with those found in existing studies, thus providing limited implications for the further development of ORA. Those reasons include the motivation to engage more actively in treatment: by using the portal, patients hope to better understand the content of medical appointments and to obtain a clearer view of themselves, their illness, and the external perspective of their doctors [29]. Although not explicitly thematized in this study, it is reasonable to infer that this motivation also leads to increased adherence to medical treatment. This assumption is supported by another study, which found that patients with SMI when using ORA, reported an improved understanding of their medication prescriptions and described feeling more comfortable and in control throughout the therapeutic process [18,30]. However, this question requires further investigation in a follow-up study.

Moreover, many patients see ORA as a way to obtain a comprehensive overview of their treatment process. This includes accessing open notes as a reminder of therapy sessions, tracking treatment progress, and understanding how their symptoms are documented in a medical context.

Critical evaluation of clinicians is another reason for the acceptance of ORA by patients which is also reported by other studies [31]. The participants in our study reported that by reading the clinical notes, they want to evaluate the transparency and accuracy of clinicians' perceptions and documentation of their needs. This critical view is also seen as an opportunity to address and correct potential misunderstandings that may occur during the consultation. However, this need for critical monitoring of the practitioners was ultimately linked to the desire to deepen trust in the practitioner and the treatment process. This need or desire to enhance trust is also highlighted in a study by Cromer et al [32].

In addition to patients appreciating gaining access to their health data through ORA, the ability to share personal health information with third parties, such as family, friends, other health care providers and public bodies, is also viewed positively. Again, this finding is consistent with preexisting literature on the perceived benefits of ORA by health care users [33].

Rejection of ORA

When considering the rejections of ORA among patients with SMI, differences compared to patients from general health settings become apparent. For example, 1 patient acknowledges having significant comprehension difficulties during direct patient-clinician interactions due to concentration problems and suspects that reading clinical notes would exacerbate the issue. This aligns with existing studies that suggest severe symptoms, which persist in daily life and tend to hinder participation in digital health interventions [34]. Furthermore, patients with SMI more often experience intersecting factors such as low educational attainment and language barriers [35], which was also repeatedly stated by study participants with regard to difficulties in dealing with technology and understanding the notes. Low educational attainment and language barriers can subject them to greater stress in patient-clinician interactions when trying to understand health care providers' explanations [35]. Collectively, these individual limitations can lead to concerns and experiences not being adequately articulated, misunderstood, or possibly forgotten within the limited time available at an appointment. Consequently, this may result in less interest in accessing the notes made by the clinician. On the other hand, these individual limitations could also serve as an argument in favor of ORA: ORA offers the opportunity to enhance understanding of the patient-clinician relationship [36]. It can contribute to mitigating the mentioned disadvantages of inequities by extending the therapeutic interaction beyond physical encounters into the digital space where patients may feel less pressured to conform with the HCP and may also express themselves more easily than in the physical space. However, it is crucial that the clinical notes are written in a language that is relatable to everyday life and nonjudgmental [8,37]. This allows patients to reread and better understand the content discussed during previous appointments in preparation for an upcoming one. Furthermore, a messaging or commenting feature enables patients to ask questions or gather any unresolved concerns. Nonetheless, concern about being emotionally burdened by reading the notes was a common reason for deciding against ORA. Remarkably, these fears correspond with those expectations of HCPs, even though they were rarely confirmed after adopting ORA [3,7,8,12].

These arguments raise the question of whether patients with SMI, who experience daily limitations due to their symptoms, should be informed about the potential benefits of ORA in a more specific or repeated manner, and whether such adapted and improved information could potentially modify the approval rate. On the other hand, it might be that just the opposite is the case and that providers are particularly reluctant to share notes with this population and do in fact not routinely discuss open notes or encourage their clients to read them [38,39]. Unfortunately, this issue did not emerge from our data and further research is needed to clarify this question.

Then there is a group of patients who reject ORA because they already feel well taken care of in face-to-face interactions for various reasons. Studies examining the willingness to adopt digital health services explain the preference for direct patient contact, among other factors, through personality traits [34]. Extraversion, in particular, is considered a predictor of a lower likelihood of engaging with digital health services [40]. Individuals who displayed higher extraversion preferred meeting and connecting with the doctor in person. Some of the statements made by the participants convey a certain persistence in favor of nondigital means of communication (see subcategory "Sufficient oral 'transmission' of notes" in Table 2). In line with this, other findings indicate that personality traits associated with resistance to change and openness to new experiences result in a lower adoption of digital health services [41]. Therefore, it would be interesting for further research to explore whether these corresponding personality traits align with the thematic trends found in our study. Beyond these considerations, the attitude of rejecting ORA seems to be explained in particular by the fact of enduring SMI. On the one hand, the hope of indirectly positively influencing one's own mental health through ORA may be reduced due to the length or severity of the course of illness [34]. On the other hand, in the examined health care system for individuals with SMI, assuming they are in a phase of predominant psychological stability, treatment contacts are rare (approximately 1-2 sessions of 15 minutes each in 3 months). As a result, the opportunities for exchange and the scope of exchangeable content through ORA are limited from the perspective of the surveyed patients [42]. However, it is worth noting that precisely because appointments are short and there are potentially many topics to be discussed (current status of well-being, medications, laboratory results, medication levels, etc), ORA could provide patients with SMI with more space to exchange a wide range of information at a later time and overall enhance the therapeutic contact beyond the physical encounter.

Other participants expressed their rejection of ORA by explaining that the burden of their mental illness in their everyday lives was already substantial, leading them to decline any additional confrontation beyond their appointments at the outpatient clinic. This aspect seems to correspond with the preceding factor, suggesting that these patients are currently unable to dedicate any further (mental) capacity to engage with their chronic mental condition beyond medical appointments.

Comparison of Acceptance and Rejection of ORA

When comparing the reasons provided by patients for or against the use of ORA, several interesting contrasts become apparent. While some reasons for approval can be interpreted as a desire to deepen the therapeutic relationship, the opposition, in certain cases, stems from the apprehension that this therapeutic alliance may be undermined and jeopardized through the introduction of control (refer to subcategory "Trust does not require control" in Table 2). Conversely, patients who embraced ORA described a high need for control, which motivates their use of ORA. Accordingly, the use of ORA is perceived as an opportunity to critically review the HCP's perspective and documentation, rather than blindly trusting them (see subcategory "Needing to verify the correctness of the notes in order to be able to trust the clinician" in Table 2). The disclosure of notes, in the optimal scenario, can thus be regarded as a demonstration of trust that allows for a deepening of the therapeutic relationship [42].

Another theme that underlies both the approval and rejection of ORA is the use of health data by third parties. This issue raises concerns about data security and the potential for unwanted control by family members when sharing information with significant others, health care providers, and public institutions. It is important to note that privacy and trustworthiness are among the most common reservations regarding ORA and digital (mental) health interventions in general, given the sensitive and potentially stigmatizing nature of the content involved [16,34,43]. A recent study conducted in Sweden provides evidence that these reservations are valid, as patients with mental illness experience significantly more attempts by unauthorized individuals to access their mental health records compared to patients in general health settings [44]. In our study, 1 patient expressed the misconception that health data are directly stored on their mobile phone. This misunderstanding highlights a knowledge gap where patients may not be aware that the data are actually securely stored remotely using 2-factor authentication, thereby aiming to prevent unauthorized access through the phone. However, the concerns expressed in our study once again emphasize the importance of data protection in the implementation of digital health platforms and the need for sufficient patient and provider education on this matter.

Differences Between Patient Groups

Generally, the proportion of patients willing to use ORA is approximately 60%, which is consistent with findings from previous studies [19,43,45]. However, the actual usage rate of ORA among patients is expected to be even lower [18]. The 2 groups, those in favor and those against ORA show little difference in terms of age, gender, distribution of diagnoses, and comorbidity, except for psychotic and depressive disorders. The higher levels of agreement and motivation among patients with depression align with findings from other studies [43,45], possibly due to a higher prevalence of socially desirable behavior in this patient group. The low approval rate among patients with schizophrenia is somewhat surprising compared to the existing literature. According to previous studies, patients with psychosis are generally very well able to use web-based interventions, exhibit positive attitudes toward them, and use the web-based more frequently to build their social networks compared to the general population [46-48]. The rejection of ORA in our study population could be attributed to reduced digital literacy, functional impairments caused by psychotic symptoms, as well as an approach to the illness characterized by internalized stigma and social withdrawal [49]. This social withdrawal has also been described as a protective mechanism against overly social and open interactions [50].

Strengths and Limitations

This is the first study that examines the reasons for the acceptance and rejection of ORA among patients with SMI in the German health care system. The investigation of these factors is crucial for advancing the implementation of ORA in the German-speaking region and can only be meaningful through a comparison with international research findings. Moreover, the study contributes to filling the research gap regarding the perspectives of individuals with SMI toward ORA.

One limitation of the study is that while a variety of reasons for rejecting ORA became apparent in the qualitative survey, raising further questions about factors such as digital literacy or the respondents' social behavior, these factors were not explored in the quantitative survey. Similarly, comprehensive sociodemographic information such as educational level, socioeconomic status, or duration of mental illness was unfortunately not available in the data corpus. A follow-up study may be useful to further validate the qualitative data and to analyze in-depth the role of other patient characteristics that contribute to the decision for or against ORA.

Another limitation of the study is that it did not present the proportion of patients who actually used ORA after having consented to do so. Since this study is based on baseline data from the PEPPPSY study [21], the analysis of usage patterns and effects of ORA is yet to be conducted. Another issue at first glance is the dichotomization of the results (see Table 2). The question arises as to how any positive attitudes of patients who reject ORA (and vice versa) were taken into account. For instance, patients who opt for ORA may still hold concerns regarding data privacy. However, this limitation was addressed by incorporating any sub-aspects of patient statements that are in opposition to their decision for or against ORA in the qualitative analysis of the data. This means that all patient statements and attitudes toward ORA were accounted for in our qualitative analysis independently from the patients' decisions for or against ORA.

Future Research

In addition to the research gaps identified above, further research is needed to address the unique needs of individuals with SMI in order to effectively facilitate maintained engagement with ORA. First, the extent to which patient characteristics and, in particular, psychiatric functional impairments, as well as concepts such as internalized stigma and social withdrawal, influence acceptance of ORA should be investigated further. Possible influences of personality traits such as extraversion or resistance to change on willingness to use ORA should also be considered.

Second, it should be investigated to a larger extent, whether the fear of possible adverse effects from reading the findings and clinical notes made available via ORA is confirmed in practice. Although studies to date tend to suggest otherwise, patients' concerns should be taken seriously. In this context, research should be conducted on how to formulate clinical notes in a way that is both understandable and empathetic to patients without overburdening the available resources of practitioners. In this respect, there are preliminary indications of promising use of generative language models [51].

Third, there is no evidence on what cues, explanations, or motivations patients with SMI need from the medical team, and especially from their clinicians, to want to use ORA more. In this context, it should be investigated whether improved patient information about the benefits of ORA increases adoption rates. At the same time, there seems to be a need to explore what skills HCPs need to acquire in order to formulate clinical notes in a way that is understood by patients and adds value, which involves adapting their communication style to align with patients' familiar vocabulary rather than relying solely on technical medical terminology. Finally, actual rates of ORA use among patients with SMI compared to adoption rates and reasons for potential discrepancies should be explored.

Conclusions

In general, patients with affective disorders (ICD 10, F3) showed high interest in ORA, whereas patients with schizophrenia, schizotypal and delusional disorders (ICD 10, F2) were less interested. It was mainly female patients of younger (18-39 years) and middle (50-59 years) age who agreed to receive ORA. Acceptance of ORA by patients with SMI stems primarily from a desire to be more actively involved in their care, to have a comprehensive view of their treatment process, and to evaluate the accuracy of physicians' perception and documentation of their needs. This critical perspective is also seen as an opportunity to address and correct any misunderstandings that may have occurred during the consultation. The value placed on access to personal health information, combined with the ability to share that information with third parties, underscores the patients' positive attitudes toward ORA.

Rejection of ORA by patients with SMI is primarily motivated by a sense of already being well supported by face-to-face interactions, as well as concerns rooted in their own insecurities. These range from fear of being unsettled by reading clinical notes to avoidance of excessive confrontation with one's condition outside of the therapeutic conversations. Patients worry that the transparency created by ORA could undermine trust in their health care providers, especially given the additional workload for clinicians. Finally, data security risks and practical barriers such as lack of digital literacy and incomprehensible medical jargon contributed to the decision not to use ORA.

Acknowledgments

This study was funded by the Brandenburg Medical School publication fund supported by the German Research Foundation and the Ministry of Science, Research and Cultural Affairs of the State of Brandenburg. ST received internal clinical scientist funding by the Brandenburg Medical School.

Authors' Contributions

JS and ST contributed to the study design and collected the data. KN, ST, and JS conducted data analysis. JS and ST wrote a first draft. EMD translated the study into the English language. Successive drafts were revised by EMD and YE. All authors critically reviewed and commented on the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guidelines.
[DOCX File , 7 KB - mental_v11i1e51126_app1.docx]

Multimedia Appendix 2
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[PDF File (Adobe PDF File), 490 KB - mental_v11i1e51126_app2.pdf]

References

1. Delbanco T, Walker J, Darer JD, Elmore JG, Feldman HJ, Leveille SG, et al. Open notes: doctors and patients signing on. Ann Intern Med 2010;153(2):121-125 [FREE Full text] [doi: 10.7326/0003-4819-153-2-201007200-00008] [Medline: 20643992]
2. Essén A, Scandurra I, Gerrits R, Humphrey G, Johansen MA, Kierkegaard P, et al. Patient access to electronic health records: differences across ten countries. Health Policy and Technology 2018 Mar;7(1):44-56. [doi: 10.1016/j.hlpt.2017.11.003]
3. Zanaboni P, Kummervold PE, Sørensen T, Johansen MA. Patient use and experience with online access to electronic health records in Norway: results from an online survey. J Med Internet Res 2020 Feb 07;22(2):e16144 [FREE Full text] [doi: 10.2196/16144] [Medline: 32031538]
4. McMillan B, Davidge G, Blease C, Watson J. Implications of the changes to patient online records access in English primary care. Br J Gen Pract 2022 Dec;72(725):558-559 [FREE Full text] [doi: 10.3399/bjgp22X721205] [Medline: 36424157]
5. Schwarz J, Esch T. OpenNotes in mental health? Pro. Nervenarzt 2022 Nov;93(11):1163-1164. [doi: 10.1007/s00115-022-01357-4] [Medline: 35913587]
6. Lauterbach: Elektronische Patientenakte ab Ende 2024 für alle verbindlich. German Ministry of Health. URL: https://www.bundesgesundheitsministerium.de/presse/interviews/interview/fas-030324-elektronische-patientenakte.html [accessed 2023-07-16]
7. Delbanco T, Walker J, Bell SK, Darer JD, Elmore JG, Farag N, et al. Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead. Ann Intern Med 2012 Oct 02;157(7):461-470 [FREE Full text] [doi: 10.7326/0003-4819-157-7-201210020-00002] [Medline: 23027317]
8. Walker J, Leveille S, Bell S, Chimowitz H, Dong Z, Elmore JG, et al. OpenNotes after 7 years: patient experiences with ongoing access to their clinicians' outpatient visit notes. J Med Internet Res 2019 May 06;21(5):e13876 [FREE Full text] [doi: 10.2196/13876] [Medline: 31066717]
9. Wolff JL, Darer JD, Berger A, Clarke D, Green JA, Stametz RA, et al. Inviting patients and care partners to read doctors' notes: OpenNotes and shared access to electronic medical records. J Am Med Inform Assoc 2017 Apr 01;24(e1):e166-e172 [FREE Full text] [doi: 10.1093/jamia/ocw108] [Medline: 27497795]
10. Bell SK, Mejilla R, Anselmo M, Darer JD, Elmore JG, Leveille S, et al. When doctors share visit notes with patients: a study of patient and doctor perceptions of documentation errors, safety opportunities and the patient-doctor relationship. BMJ Qual Saf 2017;26(4):262-270 [FREE Full text] [doi: 10.1136/bmjqs-2015-004697] [Medline: 27193032]
11. Davis Giardina T, Menon S, Parrish DE, Sittig DF, Singh H. Patient access to medical records and healthcare outcomes: a systematic review. J Am Med Inform Assoc 2014;21(4):737-741 [FREE Full text] [doi: 10.1136/amiajnl-2013-002239] [Medline: 24154835]
12. DesRoches CM, Leveille S, Bell SK, Dong ZJ, Elmore JG, Fernandez L, et al. The views and experiences of clinicians sharing medical record notes with patients. JAMA Netw Open 2020 Mar 02;3(3):e201753 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.1753] [Medline: 32219406]
13. Durocher K, Shin HD, Lo B, Chen S, Ma C, Strudwick G. Understanding the role of patient portals in fostering interprofessional collaboration within mental health care settings: mixed methods study. JMIR Hum Factors 2023 Jul 19;10:e44747 [FREE Full text] [doi: 10.2196/44747] [Medline: 37467024]
14. Schwarz J, Bärkås A, Blease C, Collins L, Hägglund M, Markham S, et al. Sharing clinical notes and electronic health records with people affected by mental health conditions: scoping review. JMIR Ment Health 2021 Dec 14;8(12):e34170 [FREE Full text] [doi: 10.2196/34170] [Medline: 34904956]
15. Blease C, Torous J, Kharko A, DesRoches CM, Harcourt K, O'Neill S, et al. Preparing patients and clinicians for open notes in mental health: qualitative inquiry of international experts. JMIR Ment Health 2021 Apr 16;8(4):e27397 [FREE Full text] [doi: 10.2196/27397] [Medline: 33861202]
16. Hägglund M, Scandurra I. Usability of the Swedish accessible electronic health record: qualitative survey study. JMIR Hum Factors 2022 Jun 23;9(2):e37192 [FREE Full text] [doi: 10.2196/37192] [Medline: 35737444]
17. Parabiaghi A, Bonetto C, Ruggeri M, Lasalvia A, Leese M. Severe and persistent mental illness: a useful definition for prioritizing community-based mental health service interventions. Soc Psychiatry Psychiatr Epidemiol 2006 Jun;41(6):457-463. [doi: 10.1007/s00127-006-0048-0] [Medline: 16565917]

18. Blease C, Dong Z, Torous J, Walker J, Hägglund M, DesRoches CM. Association of patients reading clinical notes with perception of medication adherence among persons with serious mental illness. JAMA Netw Open 2021 Mar 01;4(3):e212823 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.2823] [Medline: 33760088]
19. Onyeaka H, Ajayi KV, Muoghalu C, Eseaton PO, Azuike CO, Anugwom G, et al. Access to online patient portals among individuals with depression and anxiety. Psychiatry Res Commun 2022;2(4):100073 [FREE Full text] [doi: 10.1016/j.psycom.2022.100073]
20. Leung K, Clark C, Sakal M, Friesen M, Strudwick G. Patient and family member readiness, needs, and perceptions of a mental health patient portal: a mixed methods study. Stud Health Technol Inform 2019;257:266-270. [Medline: 30741207]
21. Hochwarter S, Fehler G, Muente C, Eisenmann Y, Heinze M, Hibsch C, et al. Design of a patient-accessible electronic health record system in mental health. In: Séroussi B, Weber P, Dhombres F, Grouin C, Liebe JD, Pelayo S, et al, editors. Challenges of Trustable AI and Added-Value on Health: Proceedings of MIE 2022. Amsterdam, Netherlands: IOS Press; 2022:583-584.
22. Gibson M, Halvorsen M, Joner MAA, Lehne M, Stautland ER, Svensson G, et al. Pilot testing and evaluation of participatory patient record in psychiatric care. 2022 Presented at: The 18th Scandinavian Conference on Health Informatics; August 22-24, 2022; Tromsø, Norway p. 216-217 URL: https://ecp.ep.liu.se/index.php/shi/article/view/467 [doi: 10.3384/ecp187039]
23. Leech NL, Onwuegbuzie AJ. A typology of mixed methods research designs. Qual Quant 2007;43(2):265-275. [doi: 10.1007/s11135-007-9105-3]
24. Kallio H, Pietilä AM, Johnson M, Kangasniemi M. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. J Adv Nurs 2016 Dec;72(12):2954-2965. [doi: 10.1111/jan.13031] [Medline: 27221824]
25. Clarke V, Braun V, Hayfield N. Thematic analysis. In: Smith JA, editor. Qualitative Psychology: A Practical Guide to Research Methods. London: SAGE Publications; 2015:222-248.
26. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
27. Chambers JM, Chambers J. Software for Data Analysis: Programming with R. New York: Springer; 2008.
28. Coolican H. Research Methods and Statistics in Psychology, Sixth Edition. London: Psychology Press; 2017.
29. Turvey CL, Fuhrmeister LA, Klein DM, Moeckli J, Howren MB, Chasco EE. Patient and provider experience of electronic patient portals and secure messaging in mental health treatment. Telemed J E Health 2022;28(2):189-198 [FREE Full text] [doi: 10.1089/tmj.2020.0395] [Medline: 33887164]
30. Klein JW, Peacock S, Tsui JI, O'Neill SF, DesRoches CM, Elmore JG. Perceptions of primary care notes by patients with mental health diagnoses. Ann Fam Med 2018;16(4):343-345 [FREE Full text] [doi: 10.1370/afm.2287] [Medline: 29987083]
31. O'Neill S, Chimowitz H, Leveille S, Walker J. Embracing the new age of transparency: mental health patients reading their psychotherapy notes online. J Ment Health 2019;28(5):527-535. [doi: 10.1080/09638237.2019.1644490] [Medline: 31364902]
32. Cromer R, Denneson LM, Pisciotta M, Williams H, Woods S, Dobscha SK. Trust in mental health clinicians among patients who access clinical notes online. Psychiatr Serv 2017;68(5):520-523 [FREE Full text] [doi: 10.1176/appi.ps.201600168] [Medline: 28142383]
33. Latulipe C, Quandt SA, Melius KA, Bertoni A, Miller DP, Smith D, et al. Insights into older adult patient concerns around the caregiver proxy portal use: qualitative interview study. J Med Internet Res 2018;20(11):e10524 [FREE Full text] [doi: 10.2196/10524] [Medline: 30389654]
34. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
35. Grossman LV, Masterson Creber RM, Benda NC, Wright D, Vawdrey DK, Ancker JS. Interventions to increase patient portal use in vulnerable populations: a systematic review. J Am Med Inform Assoc 2019 Aug 01;26(8-9):855-870 [FREE Full text] [doi: 10.1093/jamia/ocz023] [Medline: 30958532]
36. Silva JM, Durden TE, Hirsch A. Erasing inequality: examining discrepancies between electronic health records and patient narratives to uncover perceived stigma and dismissal in clinical encounters. Soc Sci Med 2023;323:115837. [doi: 10.1016/j.socscimed.2023.115837] [Medline: 36931038]
37. Gerard M, Chimowitz H, Fossa A, Bourgeois F, Fernandez L, Bell SK. The importance of visit notes on patient portals for engaging less educated or nonwhite patients: survey study. J Med Internet Res 2018 May 24;20(5):e191 [FREE Full text] [doi: 10.2196/jmir.9196] [Medline: 29793900]
38. Dobscha SK, Denneson LM, Jacobson LE, Williams HB, Cromer R, Woods S. VA mental health clinician experiences and attitudes toward OpenNotes. Gen Hosp Psychiatry 2016;38:89-93. [doi: 10.1016/j.genhosppsych.2015.08.001] [Medline: 26380876]
39. Dobscha SK, Kenyon EA, Pisciotta MK, Niederhausen M, Woods S, Denneson LM. Impacts of a web-based course on mental health clinicians' attitudes and communication behaviors related to use of OpenNotes. Psychiatr Serv 2019 Jun 01;70(6):474-479. [doi: 10.1176/appi.ps.201800416] [Medline: 30890047]

40. March S, Day J, Ritchie G, Rowe A, Gough J, Hall T, et al. Attitudes toward e-mental health services in a community sample of adults: online survey. J Med Internet Res 2018 Feb 19;20(2):e59 [FREE Full text] [doi: 10.2196/jmir.9109] [Medline: 29459357]
41. Mikolasek M, Witt CM, Barth J. Adherence to a mindfulness and relaxation self-care app for cancer patients: mixed-methods feasibility study. JMIR Mhealth Uhealth 2018;6(12):e11271 [FREE Full text] [doi: 10.2196/11271] [Medline: 30522990]
42. Shin HD, Durocher K, Lo B, Chen S, Ma C, Wiljer D, et al. Impact of a mental health patient portal on patients' views of compassion: a mixed-methods study. BMC Digit Health 2023 Jan 24;1(1):1 [FREE Full text] [doi: 10.1186/s44247-022-00002-z]
43. Matthews EB, Savoy M, Paranjape A, Washington D, Hackney T, Galis D, et al. Acceptability of health information exchange and patient portal use in depression care among underrepresented patients. J Gen Intern Med 2022;37(15):3947-3955 [FREE Full text] [doi: 10.1007/s11606-022-07427-2] [Medline: 35132548]
44. Bärkås A, Kharko A, Åhlfeldt RM, Hägglund M. Patients' experiences of unwanted access to their online health records. Stud Health Technol Inform 2023;302:356-357. [doi: 10.3233/SHTI230138] [Medline: 37203682]
45. Etingen B, Hogan TP, Martinez RN, Shimada S, Stroupe K, Nazi K, et al. How do patients with mental health diagnoses use online patient portals? An observational analysis from the veterans health administration. Adm Policy Ment Health 2019;46(5):596-608. [doi: 10.1007/s10488-019-00938-x] [Medline: 31065908]
46. Thomas N, Foley F, Lindblom K, Lee S. Are people with severe mental illness ready for online interventions? Access and use of the internet in Australian mental health service users. Australas Psychiatry 2017 Jun;25(3):257-261. [doi: 10.1177/1039856217689913] [Medline: 28139947]
47. Gay K, Torous J, Joseph A, Pandya A, Duckworth K. Digital technology use among individuals with schizophrenia: results of an online survey. JMIR Ment Health 2016;3(2):e15 [FREE Full text] [doi: 10.2196/mental.5379] [Medline: 27146094]
48. Highton-Williamson E, Priebe S, Giacco D. Online social networking in people with psychosis: a systematic review. Int J Soc Psychiatry 2015;61(1):92-101. [doi: 10.1177/0020764014556392] [Medline: 25381145]
49. Arnold C, Farhall J, Villagonzalo KA, Sharma K, Thomas N. Engagement with online psychosocial interventions for psychosis: a review and synthesis of relevant factors. Internet Interv 2021;25:100411 [FREE Full text] [doi: 10.1016/j.invent.2021.100411] [Medline: 34401370]
50. Thoma S, Schwänzl I, Galbusera L. Reopening selves: phenomenological considerations on psychiatric spaces and the therapeutic stance. Psychopathology 2022;55(3-4):156-167 [FREE Full text] [doi: 10.1159/000517888] [Medline: 34352794]
51. Blease C, Torous J, McMillan B, Hägglund M, Mandl KD. Generative language models and open notes: exploring the promise and limitations. JMIR Med Educ 2024;10:e51183 [FREE Full text] [doi: 10.2196/51183] [Medline: 38175688]

Abbreviations

COREQ: Consolidated Criteria for Reporting Qualitative research
HCP: health care professional
ICD: International Classification of Diseases
MHB: Brandenburg Medical School
NTNU: Norwegian University of Science and Technology
ORA: online record access
PEPPPSY: Piloting and Evaluation of a Participatory Patient-Accessible Electronic Health Record in Psychiatry and Somatics
SMI: severe mental illness

Edited by J Torous; submitted 22.07.23; peer-reviewed by E Matthews, T Hatlen Nøst, J Hagström; comments to author 18.08.23; revised version received 10.11.23; accepted 29.11.23; published 05.02.24.

Please cite as:
Schwarz J, Meier-Diedrich E, Neumann K, Heinze M, Eisenmann Y, Thoma S
Reasons for Acceptance or Rejection of Online Record Access Among Patients Affected by a Severe Mental Illness: Mixed Methods Study
JMIR Ment Health 2024;11:e51126
URL: https://mental.jmir.org/2024/1/e51126
doi:10.2196/51126
PMID:38315523

Original Paper

A Novel Blended Transdiagnostic Intervention (eOrygen) for Youth Psychosis and Borderline Personality Disorder: Uncontrolled Single-Group Pilot Study

Shaunagh O'Sullivan[1,2], BA, MPsychSc; Carla McEnery[1,2], BPsychSc (Hons), PhD; Daniela Cagliarini[1,2], BA, GDipPsych, PG Dip Psych, MPsych; Jordan D X Hinton[1,2,3], BA, GDip(Psych); Lee Valentine[1,2], BA, MSocWk, PhD; Jennifer Nicholas[1,2], BA, BSc (Hons), PhD; Nicola A Chen[1,2], BSc (Hons), PhD; Emily Castagnini[1,2], BSci, MPH; Jacqueline Lester[1], MPsych; Esta Kanellopoulos[1], BA/LLB, MPsych/PhD; Simon D'Alfonso[4], BASc (Hons), PhD; John F Gleeson[3], BA (Hons), MPsych, PhD; Mario Alvarez-Jimenez[1,2], DClinPsy, MAResearchMeth, PhD

[1]Orygen, Parkville, Australia

[2]Centre for Youth Mental Health, University of Melbourne, Melbourne, Australia

[3]Health Brain and Mind Research Centre, School of Behavioural and Health Sciences, Australian Catholic University, Melbourne, Australia

[4]School of Computing and Information Systems, University of Melbourne, Melbourne, Australia

Corresponding Author:
Shaunagh O'Sullivan, BA, MPsychSc
Orygen
35 Poplar Road
Parkville
Australia
Phone: 61 428282470
Email: shaunagh.osullivan@orygen.org.au

Abstract

Background: Integrating innovative digital mental health interventions within specialist services is a promising strategy to address the shortcomings of both face-to-face and web-based mental health services. However, despite young people's preferences and calls for integration of these services, current mental health services rarely offer blended models of care.

Objective: This pilot study tested an integrated digital and face-to-face transdiagnostic intervention (eOrygen) as a blended model of care for youth psychosis and borderline personality disorder. The primary aim was to evaluate the feasibility, acceptability, and safety of eOrygen. The secondary aim was to assess pre-post changes in key clinical and psychosocial outcomes. An exploratory aim was to explore the barriers and facilitators identified by young people and clinicians in implementing a blended model of care into practice.

Methods: A total of 33 young people (aged 15-25 years) and 18 clinicians were recruited over 4 months from two youth mental health services in Melbourne, Victoria, Australia: (1) the Early Psychosis Prevention and Intervention Centre, an early intervention service for first-episode psychosis; and (2) the Helping Young People Early Clinic, an early intervention service for borderline personality disorder. The feasibility, acceptability, and safety of eOrygen were evaluated via an uncontrolled single-group study. Repeated measures 2-tailed *t* tests assessed changes in clinical and psychosocial outcomes between before and after the intervention (3 months). Eight semistructured qualitative interviews were conducted with the young people, and 3 focus groups, attended by 15 (83%) of the 18 clinicians, were conducted after the intervention.

Results: eOrygen was found to be feasible, acceptable, and safe. Feasibility was established owing to a low refusal rate of 25% (15/59) and by exceeding our goal of young people recruited to the study per clinician. Acceptability was established because 93% (22/24) of the young people reported that they would recommend eOrygen to others, and safety was established because no adverse events or unlawful entries were recorded and there were no worsening of clinical and social outcome measures. Interviews with the young people identified facilitators to engagement such as peer support and personalized therapy content, as well as barriers such as low motivation, social anxiety, and privacy concerns. The clinician focus groups identified evidence-based content as an implementation facilitator, whereas a lack of familiarity with the platform was identified as a barrier owing to clinicians' competing priorities, such as concerns related to risk and handling acute presentations, as well as the challenge of being understaffed.

Conclusions: eOrygen as a blended transdiagnostic intervention has the potential to increase therapeutic continuity, engagement, alliance, and intensity. Future research will need to establish the effectiveness of blended models of care for young people with complex mental health conditions and determine how to optimize the implementation of such models into specialized services.

(JMIR Ment Health 2024;11:e49217) doi:10.2196/49217

KEYWORDS

digital intervention; blended care; youth mental health; transdiagnostic intervention; psychotic disorders; borderline personality disorder; digital health; mobile phone

Introduction

Background

The evolution of specialist early intervention services for youth represents a major global reform of mental health services [1-3]. However, there are shortcomings that remain to be addressed for these services to fully deliver on their promise; for example, 42% of young people drop out of treatment by the third therapy session [4], indicating low engagement rates with early intervention services [5]. Furthermore, those who continue treatment receive time-limited support [6], and up to 80% of young people with severe mental health conditions will incur repeated relapses, leading to long-term disability and high societal cost [7,8]. Estimates suggest that the cost associated with recurring mental ill-health is up to 5 times that of nonrelapsing presentations [9]. Even when young people receive evidence-based treatment, its effectiveness is limited [10]; for example, between one-third and two-thirds of young people do not experience symptom reduction [11], and functional impairment often remains an issue after remission [12].

Digital technologies have the potential to address these challenges and limitations by enhancing the accessibility, impact, reach, and cost-effectiveness of youth mental health (YMH) services [13,14]. Many young people, recognizing their need for help, are turning to technology, including smartphone apps, websites, and social media, to self-manage their mental well-being [15]. Research findings support the efficacy of digital mental health interventions in improving treatment outcomes in severe mental health conditions such as psychosis [16-18] and borderline personality disorder (BPD) [19]. Web-based treatment programs have also demonstrated efficacy comparable to that of face-to-face psychotherapy [20-22], and self-guided smartphone-based mental health interventions are proving to be promising self-management tools for depression and anxiety symptoms [23].

However, despite the evidence for the effectiveness of digital interventions, limitations have also been reported, such as high attrition rates [24,25], a focus on mild to moderate mental health conditions [26-28], a focus on single disorders ignoring potential comorbidity [29], and a lack of integration within clinical settings [30,31]. Factors affecting attrition and dropout from digital interventions have been identified, such as a lack of personalization within digital interventions and severe mental disorders hampering engagement with interventions [32]. Most of the first generation of digital interventions have also been deployed and evaluated without face-to-face care, generating a divide between face-to-face and digital supports [33]. Furthermore, clinical trials of digital interventions often recruit highly motivated early adopters from the community, resulting in poor generalizability of the findings in clinical settings [33,34].

Blended care refers to treatment that includes face-to-face and digital elements, both of which contribute to the treatment process and can be integrated or offered sequentially [35,36]. Blended models of care offer an innovative approach to address the limitations of both face-to-face and digital therapy for young people with serious mental illness, while maintaining the strengths of both modalities [36,37]. This integrated approach is in line with World Health Organization recommendations [38] and young people's preferences, as well as national and international calls for the integration of web-based and in-person mental health services [39]. Young people have indicated that blended models of care could enhance clinical care by increasing accessibility and continuity of care, providing access to posttherapy support, and strengthening the relationship with their clinician [40]. However, despite young people's preferences and calls for the integration of face-to-face and web-based services [16,41], current mental health services rarely provide this type of integrated web-based support [42]. Barriers to implementing digital interventions in clinical settings have also been identified, such as a lack of time for clinicians and skepticism toward digital interventions [43], and these need to be taken into account when developing blended interventions. Furthermore, there is limited research testing blended models of care as a treatment approach [37,44], despite the demonstrated efficacy of stand-alone digital interventions [45-47].

Furthermore, transdiagnostic interventions that target underlying mechanisms or symptoms that are common across multiple mental disorders have the potential to provide more effective, personalized, and engaging treatment, addressing comorbidity and being applicable to a wider range of young people [48,49]. Evidence also suggests that transdiagnostic interventions may be at least as effective, more engaging, and easier to scale up in real-world clinical settings compared with single-disorder interventions [50,51].

As blended transdiagnostic interventions for first-episode psychosis and BPD have not yet been evaluated, a 3-month pilot evaluation of a blended transdiagnostic digital intervention (eOrygen) designed to enhance the accessibility, responsiveness, and impact of face-to-face specialized YMH clinical services for youth psychosis and BPD was carried out. Pilot studies are an important first step before running a full-powered clinical trial because they focus on whether an intervention can be carried out, whether it would be worth proceeding with it, and how to proceed before focusing on evaluating the effectiveness of the intervention [52].

Objectives

The primary objective of the eOrygen pilot was to evaluate the feasibility, acceptability, and safety of an integrated web-based clinic that blends moderated online social therapy (MOST) support with face-to-face specialized YMH clinical services for youth psychosis and BPD. A secondary aim of the pilot was to assess changes in key clinical and psychosocial outcomes for young people from the point of enrollment in eOrygen to after the intervention. Furthermore, the failure to integrate digital technologies into routine practice is well documented [53], and, therefore, an additional objective of this study was to understand young people's and clinicians' experiences of barriers and facilitators to using a blended model of care in clinical practice.

Methods

Study Design

This was a 3-month multisite pre-post single-group pilot study conducted at the Early Psychosis Prevention and Intervention Centre (EPPIC) and the Helping Young People Early (HYPE) Clinic at Orygen Youth Health in Melbourne, Victoria, Australia. EPPIC provides services for young people aged 15 to 25 years experiencing their first episode of psychosis, and admission to the service is based on a clinical assessment determining the presence of full-threshold first-episode psychosis, including full-threshold psychotic symptoms such as hallucinations, delusions, or formal thought disorder [54]. The HYPE Clinic offers an early intervention program for young people aged 15 to 25 years with BPD, and intake to the service is based on meeting ≥3 BPD criteria according to the *DSM-5* (*Diagnostic and Statistical Manual of Mental Disorders* [Fifth Edition]) Text Revision [55]. These services deliver specialized early interventions, with treatment offered from 6 months to a maximum of 2 years. Each young person at Orygen Youth Health receives case management by a dedicated mental health clinician, with additional assessment and treatment support provided by a psychiatrist. Each year, approximately 155 young people access HYPE Clinic services, and approximately 250 young people access EPPIC services.

Sample Size

It is recommended that the sample size of a pilot study be approximately 10% of the sample size projected for the larger parent study [56]. At present, there is little definitive research available to determine the sample size of the larger parent randomized controlled trial for a transdiagnostic blended model of care, hence the importance of conducting this pilot feasibility study. In a recent study using the same technology, a total sample of 140 participants was determined to detect changes in social functioning at 90% power, accounting for attrition of 20% [57]. Given this, we proposed to recruit 25 clinicians and 1 to 2 young people per clinician, resulting in an anticipated 25 to 50 young people in the study.

Ethics Approval

Ethics approval was obtained from the Melbourne Health Human Research Ethics Committee (HREC/49492/MH-2019).

Participants and Procedure

The participants were 18 mental health clinicians and 33 young people recruited by participating clinicians and the Orygen research team across HYPE Clinic and EPPIC clinical services. Clinician recruitment took place over a 2-month period, and the recruitment of young people took place over a 4-month period from May 15 to September 25, 2020.

Young People

Inclusion and Exclusion Criteria

Young people were recruited via their clinician and the research team. Clinicians within each service were invited to identify potentially eligible young people based on the following criteria: (1) aged 15 to 25 years (inclusive), (2) currently receiving treatment at the HYPE Clinic or EPPIC, (3) engaged with treatment as judged by the treating clinician and not approaching discharge from service, (4) willing to nominate an emergency contact person, (5) have regular and ongoing internet and telephone access, and (6) able to give informed consent and comply with study procedures. The exclusion criteria were as follows: (1) young people with an intellectual disability who were unable to meet the cognitive demands of the web-based intervention, interfering with the likelihood of benefiting from the intervention as judged by their treating clinician; and (2) young people with an inability to converse in, or read, English. There were no specific exclusion criteria related to level of suicide risk or interpersonal hostility (ie, a consideration for harm to self or others while engaging within a web-based social network). However, clinicians were consulted on a case-by-case basis regarding participant suitability, and clinician judgment regarding suitability could be reassessed at any time. The exclusion criteria for the pilot study were kept to a minimum both to facilitate the recruitment process and to ensure that the intervention was tested and adequately mirrored the real-world characteristics of the broad population of young people accessing specialist YMH services. Furthermore, to mirror the intended real-world implementation of eOrygen, these exclusion criteria were assessed and monitored by the participating clinicians.

Recruitment Process

Once eligibility was determined, eligible young people were invited to participate in the study by the research team. Among the HYPE Clinic clients, of the 106 young people who were assessed for eligibility for this study, 42 (39.6%) met the inclusion criteria; however, 23 (55%) of these 42 young people could not be approached because of their involvement with another research study. Thus, 19 young people were approached to participate, and 15 (79%) agreed to participate and enrolled in the study, whereas 4 (21%) declined. Of the 106 young people assessed for eligibility, 64 (60.4%) were ineligible to participate in this study owing to clinical risk, poor engagement with treatment, a lack of access to technology, age, or because they were approaching discharge.

Among the EPPIC clients, of the 59 young people who were assessed for eligibility, 43 (73%) met the inclusion criteria; however, 3 (7%) of these 43 young people were already participating in another research study. Thus, 40 young people were approached to participate, and 20 (50%) agreed to participate, whereas 11 (28%) declined, and 9 (23%) could not be contacted by the research team. Of the 20 young people who agreed to participate, 2 (10%) could not be contacted to complete their baseline assessments and did not enroll in the study; the remaining 18 (90%) enrolled in the study. However, of these 18 young people, 7 (39%) were lost to follow-up during the study period (n=3, 43% before onboarding to the eOrygen platform and n=4, 57% before completing the 3-month follow-up assessments). The young people lost to follow-up were unresponsive to telephone calls and messages from the research team but were still engaged with face-to-face treatment with their clinicians. Of the 59 young people assessed for eligibility, 16 (27%) were ineligible to participate in this study for the same aforementioned reasons.

Participant consent was obtained from those interested in participating and parental or guardian consent was also obtained for young people aged <18 years.

Assessments

The consenting young people were contacted at baseline via email and telephone to complete baseline measures before setting up their eOrygen user account. The young people then continued treatment with their clinician while having access to the eOrygen platform for 3 months. They were contacted again at the end of the 3-month intervention to complete the postintervention assessments.

During postintervention follow-up telephone call assessments, the young people were asked whether they were willing to be contacted for a subsequent qualitative interview, and 19 (58%) of the 33 young people agreed to be followed up. After the intervention phase, a randomly selected subgroup comprising 12 (63%) of the 19 young people who agreed to be contacted were invited via SMS text messaging to participate. These semistructured qualitative interviews were designed to explore their experiences with the eOrygen platform. The primary goal of these interviews was to identify both barriers and facilitators to their engagement with the intervention.

Of the 12 young people approached after the intervention, 1 (8%) declined to participate (no reason provided), and 2 (17%) agreed to participate but did not attend the scheduled interviews and were not able to be contacted; thus, 9 (75%) participants successfully completed the interview process. However, a technical issue resulted in a recording failure during 1 (11%) of the 9 interviews, and it could not be included in the subsequent analysis.

The interviews were all conducted via Zoom (Zoom Video Communications, Inc), and interview times ranged from 22 to 38 minutes. Participants were recruited and interviewed by a study research assistant and author EC. Interview questions were underpinned by a user-centered design approach [24] and focused on the following aspects: what initially interested participants about using the eOrygen platform; the experience of onboarding; hopes and expectations; barriers and facilitators; and the overall experience of the therapy journeys, clinical and peer support, and community features of the platform.

Clinicians

All mental health clinicians employed at the HYPE Clinic and EPPIC were eligible for inclusion in this study. Clinicians attended a workshop focused on learning about the background of the intervention, including previous empirical findings using the same technology and how to use the eOrygen platform. The latter aspect concerned how to use the intervention functions and set up an account, with suggestions provided on how to integrate this into the clinicians' work with young people, with the possibility of using it within and between face-to-face sessions as they felt appropriate. This included clinical case studies that applied to the populations of both HYPE Clinic and EPPIC services and were coauthored by clinicians at these services. Clinicians were also provided a training manual that was used to help them navigate the platform during the workshop and also to keep and reuse as necessary when navigating the intervention platform independently. As the workshop was held before recruiting young people to the study, clinicians were provided with a training video at a later date describing once again how to use the intervention platform features.

Eligible clinicians were then invited to identify eligible young people who met the aforementioned inclusion criteria. All participating clinicians had at least 1 young person using the eOrygen platform. The clinicians were contacted via email at baseline to complete the clinician-rated measures. They were then invited to use the eOrygen platform with their clients for 3 months. At the end of the intervention, they were contacted again via email to complete the postintervention clinician-rated measures.

After the postintervention phase, an invitation was extended to all 18 clinicians to participate in a structured focus group session. The aim of this session was to delve into the obstructive and conducive factors influencing the effective implementation of the eOrygen platform into routine clinical practice. This evaluative process was grounded in the Consolidated Framework for Implementation Research (CFIR), a recognized and widely used framework for assessing the determinants influencing implementation in health care settings [58]. The interview schedule was designed by author LV and based on the CFIR constructs that were identified via both formal and informal consultation with the specialist service settings throughout the intervention period. Among the notable constructs under consideration were those related to evidence, adaptability, complexity, needs and resources, and self-efficacy.

Clinicians were invited via email and supported by their line managers to attend. Of the 18 clinicians, 15 (83%) were available to participate in the focus groups that were conducted over Zoom. Because of the number of clinicians available to participate, the focus groups were divided into 3 distinct sessions to facilitate more extensive discussions and allow individual clinicians ample opportunity to share their insights and experiences. Authors LV, DC, and EK each conducted 1 of the 3 parallel sessions.

Intervention: eOrygen

The eOrygen intervention was based on Orygen Digital's MOST model, which was the first digital solution to offer continuous integrated face-to-face and digital care to young people across the mental health diagnostic and severity spectrum and stages of treatment [59-61]. In partnership with young people, the MOST model was iteratively developed by a multidisciplinary team of researchers, clinical psychologists, programmers, creative writers, graphic artists, and experts in human-computer interaction [61-63]. A recent clinical trial with young people with psychosis demonstrated that an intervention based on the MOST model was effective in improving vocational and educational outcomes as well as reducing the use of emergency services; in addition, it was cost-effective, with evidence of a dose-response effect [60,64,65].

The eOrygen intervention was a purpose-built web-based platform designed to integrate face-to-face and web-based support for young people experiencing mental ill-health (Figure 1). This was achieved through the use of both clinician and young person user accounts, making it possible for young people and their treating clinician to use the platform throughout the treatment process, during their face-to-face sessions or between sessions. Young people could also use the platform in a self-directed way, but no automated prompts or reminders were provided for young people to use eOrygen between sessions, unless clinicians made suggestions to their young clients using the platform.

The platform was designed to enhance, not replace, recommended treatments for mental health conditions (eg, clinician-administered cognitive behavioral therapy). eOrygen comprised interactive user-directed psychosocial interventions (*therapy journeys*), a social network, clinical moderation, and peer support.

Therapy journeys comprised collections of *therapy activities* relating to different themes. Themes related to the treatment of mental ill-health, such as managing social anxiety, anxiety, and depression, as well as social functioning. Users were assigned a suggested *therapy journey* based on their responses to a questionnaire they completed after being onboarded to the eOrygen platform, providing personalized content specific to their individual mental health concerns (Figure 2). Users could complete multiple *therapy journeys*, and clinicians or young people could change the assigned journey.

Therapy activities could be accessed as part of a *therapy journey* or as stand-alone activities via the *explore* function. The *explore* function enabled young people to use a search bar to locate therapy content of interest, and eOrygen clinicians could also recommend personalized content to young people using this function. *Therapy activities* included *activities*, *comics*, *talking points*, and *actions*. *Activities* comprised written content, and *comics* comprised storyboard panels focusing on a particular therapeutic theme and target related to the treatment of mental ill-health challenges. *Talking points* enabled participants to propose a solution to identified problems (eg, how to incorporate mindfulness into everyday activities), which encouraged social problem-solving and effective peer modeling. *Actions* were behavioral prompts that young people could complete to translate learning on a mental health topic into behavior change.

The eOrygen *social network* was moderated by trained peer workers, who were young people who had a lived experience of mental illness. The *social network* included a community newsfeed and individual profile pages where participants and peer workers could create *posts* to share thoughts, information, pictures, and videos (Figure 3). They could also respond to other users' posts through *comments* or *reactions*. *Reactions* were designed to facilitate social support (eg, "I get you" and "Thinking of you"). Likewise, *talking points* were designed as collaborative spaces to discuss specific topics by leaving *comments*. Young people were also able to receive direct support from peer workers and clinicians on the platform via private *messages*.

Figure 1. eOrygen as a blended model of care.

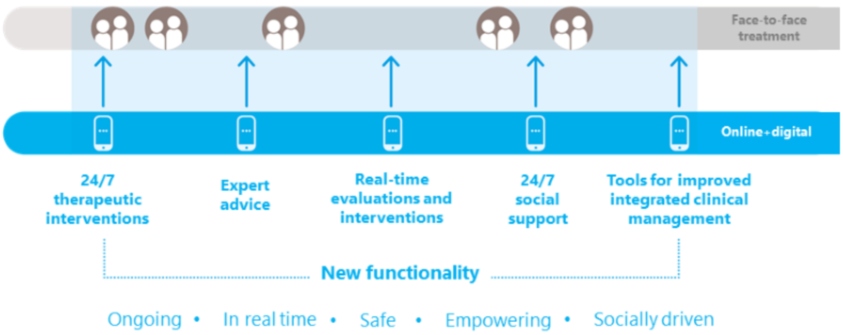


Figure 2. eOrygen therapy journey.

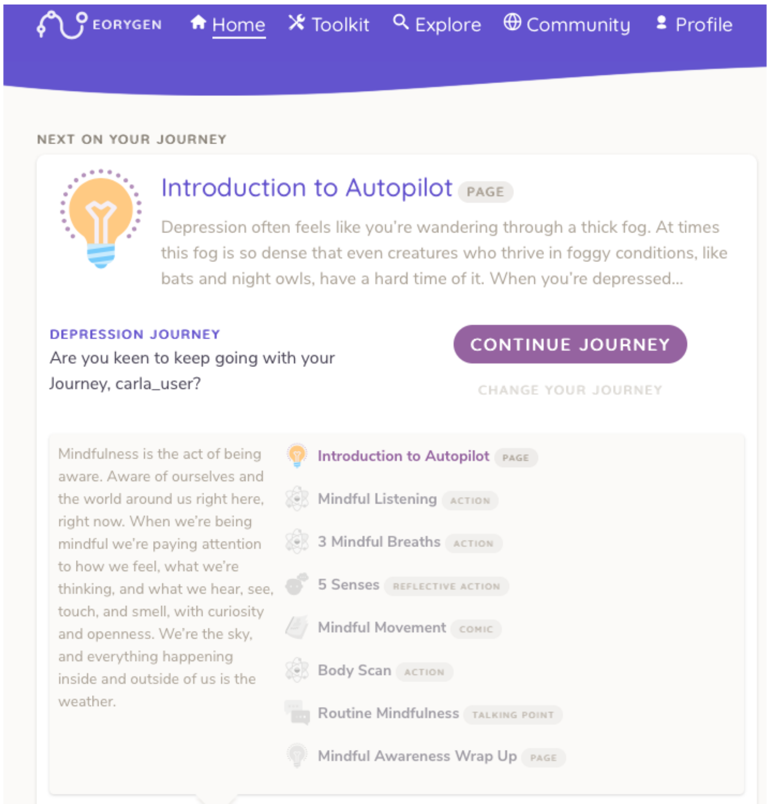
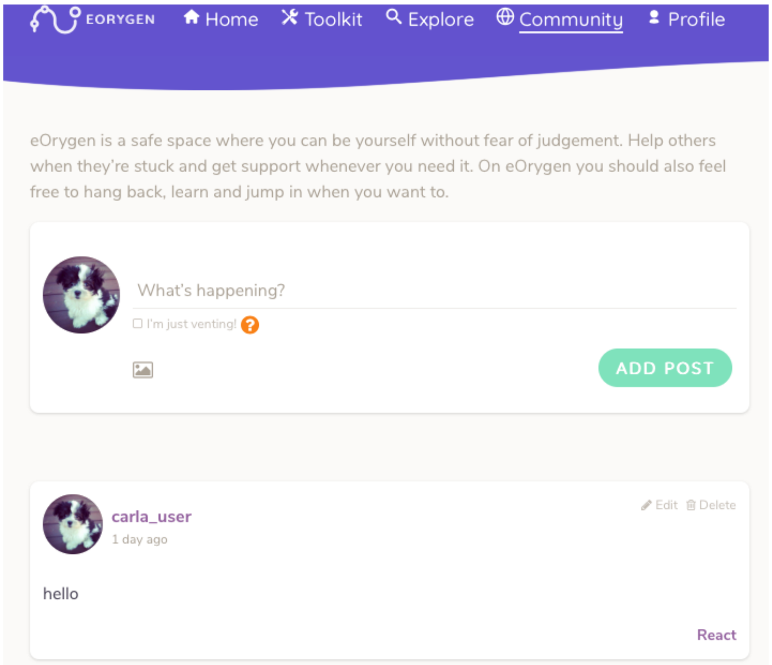


Figure 3. eOrygen social network.



Outcome Measures

Feasibility

The feasibility of eOrygen was measured by tracking recruitment to the study, which is in line with other studies testing the feasibility of digital interventions [66]. Although no a priori minimum or maximum number of participants per clinician or per clinic was specified, to accurately assess feasibility, we anticipated recruiting approximately 15 clinicians from EPPIC and 1 to 2 young people per each participating clinician as well as 10 clinicians from the HYPE Clinic and 1 to 2 young people per clinician.

Feasibility was indicated if (1) the recruitment goal was met and (2) the participant refusal rate was <50%. If the recruitment goal was not met after 2 months of recruitment, barriers to recruitment were to be identified.

Acceptability

Intervention acceptability was measured via responses to a feedback questionnaire. The pilot was considered to indicate acceptability of the eOrygen intervention if at least 90% of the young people reported that they would recommend it to others, which is in line with a previous pilot study testing the acceptability of a digital mental health intervention [59].

Safety

Intervention safety was measured by analyzing reports of any adverse events, tracking the security of the web-based system, and analyzing responses to a feedback questionnaire, following a similar protocol for a previous study [7,67]. An adverse event was defined as any unfavorable or unintended sign, symptom, or disease temporally associated with the use of the intervention, whether or not it was related to the intervention. A serious adverse event was defined as any untoward medical occurrence that could be life threatening, result in death, require inpatient hospitalization, or result in persistent or significant disability. All participants were closely monitored by clinical moderators for adverse events and serious adverse events. The research team members were trained in study procedures, including adverse event assessments, and attended good clinical practice training. In addition, treating clinicians were asked to report any adverse events, including suicide attempts and serious self-harm, to the research team.

The pilot was considered to indicate the safety of eOrygen if (1) there were no unlawful entries recorded in the eOrygen system during the pilot, (2) no young people experienced a serious adverse event as a result of their engagement with the system during the 3-month intervention period, (3) at least 95% of the young people reported it to be safe via the feedback questionnaire, and (4) clinical and social measures did not show a worsening pattern over the course of the study.

Safety was also reported by assessing pre-post changes in BPD symptomatology for HYPE Clinic participants as measured by the 23-item version of the Borderline Symptom List [68], and pre-post changes in psychotic symptoms for EPPIC participants as measured by 3 items of the expanded Brief Psychiatric Rating Scale (version 4.0), including suspiciousness, hallucinations, and unusual thought content [69].

Potential Clinical Effects

Potential clinical effects were assessed by measuring pre- to postintervention changes in clinical and psychosocial outcomes at baseline and at 3 months. Clinician-rated measures included social and occupational functioning as measured by the Social and Occupational Functioning Assessment Scale [70] and therapeutic alliance (TA) as measured by the Working Alliance Inventory–Short Revised (therapist version) [71].

Young people self-report measures included depression as measured by the Patient Health Questionnaire-9 [72]; TA with their face-to-face clinician as measured by the Working Alliance Inventory–Short Revised (client version) [73]; psychological well-being as measured by the Flourishing Scale [74]; self-determination as measured by the Basic Psychological Needs Satisfaction Questionnaire [75]; loneliness as measured by the University of California, Los Angeles Loneliness Scale (version 3) [76]; social isolation as measured by the Friendship Scale [77]; social anxiety as measured by the Social Interaction Anxiety Scale [78]; stress as measured by the Perceived Stress Scale [79]; and psychological distress as measured by the 10-item Kessler Psychological Distress Scale [80]. These measures have been validated in a youth population and were chosen for their demonstrated reliability. All baseline measures were completed before onboarding participants to the eOrygen platform, and all assessments were completed via Qualtrics (Qualtrics International Inc) where possible or otherwise administered by a research assistant over the telephone.

Satisfaction Survey Feedback

Purpose-designed questionnaires administered via Qualtrics were used to assess user satisfaction and user feedback.

Statistical Analyses

Overview

This study used a mixed methods design involving both quantitative and qualitative data, which allowed for a more robust analysis [81]. Quantitative data assessing the feasibility, acceptability, safety, and potential clinical effects of the intervention were measured at baseline and at 3-month follow-up. Qualitative data were used to assess young people's and clinicians' barriers and facilitators to implementing eOrygen within clinical services, which enhanced our understanding of the feasibility and acceptability aspects of the intervention and provided valuable information for designing a full-powered trial that would have not been achieved through quantitative data strands alone.

Quantitative Analyses

Chi-square tests showed no differences between the clinical sites in baseline demographic and clinical and psychosocial outcomes. Therefore, data were pooled, repeated measures 2-tailed *t* tests were conducted, and within-group effect sizes (Cohen *d*) were reported for changes in pre- to postintervention scores on effectiveness outcome measures. Parametric and nonparametric correlations were conducted to explore the association between the use of eOrygen and the degree of change between before and after the intervention on effectiveness outcome measures, but no relationships between use and effectiveness outcomes were found.

Aggregated data from the user satisfaction questionnaire and descriptive statistics from insights into using eOrygen were reported as exploratory findings. The data for the acceptability criterion of whether young people would recommend eOrygen to others were derived from insights into eOrygen descriptive statistics. The data for 1 of the safety criteria regarding whether young people felt safe using eOrygen were derived from the user satisfaction questionnaire. Statistical analyses were performed using SPSS (version 27.0; IBM Corp).

Qualitative Analyses

Young People

Given the user-centered design approach, thematic analysis was considered the most appropriate method of data analysis [82].

The data analysis was conducted by author EC under the supervision of author LV.

The analysis process used an inductive approach in which EC explored young people's experiences for factors relating to barriers and facilitators to young people's engagement with the blended model of care. To gain familiarity with the data set, the interview transcripts were read and reread. Subsequently, initial codes were applied to the transcripts to identify relevant factors to engagement. Any recurring codes, both within and across different transcripts, were identified and recorded.

Following the principles of thematic analysis, these codes were then grouped into preliminary themes and subjected to thorough review in relation to all other identified themes. Some themes were identified as superordinate, representing broader categories of experience, whereas others assumed subordinate positions, delineating into subthemes.

Clinicians

Three focus groups were conducted by authors DC, EK, and LV with 15 service clinicians overall. Focus groups comprised a mix of both HYPE Clinic and EPPIC clinicians and lasted between 31 and 44 minutes. Interview questions were underpinned by the CFIR [58], which is one of the most widely used frameworks for identifying factors impacting implementation outcomes [83]. The CFIR comprises 39 constructs across 5 domains. The domains identified as most relevant to this implementation setting in the preimplementation phase were *inner and outer settings*, *individual characteristics*, and *innovation characteristics*. The clinician focus groups underwent deductive coding by authors JN and LV in accordance with the CFIR. JN and LV then engaged in a rigorous discussion to examine how the attributes within the identified domains acted as either obstacles or enablers to the implementation of eOrygen in this clinical setting.

Use Metrics

Use metrics were used to measure engagement with eOrygen. *Number of active days* was used as an overall metric for platform use and referred to the number of days a young person accessed eOrygen after completing onboarding until the end of the 3-month intervention period.

Therapy views comprised the number of times a young person opened *therapy activities* via a *therapy journey* or via a search function. Users were encouraged to revisit activities, and repeat views were counted within the number of *therapy views*. *Journey components completed* was a count of the number of unique *therapy activities* a user completed within a *therapy journey*.

Engagement with the social networking component of eOrygen was measured by the *number of posts*, *comments*, and *reactions made* by young people on the social network. This included *posts made* on the newsfeed and on individual users' profiles. *Comments made* could be in response to *posts made* by other users or peer workers or *talking point* therapy activities. *Reactions made* could be in response to any post or comment on the social network.

There was also a chat function where young people could communicate with eOrygen staff (clinicians and peer workers) through private direct messages on the platform. Engagement with clinicians and peer workers was measured by the *number of messages sent* (by young people) and the *number of messages received* (from clinicians or peer workers).

Results

Demographics

Participants were aged between 15 and 24 (mean 19.48, SD 2.84) years, and the majority were female individuals (21/33, 64%). Of the 33 participants, 27 (82%) were born in Australia, 7 (21%) spoke languages other than English at home, 3 (9%) identified as Aboriginal or Torres Strait Islander, 9 (27%) were engaged in paid work, and 22 (66%) were studying part time or full time (Table 1).

Table 1. Participants' descriptive statistics (total and by each clinical site).

	Total sample (n=33)	Clinical sites	
		HYPE[a] Clinic (n=15)	EPPIC[b] (n=18)
Age (years), mean (SD)	19.48 (2.84)	18.8 (2.70)	20.21 (2.89)
Age (years), n (%)			
≤18	13 (45)[c]	8 (53)[d]	5 (36)[e]
>18	16 (55)[c]	7 (47)[d]	9 (64)[e]
Gender, n (%)			
Man	7 (21)	3 (20)	4 (22)
Woman	21 (64)	11 (73)	10 (56)
Transgender, genderqueer or nonconforming, and other (including multiple gender selections)	5 (15)	1 (7)	4 (22)
Aboriginal or Torres Strait Islander			
Yes	3 (9)	2 (13)	1 (6)
No	29 (88)	13 (87)	16 (89)
Unsure	1 (3)	0 (0)	1 (6)
Currently studying			
Yes	22 (67)	10 (67)	12 (67)
No	11 (33)	5 (33)	6 (33)
Currently employed			
Yes	9 (27)	4 (27)	5 (28)
No	24 (73)	11 (73)	13 (72)

[a]HYPE: Helping Young People Early.

[b]EPPIC: Early Psychosis Prevention and Intervention Centre.

[c]n=29.

[d]n=15.

[e]n=14.

Feasibility

We sought to recruit 25 clinicians; the final number of clinicians enrolled in the study was 18 (78%). The final number of young people recruited was 33 (which exceeded our goal of 1 young person per clinician). We also anticipated that the refusal rate would be <50%, which it was at 25% (15/59).

Acceptability

We exceeded our acceptability goal, with 92% (22/24) of the young people reporting that they would recommend eOrygen to others (Multimedia Appendix 1).

Safety

There were no unlawful entries recorded on the eOrygen platform during the study, there were no serious adverse events experienced by participants, there was no worsening of clinical and social outcome measures, and 96% (24/25) of the young people stated that they felt safe using the platform (Multimedia Appendix 2).

In terms of symptom monitoring, the mean score for BPD symptomatology for HYPE Clinic participants reduced from 2.65 (SD 0.75) at baseline to 1.99 (SD 0.83) after the intervention. For the assessment of psychotic symptoms in EPPIC participants, the mean score for suspiciousness reduced from 2.72 (SD 1.64) at baseline to 0.94 (SD 3.78) after the intervention. In addition, the mean score for hallucinations reduced from 2.89 (SD 1.97) at baseline to 1.22 (SD 4.11) after the intervention, and the mean score for unusual thought content reduced from 1.72 (SD 1.13) at baseline to 0.67 (SD 3.74) after the intervention.

Potential Clinical Effects

Significant pre- to postintervention improvements were observed for 9 (82%) of the 11 clinical outcome measures, including social and occupational functioning, depression, psychological distress, social anxiety, social isolation, stress, borderline symptoms, and loneliness, as well as all aspects of therapist-rated working alliance, including goal, task, and bond, with effect sizes ranging from 0.56 to 0.89 (Table 2).

Table 2. Inferential statistics exploring the differences between before and after the intervention in well-being and mental health symptomology.

	Participants, n	Before the intervention, mean (SD)	After the intervention, mean (SD)	Mean difference (95% CI)	Repeated measures *t* test (*df*)	*P* value	Cohen *d*
Autonomy (BPNS[a] Questionnaire)	26	4.38 (1.08)	4.36 (1.02)	0.02 (−0.40 to 0.46)	0.13 (25)	.90	0.03
Competence (BPNS Questionnaire)	25	3.70 (0.89)	3.95 (1.01)	−0.25 (−0.64 to 0.15)	−1.30 (24)	.21	−0.26
Relatedness (BPNS Questionnaire)	26	4.44 (1.18)	4.62 (1.01)	−0.18 (−0.54 to 0.17)	−1.05 (25)	.30	−0.21
Depression (PHQ-9[b])	25	15.92 (7.57)	13.44 (7.93)	2.48 (0.82 to 4.14)	3.08 (24)	.005	0.62
Loneliness (University of California, Los Angeles Loneliness Scale)	25	54.92 (10.54)	50.38 (10.55)	4.54 (0.75 to 8.32)	2.47 (24)	.02	0.50
Psychological well-being (Flourishing Scale)	24	33.62 (10.04)	36.79 (9.63)	−3.17 (−6.51 to 0.16)	−1.97 (23)	.06	−0.40
Social isolation (Friendship Scale)	24	13.08 (5.71)	9.83 (5.51)	3.25 (1.19 to 5.30)	3.28 (23)	.003	0.70
Social anxiety (SIAS[c])	24	44.55 (14.43)	39.07 (15.28)	5.48 (2.34 to 8.62)	3.61 (23)	.001	0.74
Stress (PSS[d])	24	25.33 (6.89)	19.92 (6.43)	5.42 (3.01 to 7.83)	4.65 (23)	<.001	0.95
Psychological distress (K10[e])	24	33.24 (9.82)	29.30 (9.31)	3.94 (1.09 to 6.80)	2.86 (23)	.009	0.58
Client working alliance							
Goal (WAI-SRC[f])	24	9.63 (2.12)	9.67 (1.61)	−0.04 (−1.11 to 1.03)	−0.08 (23)	.94	−0.02
Task (WAI-SRC)	24	11.83 (1.17)	11.83 (1.46)	0.00 (−0.63 to 0.63)	0.00 (23)	.99	0.00
Bond (WAI-SRC)	23	12.39 (2.73)	12.78 (2.95)	−0.39 (−1.59 to 0.81)	−0.68 (22)	.51	−0.14
Therapist working alliance							
Goal (WAI-SRT[g])	17	10.76 (2.08)	11.88 (2.45)	−1.11 (−1.86 to −0.37)	−3.17 (16)	.006	−0.77
Task (WAI-SRT)	17	10.00 (1.66)	11.00 (2.09)	−1.00 (−1.93 to −0.07)	−2.29 (16)	.04	−0.56
Bond (WAI-SRT)	17	16.65 (2.55)	17.53 (2.03)	−0.88 (−1.58 to −0.18)	−2.67 (16)	.002	−0.65
Functioning (SOFAS[h])	18	57.50 (14.63)	64.28 (17.22)	−6.78 (−10.71 to −2.85)	−3.92 (17)	<.001	−0.86

[a]BPNS: Basic Psychological Needs Satisfaction.
[b]PHQ-9: Patient Health Questionnaire (9-item version).
[c]SIAS: Social Interaction Anxiety Scale.
[d]PSS: Perceived Stress Scale.
[e]K10: Kessler Psychological Distress Scale (10-item version).
[f]WAI-SRC: Working Alliance Inventory–Short Revised (client version).
[g]WAI-SRT: Working Alliance Inventory–Short Revised (therapist version).
[h]SOFAS: Social and Occupational Functioning Assessment Scale (clinician rated).

Young People's Satisfaction Survey Feedback

In terms of client feedback, 88% (21/24) of the young people reported that they would use eOrygen again. The top initial reasons of interest in using eOrygen included (1) practicing well-being skills (22/24, 92%), (2) contributing to research (22/24, 92%), (3) receiving support from clinicians (17/24, 71%), (4) connecting with others with similar mental health experiences (17/24, 71%), and (5) learning and well-being (17/24, 71%). Multimedia Appendix 1 presents a full list of reasons for the young people's use of eOrygen.

In terms of the young people's satisfaction with eOrygen, 96% (23/24) rated it as a positive experience, 88% (22/25) rated it as easy to use, and 83% (20/25) rated it as helpful (Multimedia Appendix 2).

Engagement

A total of 30 young people were onboarded to eOrygen (n=15, 50% from the HYPE Clinic and n=15, 50% from EPPIC). The mean *number of active days* on eOrygen was 8.4 (SD 7.5) days; 50% (15/30) of the young people had between 6 and 30 *active days*, 43% (13/30) had between 2 and 5 *active days*, and 7% (2/30) had 1 *active day* (Table 3). Of the 30 young people, 12 (40%) used the platform at least once per fortnight during the initial 6 weeks of the intervention, and 6 (20%) maintained fortnightly access across the entire 12 weeks.

In terms of therapy engagement specifically, 40% (12/30) of the young people viewed ≥3 therapy activities, 23% (7/30) viewed 1 to 2 therapy activities, and 37% (11/30) did not view any therapy activities. Half of the onboarded young people (15/30, 50%) began a therapy journey, whereas 40% (12/30) of the participants completed at least 1 journey component (ie, therapy activity within a therapy journey), 20% (6/30) completed 2 to 7 journey components, and 20% (6/30) completed ≥10 journey components (refer to Table 3 for a summary of all therapy engagement metrics).

In terms of social network engagement, it was mandatory for the young people to make an introductory post to the community as part of the onboarding process. However, almost half (14/30, 47%) of the young people made at least 1 additional post or comment, whereas 10% (3/30) of them made 2 additional posts or comments, 20% (6/30) of them made 3 to 5 additional posts or comments, and 17% (5/30) of them made >6 additional posts or comments. Furthermore, 47% (14/30) of the young people reacted to at least 1 post or comment on the social network, whereas 27% (8/30) of them reacted to >1 post or comment (refer to Table 3 for a summary of all therapy engagement metrics).

In relation to exchanging private messages with eOrygen staff using the chat function, 47% (14/30) of the young people sent at least 1 message, whereas 13% (4/30) of them sent 1 to 2 messages, 13% (4/30) of them sent 3 to 6 messages, and 20% (6/30) of them sent ≥10 messages. By contrast, 97% (29/30) of the young people received at least 1 private chat message from a clinician or peer worker, whereas 37% (11/30) of them received 1 to 3 messages, 37% (11/30) of them received 4 to 9 messages, and 23% (7/30) of them received ≥14 messages (refer to Table 3 for a summary of these metrics).

Table 3. User engagement with eOrygen.

Variables, n	All sites (n=30), mean (SD; range)	EPPIC[a] (n=15), mean (SD; range)	HYPE[b] Clinic (n=15), mean (SD; range)
Active days	8.4 (7.5; 1-30)	10.4 (8.6; 1-30)	6.3 (5.7; 1-23)
Therapy views	9.1 (21.5; 0-115)	12.9 (29.2; 0-115)	5.3 (8.5; 0-29)
Journey components completed	6.0 (18.3; 0-100)	9.3 (25.4; 0-100)	2.7 (5.2; 0-18)
Posts and comments made	3.3 (4.3; 1-22)	4.1 (5.4; 1-22)	2.6 (2.9; 1-11)
Reactions made	2.2 (4.3; 0-16)	3.3 (5.1; 0-16)	1.1 (3.1; 0-12)
Messages sent	4.6 (11.6; 0-62)	8.0 (15.8; 0-62)	1.1 (1.8; 0-6)
Messages received	8.1 (11.3; 0-60)	11.1 (15.3; 0-60)	5.2 (3.6; 1-15)

[a]EPPIC: Early Psychosis Prevention and Intervention Centre.
[b]HYPE: Helping Young People Early.

Qualitative Findings

Young People

A thematic analysis of the interviews with the young people was conducted, regarding their experiences with eOrygen, with the aim of identifying facilitators and barriers to their engagement with the platform (Table 4). Facilitators included clinician endorsement, which increased trust in the platform; the presence of peer support workers, fostering a sense of safety and support; a sense of community and connection with other users; personalized therapy content recommended by clinicians; the use of eOrygen for between-sessions work, supported by follow-up discussions; and an easy-to-use interface. Conversely, barriers to engagement included general low motivation, social anxiety hindering social interactions, privacy concerns, inflexible progression in modules, and periods of limited content and interactions on the platform. Addressing these barriers and leveraging the facilitators can enhance the platform’s appeal and effectiveness for a broader range of users seeking mental health support.

Table 4. Facilitators and barriers to young people's engagement with eOrygen as a blended model of care integrated into specialized services.

	Description
Facilitators	
Clinician endorsement	Clinician endorsement of eOrygen seemed to increase young people's trust in the platform. Several participants mentioned that their clinician's recommendation to use eOrygen was central to their signing up to use the platform.
Professional peer support workers	The presence of professional peer support workers on the platform was engaging and increased young people's sense of safety in the community space.
Community, connection, and belonging	Participants appreciated interacting with other young people or sharing space on the platform, even if they did not post much themselves. The sense of community and connection was a facilitator for engagement.
Personalized therapy content	Young people found it helpful when clinicians recommended specific content aligned with their therapeutic work in face-to-face sessions. Therapeutic modules and resources within the platform were seen as valuable, trustworthy, and informative.
Blended care	Young people commonly reported that eOrygen was used for between-sessions work. Follow-up discussions in sessions regarding the *homework* were described as helpful and valuable for consolidating learning and providing accountability to support homework completion.
Ease of use	When participants found the platform easy to navigate and understand, it positively impacted their engagement.
Barriers	
Low motivation	Several participants mentioned a general lack of motivation as a barrier to engagement.
Social anxiety	Some participants experienced anxiety or insecurity related to social interactions on the platform. This social anxiety acted as a barrier, and participants reported feeling too anxious to post or reply to other young people's posts. Several young people felt anxiety about the perceived lack of clarity regarding the rules of the online community.
Privacy concerns	Privacy and confidentiality concerns were mentioned by some participants, impacting their willingness to engage fully with the platform. Some young people reported that uncertainty about who could see their posts or information raised anxiety. Several young people expressed discomfort at the thought of their face-to-face clinician viewing personal content that they may post on the platform's social network.
Inflexible module progression	One interviewee identified the inability to skip sections or activities within a therapy module as a barrier and requested more autonomy when engaging with therapy modules. Some participants described desiring more flexibility with how they engaged with content, such as wanting to change therapy journeys (eg, from depression to anxiety) but not being able to.
Limited content and interactions	Periods of low activity or limited content being posted by other young people on the platform could lead to reduced engagement. Participants mentioned that the lack of new posts or interactions could be discouraging. More young people and greater interaction was requested by several young people.

Clinicians

The analysis of the clinician focus group grounded in the CFIR identified various barriers and facilitators to the successful implementation and use of the eOrygen intervention (Table 5). Among the notable barriers were the length of the onboarding process for young people, a need for increased confidence in using the platform, and a perceived lack of practical knowledge regarding its features and how these features are related to benefits for young people; in addition, competing priorities, such as addressing risk and acute presentations amid understaffing, consistently disrupted platform use. Conversely, the platform's positive reputation and alignment with evidence-based frameworks emerged as facilitators. In addition, the ability to access trustworthy content was highlighted as advantageous for clinicians. These findings underscore the multifaceted nature of implementing digital mental health resources and the importance of addressing both barriers and facilitators to optimize their effectiveness.

Table 5. Summary of factors impacting clinician implementation of eOrygen using the Consolidated Framework for Implementation Research (CFIR).

CFIR domain	CFIR construct	Description	Illustrative quote	Implementation impact
Characteristics of individuals	Self-efficacy	Clinicians expressed varying levels of confidence when it came to implementing the intervention. This variance in confidence had a notable influence on their motivation and capacity to use the platform as part of their clinical practice.	“[H]ard when I didn’t know exactly what I was prompting them to do sometimes.” (Focus group 1)	Barrier
Inner setting	Implementation climate: relative priority	Clinicians frequently cited competing priorities, including concerns related to young people’s risk, handling acute presentations, and the challenge of being understaffed, as significant barriers that impeded their ability to engage with, and become familiar with, the eOrygen platform.	“Might’ve been a little bit me as well of not feeling like I had the time or just forgetting sometimes to just the other priorities I have.” (Focus group 3)	Barrier
Outer setting	Young people’s needs and resources	Clinicians demonstrated a strong understanding of young people’s needs, including the complexities associated with addressing these needs effectively. This encompassed underaddressed comorbid conditions, clinical complexity, and barriers related to motivation and the specific age group within the cohort.	“All the comorbidities...we don’t always get to do the, you know, anxiety management stuff or like skills-based stuff as much as, um, I feel like as much as I did in other clinics.” (Focus group 3)	Facilitator
Intervention characteristics	Complexity	Clinicians considered the process for connecting and onboarding young people to the eOrygen platform to be lengthier and more in-depth than what young people typically expect when engaging with technology.	“’Cause young people want, unfortunately, they want everything now and right there and then. So, if they click and it doesn’t work, they’ll be like, ‘now there’s nothing, it’s not working.’” (Focus group 3)	Barrier
Characteristics of individuals	Knowledge and beliefs about intervention	Clinicians had positive regard for the eOrygen platform overall. However, their implementation was hindered by a lack of comprehensive knowledge about how the platform practically worked.	“We’re all probably a bit limited in thinking about how we’re using it.” (Focus group 2)	Facilitator and barrier
Intervention characteristics	Evidence strength and quality	Clinicians had the perception that the intervention had a strong evidence base and provided quality information to support young people. However, they also had a limited understanding of which components of the intervention led to effective outcomes.	“I didn’t have to go and do my own homework or ensure that I was using like, up-to-date and relevant information.” (Focus group 3)	Facilitator and barrier

Discussion

Principal Findings

To the best of our knowledge, this was the first study to test an integrated blended model of care for youth psychosis and BPD in young people aged 15 to 25 years. The results of this study showed that eOrygen was feasible, acceptable, and safe. In terms of feasibility, we anticipated recruiting approximately 25 clinicians and 1 to 2 young people per clinician to the intervention and expected that the refusal rate would be <50%. Our refusal rate was 25% (15/59), which indicated 1 element of feasibility. We sought to recruit 25 clinicians; the final number of clinicians enrolled in the study was 18 (78%). In addition, we recruited 33 young people (which exceeded our goal) to the study over a 4-month period. Although we did not meet our clinician goal, our recruitment took place at the beginning of the COVID-19 pandemic, which proved a difficult period to introduce a new digital intervention to YMH services and to train clinicians in the use of a new digital platform. Despite these challenges, we still recruited a relatively high percentage of clinicians to the study and exceeded our goal in recruiting young people.

In terms of acceptability, 92% (22/24) of the young people onboarded reported that they would recommend eOrygen to others, exceeding our goal of 90%. Furthermore, 40% (12/30) of the participants used the platform at least once per fortnight during the initial 6 weeks of the intervention period, although only 20% (6/30) maintained fortnightly access across the entire 12 weeks. These findings compare well with another study reporting decreased engagement over time, with retention rates of only 3.9% over 15 days and 3.3% over 30 days for the use of mental health apps in the general population [84]. Although engagement was strong during the first 6 weeks, more strategies are needed to sustain engagement over longer periods. We did not use any strategies in this pilot to promote young people’s or clinicians’ engagement and left this to the discretion of the participating clinicians because it was purely an ecological study. Therefore, future studies should implement scalable strategies to sustain engagement for clinicians and young people, such as external support, coaching, and automated prompts or reminders.

Higher engagement rates with eOrygen were also observed for young people attending EPPIC (mean 10.4, SD 8.6 active days) versus those attending the HYPE Clinic (mean 6.3, SD 5.7 active days). To the best of our knowledge, this was the first blended model tested for young people with BPD. The MOST model was originally developed and optimized for young people with first-episode psychosis [57,60,85]. Therefore, lower engagement rates for young people attending the HYPE Clinic could be because the therapeutic model used by clinicians in face-to-face care was slightly different than the content in eOrygen, and the MOST model may need further refinement and optimization for young people with BPD. Research has indicated that young people with BPD are difficult to engage in face-to-face treatment [86-88], and it is possible that this extends to digital mental health care.

In terms of safety, there were no unlawful entries recorded on the web-based platform, no serious adverse events were experienced by participants, and there was no worsening of clinical or social outcome measures during the intervention. We also anticipated that at least 95% of young people would report feeling safe using the platform, and this goal was exceeded with 96% of the participants reporting feeling safe. Our primary findings are also in line with a previous pilot study testing Orygen Digital's MOST platform with real-time clinician-delivered web chat counseling, which found that all acceptability and safety indicators exceeded their a priori established criteria [59].

The secondary outcome variables showed significant pre-post improvements in 9 (75%) of the 12 outcomes assessed. These included borderline symptoms, depression, loneliness, social isolation, social anxiety, stress, psychological distress, social and occupational functioning, and the therapist-reported working alliance. It is important to note that although our study included 2 clinical sites treating young people with complex mental health disorders (eg, psychotic disorders and BPD), there were no significant differences between the sites on outcome variables at baseline. Therefore, improvements in clinical outcomes relate to all participants in the study. The findings also support a previous pilot study (MOST+) that integrated MOST with real-time clinician-delivered web chat counseling [59]. MOST+ also found statistically significant improvements in psychological distress, depression, and stress. However, both studies are single-group pilot studies, and we cannot make causal inferences from the findings because it cannot be determined from uncontrolled studies whether the observed effects are related to the intervention or to external factors such as individual or in-person treatment characteristics. Future research should confirm these findings by conducting controlled studies with larger sample sizes and greater power.

Findings from a recent qualitative study also suggested that blended care has the potential to enhance the therapeutic relationship [40]. The study suggested that the TA developed through blended care can enhance engagement with both face-to-face and web-based treatment modalities by offering treatment continuity and personalization as well as enhancing therapeutic intensity, which are key areas of concern in the field [24,40]. In our study, we observed statistically significant improvements for therapist-reported TA but no improvements in client-reported TA. Research has also indicated that TA has moderate but reliable correlations with mental health outcomes [89-91]. Although we observed improvements in therapist-rated TA, they did not correlate with improvements in clinical outcomes; therefore, future research should further explore this, along with the importance of client-rated TA in relation to outcomes.

By contrast, qualitative feedback from participants in our study indicated that eOrygen was beneficial when used in a blended way; for example, young people found it helpful when their clinician recommended content to them that aligned with their in-session work, and they also found the homework to be completed on the eOrygen platform to be helpful for between-sessions work. These findings are in line with other research indicating that blended care is beneficial when it is integrated and intensifies treatment [92]. However, research has indicated that blended care may not be effective if perceived as burdensome or time consuming by clinicians, but this may be related to trial-related factors, such as reporting to a research team if using the intervention during a clinical trial or inflexible intervention structure [93]. One way to overcome this could be for developers to work with clinicians to ensure suitable content and for features to be provided within the intervention, which would enable clinicians to use the intervention with young people in a way that is meaningful, relevant, and related to the face-to-face treatment they provide [94].

Limitations

A number of limitations should be noted. The single-group design was chosen to enhance external validity by maximizing real-world uptake of the intervention. Therefore, it was important to expose as many young people and clinicians to the intervention as possible to determine real-world uptake with eOrygen, and the inclusion of a control group may have negatively impacted the number of clinicians and young people who signed up to the intervention, potentially limiting our understanding of feasibility in this context. However, this study design limited our ability to determine a cause-and-effect relationship between the intervention and outcomes [95]. Furthermore, a 3-month time frame was chosen because this is an acceptable time frame that is comparable to the time frames of other pilot studies testing the feasibility, safety, and acceptability of digital interventions [59,66,96].

Although this study tested an integrated blended model of care, we did not collect data on how young people engaged with face-to-face treatment or how participating clinicians used the eOrygen platform, and future studies should consider this when evaluating blended models of care. Furthermore, our goal to provide flexibility and autonomy to clinicians may have negatively impacted competence in the use of the platform; for example, clinicians attended a 1-day workshop and received a printed user manual and training videos on how to use the eOrygen platform. However, as noted in the clinician-identified barriers to implementation, attending 1 workshop may be inadequate to gain competence, and there was also a substantial gap in time between clinician training and the implementation of eOrygen owing to the COVID-19 pandemic. Therefore, the training may have been forgotten, and a lack of time to review training materials may have also been an issue. Future studies should consider providing ongoing clinician support in this regard, while also remaining flexible to the needs of clinicians. Furthermore, although the recruitment goal was met for this study, it must be noted that the sample of male participants in the study was small at only 21% (7/33) of the total sample, limiting the generalizability of these findings and highlighting that significant barriers may still exist for young men to access mental health treatment [97] and that difficulties may also exist in tailoring interventions to young men [98].

Conclusions

In conclusion, our pilot study was an important first step in testing a transdiagnostic blended model of care for youth psychosis and BPD in young people aged 15 to 25 years. We found that eOrygen was feasible, acceptable, and safe; there were indications that eOrygen may improve treatment outcomes if tested in a full-powered trial; and the majority of participants and clinicians reported positive experiences of using eOrygen as a blended model of care. However, some participants misunderstood the meaning of blended care, and future research should ensure that this is clearly outlined before integrating a digital tool into clinical practice. Furthermore, some clinicians reported a lack of knowledge and confidence in their ability to implement the intervention, and future research should aim to understand the possible barriers and address them to ensure clinician competence and confidence with the intervention itself. Overall, this pilot study provides promise for integrating blended models of care into specialized services for young people with complex mental health conditions, but a full-scale trial will be needed to test the effectiveness of such an intervention. This study also reaffirms prior findings indicating that blended models of care have the potential to increase therapeutic intensity, continuity, engagement, and effectiveness. Future research needs to focus on the development of tools to integrate blended care into practice specifically [36], as well as strategies to support both clinicians and young people in continuous use of the platform.

Acknowledgments

The authors would like to thank the young people who participated in the eOrygen pilot study and agreed to share their data for research purposes. The study was funded by generous support from the Telstra Foundation. MA-J was supported by an investigator grant (APP1177235) from the National Health and Medical Research Council and a Dame Kate Campbell Fellowship from The University of Melbourne. JN was supported by a National Health and Medical Research Council Emerging Leader Fellowship (ID 2009782).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Descriptive insights into the participants' willingness to use eOrygen again, whether they would recommend it to others, and their initial reasons of interest in using eOrygen.

[DOCX File , 15 KB - mental_v11i1e49217_app1.docx]

Multimedia Appendix 2

Young people's satisfaction with eOrygen (n=25).

[DOCX File , 13 KB - mental_v11i1e49217_app2.docx]

References

1. McGorry PD, Mei C, Chanen A, Hodges C, Alvarez-Jimenez M, Killackey E. Designing and scaling up integrated youth mental health care. World Psychiatry 2022 Feb 11;21(1):61-76 [FREE Full text] [doi: 10.1002/wps.20938] [Medline: 35015367]
2. Barrett S. From adult lunatic asylums to CAMHS community care: the evolution of specialist mental health care for children and adolescents 1948-2018. Fr J Br Stud 2019 Aug 30;XXIV(3):1-17 [FREE Full text] [doi: 10.4000/rfcb.4138]
3. Malla A, Iyer S, McGorry P, Cannon M, Coughlan H, Singh S, et al. From early intervention in psychosis to youth mental health reform: a review of the evolution and transformation of mental health services for young people. Soc Psychiatry Psychiatr Epidemiol 2016 Mar;51(3):319-326. [doi: 10.1007/s00127-015-1165-4] [Medline: 26687237]
4. Seidler ZE, Rice SM, Dhillon HM, Cotton SM, Telford NR, McEachran J, et al. Patterns of youth mental health service use and discontinuation: population data from Australia's headspace model of care. Psychiatr Serv 2020 Nov 01;71(11):1104-1113. [doi: 10.1176/appi.ps.201900491] [Medline: 32790590]

5. Robson E, Greenwood K. Rates and predictors of disengagement and strength of engagement for people with a first episode of psychosis using early intervention services: a systematic review of predictors and meta-analysis of disengagement rates. Schizophr Bull Open 2023 Jan;4(1):sgad033 [FREE Full text] [doi: 10.1093/schizbullopen/sgac012]
6. headspace centre young person follow up study. headspace. 2019. URL: https://headspace.org.au/assets/headspace-centre-young-person-follow-up-study-Sept-2019.PDF [accessed 2024-02-20]
7. Alvarez-Jimenez M, Priede A, Hetrick S, Bendall S, Killackey E, Parker A, et al. Risk factors for relapse following treatment for first episode psychosis: a systematic review and meta-analysis of longitudinal studies. Schizophr Res 2012 Aug;139(1-3):116-128. [doi: 10.1016/j.schres.2012.05.007] [Medline: 22658527]
8. Kennard BD, Emslie GJ, Mayes TL, Nakonezny PA, Jones JM, Foxwell AA, et al. Sequential treatment with fluoxetine and relapse--prevention CBT to improve outcomes in pediatric depression. Am J Psychiatry 2014 Oct;171(10):1083-1090 [FREE Full text] [doi: 10.1176/appi.ajp.2014.13111460] [Medline: 24935082]
9. Ascher-Svanum H, Zhu B, Faries DE, Salkever D, Slade EP, Peng X, et al. The cost of relapse and the predictors of relapse in the treatment of schizophrenia. BMC Psychiatry 2010 Jan 07;10:2 [FREE Full text] [doi: 10.1186/1471-244X-10-2] [Medline: 20059765]
10. Weisz JR, Kuppens S, Eckshtain D, Ugueto AM, Hawley KM, Jensen-Doss A. Performance of evidence-based youth psychotherapies compared with usual clinical care: a multilevel meta-analysis. JAMA Psychiatry 2013 Jul 01;70(7):750-761 [FREE Full text] [doi: 10.1001/jamapsychiatry.2013.1176] [Medline: 23754332]
11. Moses EB, Barlow DH. A new unified treatment approach for emotional disorders based on emotion science. Curr Dir Psychol Sci 2016 Jun 22;15(3):146-150. [doi: 10.1111/j.0963-7214.2006.00425.x]
12. Yang H, Gao S, Li J, Yu H, Xu J, Lin C, et al. Remission of symptoms is not equal to functional recovery: psychosocial functioning impairment in major depression. Front Psychiatry 2022 Jul 26;13:915689 [FREE Full text] [doi: 10.3389/fpsyt.2022.915689] [Medline: 35958633]
13. Andrews G, Issakidis C, Sanderson K, Corry J, Lapsley H. Utilising survey data to inform public policy: comparison of the cost-effectiveness of treatment of ten mental disorders. Br J Psychiatry 2004 Jun 02;184(6):526-533. [doi: 10.1192/bjp.184.6.526] [Medline: 15172947]
14. Torous J, Woodyatt J, Keshavan M, Tully LM. A new hope for early psychosis care: the evolving landscape of digital care tools. Br J Psychiatry 2019 May 11;214(5):269-272 [FREE Full text] [doi: 10.1192/bjp.2019.8] [Medline: 30739613]
15. Birnbaum ML, Rizvi AF, Confino J, Correll CU, Kane JM. Role of social media and the internet in pathways to care for adolescents and young adults with psychotic disorders and non-psychotic mood disorders. Early Interv Psychiatry 2017 Aug 23;11(4):290-295 [FREE Full text] [doi: 10.1111/eip.12237] [Medline: 25808317]
16. Alvarez-Jimenez M, Alcazar-Corcoles M, González-Blanch C, Bendall S, McGorry P, Gleeson J. Online, social media and mobile technologies for psychosis treatment: a systematic review on novel user-led interventions. Schizophr Res 2014 Jun;156(1):96-106. [doi: 10.1016/j.schres.2014.03.021] [Medline: 24746468]
17. Gottlieb JD, Romeo KH, Penn DL, Mueser KT, Chiko BP. Web-based cognitive-behavioral therapy for auditory hallucinations in persons with psychosis: a pilot study. Schizophr Res 2013 Apr;145(1-3):82-87. [doi: 10.1016/j.schres.2013.01.002] [Medline: 23410709]
18. Granholm E, Ben-Zeev D, Link PC, Bradshaw KR, Holden JL. Mobile Assessment and Treatment for Schizophrenia (MATS): a pilot trial of an interactive text-messaging intervention for medication adherence, socialization, and auditory hallucinations. Schizophr Bull 2012 May 10;38(3):414-425 [FREE Full text] [doi: 10.1093/schbul/sbr155] [Medline: 22080492]
19. Drews-Windeck E, Greenwood K, Cavanagh K. A systematic review and meta-analysis of digital interventions targeted at individuals with borderline personality disorder (BPD), emotionally unstable personality disorder (EUPD), and related symptoms. J Clin Psychol 2023 Sep 26;79(9):2155-2185. [doi: 10.1002/jclp.23523] [Medline: 37185891]
20. Christensen H, Batterham P, Calear A. Online interventions for anxiety disorders. Curr Opin Psychiatry 2014 Jan;27(1):7-13. [doi: 10.1097/YCO.0000000000000019] [Medline: 24257123]
21. Sethi S. Treating youth depression and anxiety: a randomised controlled trial examining the efficacy of computerised versus face - to - face cognitive behaviour therapy. Australian Psychologist 2020 Nov 12;48(4):249-257. [doi: 10.1111/ap.12006]
22. Hedman-Lagerlöf E, Carlbring P, Svärdman F, Riper H, Cuijpers P, Andersson G. Therapist-supported internet-based cognitive behaviour therapy yields similar effects as face-to-face therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. World Psychiatry 2023 Jun 09;22(2):305-314 [FREE Full text] [doi: 10.1002/wps.21088] [Medline: 37159350]
23. Firth J, Torous J, Nicholas J, Carney R, Pratap A, Rosenbaum S, et al. The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry 2017 Oct 21;16(3):287-298 [FREE Full text] [doi: 10.1002/wps.20472] [Medline: 28941113]
24. de Beurs D, van Bruinessen I, Noordman J, Friele R, van Dulmen S. Active involvement of end users when developing web-based mental health interventions. Front Psychiatry 2017 May 03;8:72 [FREE Full text] [doi: 10.3389/fpsyt.2017.00072] [Medline: 28515699]

25. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019 Jul 22;21(7):e12869 [FREE Full text] [doi: 10.2196/12869] [Medline: 31333198]
26. Moritz S, Schröder J, Meyer B, Hauschildt M. The more it is needed, the less it is wanted: attitudes toward face-to-face intervention among depressed patients undergoing online treatment. Depress Anxiety 2013 Feb 28;30(2):157-167. [doi: 10.1002/da.21988] [Medline: 22930656]
27. Carlbring P, Andersson G, Cuijpers P, Riper H, Hedman-Lagerlöf E. Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. Cogn Behav Ther 2018 Jan 07;47(1):1-18. [doi: 10.1080/16506073.2017.1401115] [Medline: 29215315]
28. Christensen H, Hickie IB. Using e-health applications to deliver new mental health services. Med J Aust 2010 Jun 07;192(S11):S53-S56. [doi: 10.5694/j.1326-5377.2010.tb03695.x] [Medline: 20528711]
29. Black M, Hitchcock C, Bevan A, O Leary C, Clarke J, Elliott R, et al. The HARMONIC trial: study protocol for a randomised controlled feasibility trial of Shaping Healthy Minds-a modular transdiagnostic intervention for mood, stressor-related and anxiety disorders in adults. BMJ Open 2018 Aug 05;8(8):e024546 [FREE Full text] [doi: 10.1136/bmjopen-2018-024546] [Medline: 30082367]
30. Torous J, Roberts LW. Needed innovation in digital health and smartphone applications for mental health: transparency and trust. JAMA Psychiatry 2017 May 01;74(5):437-438. [doi: 10.1001/jamapsychiatry.2017.0262] [Medline: 28384700]
31. Proudfoot J, Klein B, Barak A, Carlbring P, Cuijpers P, Lange A, et al. Establishing guidelines for executing and reporting internet intervention research. Cogn Behav Ther 2011 Jun;40(2):82-97. [doi: 10.1080/16506073.2011.573807] [Medline: 25155812]
32. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
33. Andersson G, Titov N. Advantages and limitations of internet-based interventions for common mental disorders. World Psychiatry 2014 Feb 04;13(1):4-11 [FREE Full text] [doi: 10.1002/wps.20083] [Medline: 24497236]
34. Donker T, Petrie K, Proudfoot J, Clarke J, Birch M, Christensen H. Smartphones for smarter delivery of mental health programs: a systematic review. J Med Internet Res 2013 Nov 15;15(11):e247 [FREE Full text] [doi: 10.2196/jmir.2791] [Medline: 24240579]
35. Erbe D, Eichert H, Riper H, Ebert DD. Blending face-to-face and internet-based interventions for the treatment of mental disorders in adults: systematic review. J Med Internet Res 2017 Sep 15;19(9):e306 [FREE Full text] [doi: 10.2196/jmir.6588] [Medline: 28916506]
36. Wentzel J, van der Vaart R, Bohlmeijer ET, van Gemert-Pijnen JE. Mixing online and face-to-face therapy: how to benefit from blended care in mental health care. JMIR Ment Health 2016 Feb 09;3(1):e9 [FREE Full text] [doi: 10.2196/mental.4534] [Medline: 26860537]
37. Kooistra LC, Wiersma JE, Ruwaard J, van Oppen P, Smit F, Lokkerbol J, et al. Blended vs. face-to-face cognitive behavioural treatment for major depression in specialized mental health care: study protocol of a randomized controlled cost-effectiveness trial. BMC Psychiatry 2014 Oct 18;14(1):290 [FREE Full text] [doi: 10.1186/s12888-014-0290-z] [Medline: 25326035]
38. Saxena S, Funk M, Chisholm D. Comprehensive mental health action plan 2013–2020. East Mediterr Health J 2015 Jul 01;12(7):461-463. [doi: 10.26719/2015.21.7.461]
39. Batterham PJ, Sunderland M, Calear AL, Davey CG, Christensen H, Teesson M, et al. Developing a roadmap for the translation of e-mental health services for depression. Aust N Z J Psychiatry 2015 Sep;49(9):776-784. [doi: 10.1177/0004867415582054] [Medline: 25907269]
40. Valentine L, McEnery C, Bell I, O'Sullivan S, Pryor I, Gleeson J, et al. Blended digital and face-to-face care for first-episode psychosis treatment in young people: qualitative study. JMIR Ment Health 2020 Jul 28;7(7):e18990 [FREE Full text] [doi: 10.2196/18990] [Medline: 32720904]
41. Alvarez-Jimenez M, Gleeson JF, Rice S, Gonzalez-Blanch C, Bendall S. Online peer-to-peer support in youth mental health: seizing the opportunity. Epidemiol Psychiatr Sci 2016 Feb 16;25(2):123-126. [doi: 10.1017/s2045796015001092]
42. Mohr DC, Riper H, Schueller SM. A solution-focused research approach to achieve an implementable revolution in digital mental health. JAMA Psychiatry 2018 Feb 01;75(2):113-114. [doi: 10.1001/jamapsychiatry.2017.3838] [Medline: 29238805]
43. Aref-Adib G, McCloud T, Ross J, O'Hanlon P, Appleton V, Rowe S, et al. Factors affecting implementation of digital health interventions for people with psychosis or bipolar disorder, and their family and friends: a systematic review. Lancet Psychiatry 2019 Mar;6(3):257-266. [doi: 10.1016/S2215-0366(18)30302-X] [Medline: 30522979]
44. Richards D, Richardson T, Timulak L, McElvaney J. The efficacy of internet-delivered treatment for generalized anxiety disorder: a systematic review and meta-analysis. Internet Interv 2015 Sep;2(3):272-282. [doi: 10.1016/j.invent.2015.07.003]
45. Weisel KK, Fuhrmann LM, Berking M, Baumeister H, Cuijpers P, Ebert DD. Standalone smartphone apps for mental health-a systematic review and meta-analysis. NPJ Digit Med 2019 Dec 02;2(1):118 [FREE Full text] [doi: 10.1038/s41746-019-0188-8] [Medline: 31815193]

46. Andrews G, Basu A, Cuijpers P, Craske M, McEvoy P, English C, et al. Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: an updated meta-analysis. J Anxiety Disord 2018 Apr;55:70-78 [FREE Full text] [doi: 10.1016/j.janxdis.2018.01.001] [Medline: 29422409]
47. Cuijpers P, Sijbrandij M, Koole SL, Andersson G, Beekman AT, Reynolds CF. The efficacy of psychotherapy and pharmacotherapy in treating depressive and anxiety disorders: a meta-analysis of direct comparisons. World Psychiatry 2013 Jun 04;12(2):137-148 [FREE Full text] [doi: 10.1002/wps.20038] [Medline: 23737423]
48. Sauer-Zavala S, Cassiello-Robbins C, Ametaj AA, Wilner JG, Pagan D. Transdiagnostic treatment personalization: the feasibility of ordering unified protocol modules according to patient strengths and weaknesses. Behav Modif 2019 Jul 10;43(4):518-543. [doi: 10.1177/0145445518774914] [Medline: 29742904]
49. Sauer-Zavala S, Gutner CA, Farchione TJ, Boettcher HT, Bullis JR, Barlow DH. Current definitions of "transdiagnostic" in treatment development: a search for consensus. Behav Ther 2017 Jan;48(1):128-138. [doi: 10.1016/j.beth.2016.09.004] [Medline: 28077216]
50. Newby JM, McKinnon A, Kuyken W, Gilbody S, Dalgleish T. Systematic review and meta-analysis of transdiagnostic psychological treatments for anxiety and depressive disorders in adulthood. Clin Psychol Rev 2015 Aug;40:91-110 [FREE Full text] [doi: 10.1016/j.cpr.2015.06.002] [Medline: 26094079]
51. Newby JM, Twomey C, Yuan Li SS, Andrews G. Transdiagnostic computerised cognitive behavioural therapy for depression and anxiety: a systematic review and meta-analysis. J Affect Disord 2016 Jul 15;199:30-41. [doi: 10.1016/j.jad.2016.03.018] [Medline: 27060430]
52. Eldridge SM, Lancaster GA, Campbell MJ, Thabane L, Hopewell S, Coleman CL, et al. Defining feasibility and pilot studies in preparation for randomised controlled trials: development of a conceptual framework. PLoS One 2016 Mar 15;11(3):e0150205 [FREE Full text] [doi: 10.1371/journal.pone.0150205] [Medline: 26978655]
53. Mohr DC, Lyon AR, Lattie EG, Reddy M, Schueller SM. Accelerating digital mental health research from early design and creation to successful implementation and sustainment. J Med Internet Res 2017 May 10;19(5):e153 [FREE Full text] [doi: 10.2196/jmir.7725] [Medline: 28490417]
54. Cotton SM, Filia KM, Ratheesh A, Pennell K, Goldstone S, McGorry PD. Early psychosis research at Orygen, the national centre of excellence in youth mental health. Soc Psychiatry Psychiatr Epidemiol 2016 Jan 23;51(1):1-13. [doi: 10.1007/s00127-015-1140-0] [Medline: 26498752]
55. Chanen AM, McCutcheon LK, Germano D, Nistico H, Jackson HJ, McGorry PD. The HYPE Clinic: an early intervention service for borderline personality disorder. J Psychiatr Pract 2009 May;15(3):163-172. [doi: 10.1097/01.pra.0000351876.51098.f0] [Medline: 19461389]
56. Hertzog MA. Considerations in determining sample size for pilot studies. Res Nurs Health 2008 Apr 08;31(2):180-191. [doi: 10.1002/nur.20247] [Medline: 18183564]
57. Alvarez-Jimenez M, Bendall S, Koval P, Rice S, Cagliarini D, Valentine L, et al. HORYZONS trial: protocol for a randomised controlled trial of a moderated online social therapy to maintain treatment effects from first-episode psychosis services. BMJ Open 2019 Feb 19;9(2):e024104 [FREE Full text] [doi: 10.1136/bmjopen-2018-024104] [Medline: 30782893]
58. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 07;4(1):50 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
59. Alvarez-Jimenez M, Rice S, D'Alfonso S, Leicester S, Bendall S, Pryor I, et al. A novel multimodal digital service (moderated online social therapy+) for help-seeking young people experiencing mental ill-health: pilot evaluation within a national youth e-mental health service. J Med Internet Res 2020 Aug 13;22(8):e17155 [FREE Full text] [doi: 10.2196/17155] [Medline: 32788151]
60. Alvarez-Jimenez M, Koval P, Schmaal L, Bendall S, O'Sullivan S, Cagliarini D, et al. The Horyzons project: a randomized controlled trial of a novel online social therapy to maintain treatment effects from specialist first-episode psychosis services. World Psychiatry 2021 Jun 18;20(2):233-243 [FREE Full text] [doi: 10.1002/wps.20858] [Medline: 34002511]
61. Rice S, Gleeson J, Davey C, Hetrick S, Parker A, Lederman R, et al. Moderated online social therapy for depression relapse prevention in young people: pilot study of a 'next generation' online intervention. Early Interv Psychiatry 2018 Aug 17;12(4):613-625. [doi: 10.1111/eip.12354] [Medline: 27311581]
62. Lederman R, Wadley G, Gleeson J, Bendall S, Álvarez-Jiménez M. Moderated online social therapy: designing and evaluating technology for mental health. ACM Trans Comput Hum Interact 2014 Feb;21(1):1-26. [doi: 10.1145/2513179]
63. Alvarez-Jimenez M, Gleeson J, Bendall S, Penn D, Yung A, Ryan R, et al. Enhancing social functioning in young people at Ultra High Risk (UHR) for psychosis: a pilot study of a novel strengths and mindfulness-based online social therapy. Schizophr Res 2018 Dec;202:369-377. [doi: 10.1016/j.schres.2018.07.022] [Medline: 30031616]
64. O'Sullivan S, Schmaal L, D'Alfonso S, Toenders YJ, Valentine L, McEnery C, et al. Characterizing use of a multicomponent digital intervention to predict treatment outcomes in first-episode psychosis: cluster analysis. JMIR Ment Health 2022 Apr 07;9(4):e29211 [FREE Full text] [doi: 10.2196/29211] [Medline: 35389351]
65. Engel L, Alvarez-Jimenez M, Cagliarini D, D'Alfonso S, Faller J, Valentine L, et al. The cost-effectiveness of a novel online social therapy to maintain treatment effects from first-episode psychosis services: results from the Horyzons randomized controlled trial. Schizophr Bull (Forthcoming) 2023 Jun 01. [doi: 10.1093/schbul/sbad071] [Medline: 37261464]

66. Gilchrist G, Landau S, Dheensa S, Henderson J, Johnson A, Love B, et al. The feasibility of delivering the ADVANCE digital intervention to reduce intimate partner abuse by men receiving substance use treatment: protocol for a non-randomised multi-centre feasibility study and embedded process evaluation. Pilot Feasibility Stud 2022 Jul 30;8(1):163 [FREE Full text] [doi: 10.1186/s40814-022-01116-x] [Medline: 35907900]
67. Rice S, Gleeson J, Leicester S, Bendall S, D'Alfonso S, Gilbertson T, et al. Implementation of the enhanced Moderated Online Social Therapy (MOST+) model within a national youth e-mental health service (eheadspace): protocol for a single group pilot study for help-seeking young people. JMIR Res Protoc 2018 Feb 22;7(2):e48 [FREE Full text] [doi: 10.2196/resprot.8813] [Medline: 29472177]
68. Bohus M, Kleindienst N, Limberger MF, Stieglitz R, Domsalla M, Chapman AL, et al. The short version of the Borderline Symptom List (BSL-23): development and initial data on psychometric properties. Psychopathology 2009 Nov 20;42(1):32-39. [doi: 10.1159/000173701] [Medline: 19023232]
69. Zanello A, Berthoud L, Ventura J, Merlo MC. The Brief Psychiatric Rating Scale (version 4.0) factorial structure and its sensitivity in the treatment of outpatients with unipolar depression. Psychiatry Res 2013 Dec 15;210(2):626-633 [FREE Full text] [doi: 10.1016/j.psychres.2013.07.001] [Medline: 23890713]
70. Goldman HH, Skodol AE, Lave TR. Revising axis V for DSM-IV: a review of measures of social functioning. Am J Psychiatry 1992 Sep;149(9):1148-1156. [doi: 10.1176/ajp.149.9.1148] [Medline: 1386964]
71. Hatcher RL, Lindqvist K, Falkenström F. Psychometric evaluation of the Working Alliance Inventory-Therapist version: current and new short forms. Psychother Res 2020 Jul 17;30(6):706-717. [doi: 10.1080/10503307.2019.1677964] [Medline: 31621525]
72. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
73. Munder T, Wilmers F, Leonhart R, Linster HW, Barth J. Working alliance inventory-short revised (WAI-SR): psychometric properties in outpatients and inpatients. Clin Psychol Psychother 2010 Dec 09;17(3):231-239. [doi: 10.1002/cpp.658] [Medline: 20013760]
74. Diener E, Wirtz D, Tov W, Kim-Prieto C, Choi D, Oishi S, et al. New well-being measures: short scales to assess flourishing and positive and negative feelings. Soc Indic Res 2009 May 28;97(2):143-156. [doi: 10.1007/s11205-009-9493-y]
75. Chen B, Vansteenkiste M, Beyers W, Boone L, Deci EL, Van der Kaap-Deeder J, et al. Basic psychological need satisfaction, need frustration, and need strength across four cultures. Motiv Emot 2014 Nov 12;39(2):216-236. [doi: 10.1007/s11031-014-9450-1]
76. Russell DW. UCLA Loneliness Scale (Version 3): reliability, validity, and factor structure. J Pers Assess 1996 Feb;66(1):20-40. [doi: 10.1207/s15327752jpa6601_2] [Medline: 8576833]
77. Hawthorne G. Measuring social isolation in older adults: development and initial validation of the friendship scale. Soc Indic Res 2006 Feb 13;77(3):521-548. [doi: 10.1007/s11205-005-7746-y]
78. Mattick RP, Clarke JC. Development and validation of measures of social phobia scrutiny fear and social interaction anxiety. Behav Res Ther 1998 Apr;36(4):455-470. [doi: 10.1016/s0005-7967(97)10031-6] [Medline: 9670605]
79. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385. [doi: 10.2307/2136404]
80. Andrews G, Slade T. Interpreting scores on the Kessler Psychological Distress Scale (K10). Aust N Z J Public Health 2001 Dec;25(6):494-497 [FREE Full text] [doi: 10.1111/j.1467-842x.2001.tb00310.x] [Medline: 11824981]
81. Palinkas LA, Horwitz SM, Chamberlain P, Hurlburt MS, Landsverk J. Mixed-methods designs in mental health services research: a review. Psychiatr Serv 2011 Mar;62(3):255-263. [doi: 10.1176/ps.62.3.pss6203_0255] [Medline: 21363896]
82. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
83. Damschroder LJ, Reardon CM, Opra Widerquist MA, Lowery J. Conceptualizing outcomes for use with the Consolidated Framework for Implementation Research (CFIR): the CFIR outcomes addendum. Implement Sci 2022 Jan 22;17(1):7 [FREE Full text] [doi: 10.1186/s13012-021-01181-5] [Medline: 35065675]
84. Fleming T, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018 Jun 06;20(6):e199 [FREE Full text] [doi: 10.2196/jmir.9275] [Medline: 29875089]
85. Alvarez-Jimenez M, Bendall S, Lederman R, Wadley G, Chinnery G, Vargas S, et al. On the HORYZON: moderated online social therapy for long-term recovery in first episode psychosis. Schizophr Res 2013 Jan;143(1):143-149. [doi: 10.1016/j.schres.2012.10.009] [Medline: 23146146]
86. Chanen AM. Borderline personality disorder in young people: are we there yet? J Clin Psychol 2015 Aug 20;71(8):778-791. [doi: 10.1002/jclp.22205] [Medline: 26192914]
87. Desrosiers L, Saint-Jean M, Laporte L, Lord M. Engagement complications of adolescents with borderline personality disorder: navigating through a zone of turbulence. Borderline Personal Disord Emot Dysregul 2020 Sep 01;7(1):18 [FREE Full text] [doi: 10.1186/s40479-020-00134-6] [Medline: 32884818]

88. De Panfilis C, Marchesi C, Cabrino C, Monici A, Politi V, Rossi M, et al. Patient factors predicting early dropout from psychiatric outpatient care for borderline personality disorder. Psychiatry Res 2012 Dec 30;200(2-3):422-429. [doi: 10.1016/j.psychres.2012.03.016] [Medline: 22503328]
89. Flückiger C, Del Re AC, Wampold BE, Horvath AO. The alliance in adult psychotherapy: a meta-analytic synthesis. Psychotherapy (Chic) 2018 Dec;55(4):316-340 [FREE Full text] [doi: 10.1037/pst0000172] [Medline: 29792475]
90. Karver MS, De Nadai AS, Monahan M, Shirk SR. Meta-analysis of the prospective relation between alliance and outcome in child and adolescent psychotherapy. Psychotherapy (Chic) 2018 Dec;55(4):341-355. [doi: 10.1037/pst0000176] [Medline: 30335449]
91. Mander J, Neubauer AB, Schlarb A, Teufel M, Bents H, Hautzinger M, et al. The therapeutic alliance in different mental disorders: a comparison of patients with depression, somatoform, and eating disorders. Psychol Psychother 2017 Dec 12;90(4):649-667. [doi: 10.1111/papt.12131] [Medline: 28497909]
92. Schuster R, Fichtenbauer I, Sparr VM, Berger T, Laireiter A. Feasibility of a blended group treatment (bGT) for major depression: uncontrolled interventional study in a university setting. BMJ Open 2018 Mar 12;8(3):e018412 [FREE Full text] [doi: 10.1136/bmjopen-2017-018412] [Medline: 29530905]
93. Cerga-Pashoja A, Doukani A, Gega L, Walke J, Araya R. Added value or added burden? A qualitative investigation of blending internet self-help with face-to-face cognitive behaviour therapy for depression. Psychother Res 2020 Nov 05;30(8):998-1010. [doi: 10.1080/10503307.2020.1720932] [Medline: 32024447]
94. Liverpool S, Mota CP, Sales CM, Čuš A, Carletto S, Hancheva C, et al. Engaging children and young people in digital mental health interventions: systematic review of modes of delivery, facilitators, and barriers. J Med Internet Res 2020 Jun 23;22(6):e16317 [FREE Full text] [doi: 10.2196/16317] [Medline: 32442160]
95. Tofthagen C. Threats to validity in retrospective studies. J Adv Pract Oncol 2012 May;3(3):181-183 [FREE Full text] [Medline: 25031944]
96. van Schaik DJ, Schotanus AY, Dozeman E, Huibers MJ, Cuijpers P, Donker T. Pilot study of blended-format interpersonal psychotherapy for major depressive disorder. Am J Psychother 2023 Feb 01;76(2):69-74. [doi: 10.1176/appi.psychotherapy.20210061] [Medline: 36541072]
97. Seidler ZE, Rice SM, Kealy D, Oliffe JL, Ogrodniczuk JS. What gets in the way? Men's perspectives of barriers to mental health services. Int J Soc Psychiatry 2020 Mar 06;66(2):105-110. [doi: 10.1177/0020764019886336] [Medline: 31692401]
98. Ellis LA, Collin P, Hurley PJ, Davenport TA, Burns JM, Hickie IB. Young men’s attitudes and behaviour in relation to mental health and technology: implications for the development of online mental health services. BMC Psychiatry 2013 Apr 20;13(1):119. [doi: 10.1186/1471-244x-13-119]

Abbreviations

BPD: borderline personality disorder
CFIR: Consolidated Framework for Implementation Research
DSM-5: Diagnostic and Statistical Manual of Mental Disorders
EPPIC: Early Psychosis Prevention and Intervention Centre
HYPE: Helping Young People Early
MOST: moderated online social therapy
TA: therapeutic alliance
YMH: youth mental health

Edited by J Torous; submitted 22.05.23; peer-reviewed by L Balcombe, X Zhang, M Cederberg; comments to author 29.06.23; revised version received 05.12.23; accepted 12.01.24; published 01.04.24.

Please cite as:
O'Sullivan S, McEnery C, Cagliarini D, Hinton JDX, Valentine L, Nicholas J, Chen NA, Castagnini E, Lester J, Kanellopoulos E, D'Alfonso S, Gleeson JF, Alvarez-Jimenez M
A Novel Blended Transdiagnostic Intervention (eOrygen) for Youth Psychosis and Borderline Personality Disorder: Uncontrolled Single-Group Pilot Study
JMIR Ment Health 2024;11:e49217
URL: https://mental.jmir.org/2024/1/e49217
doi:10.2196/49217
PMID:38557432

Original Paper

A Web-Based Intervention Using "Five Ways to Wellbeing" to Promote Well-Being and Mental Health: Randomized Controlled Trial

Monica Beer Prydz[1], MS; Nikolai Olavi Czajkowski[1,2], PhD; Maja Eilertsen[1,2], MD; Espen Røysamb[1,3], PhD; Ragnhild Bang Nes[1,2,4], PhD

[1]Promenta Research Centre, Department of Psychology, University of Oslo, Oslo, Norway
[2]Department of Mental Health and Suicide, Norwegian Institute of Public Health, Oslo, Norway
[3]Department of Childhood and Families, Norwegian Institute of Public Health, Oslo, Norway
[4]Department of Philosophy, Classics, and History of Arts and Ideas, University of Oslo, Oslo, Norway

Corresponding Author:
Monica Beer Prydz, MS
Promenta Research Centre
Department of Psychology
University of Oslo
Forskningsveien 3 a
PO Box 1020
Oslo, 0315
Norway
Phone: 47 95856400
Email: m.b.prydz@psykologi.uio.no

Abstract

Background: Compromised well-being and mental health problems pose a significant threat to individuals and societies worldwide. Resource-intensive psychological treatments alone cannot alleviate this burden. There is a need for low-cost, evidence-based interventions aimed at preventing illness and promoting well-being. Five activity domains appear to be linked with well-being promotion across populations: connecting with others, being active, taking notice, learning, and being generous/giving. The activities mentioned are part of the Five Ways to Wellbeing framework and the web-based intervention Five Ways to Wellbeing for All (5waysA).

Objective: This randomized controlled trial aims to test the effects of the 5waysA intervention, a web-based, low-cost, well-being–promoting measure targeting the general population. To date, the Five Ways to Wellbeing framework has not been tested in this specific format. The 5waysA intervention comprises 2 webinars and SMS text message reminders delivered over a 10-week period.

Methods: In 2021, a total of 969 study participants from various regions across Norway were openly recruited through a web page. They were then randomly assigned to either an intervention group or 1 of 2 waiting list control groups, namely, active or passive. Self-reported life satisfaction (Satisfaction With Life Scale [SWLS]), flourishing (Flourishing Scale [FS]), positive emotions, anxiety, and depression symptoms (Hopkins Symptom Checklist-8 [HSCL-8]) were assessed before the intervention, at 4 weeks into the intervention, and 1-2 weeks after the intervention (over 10 weeks). Data analysis was conducted using linear mixed (multilevel) models.

Results: After 10 weeks, 453 participants (171 in the intervention group and 282 in the waiting list control group) were assessed on outcome variables, with a dropout rate of 53.2% (516/969). Results revealed a significantly greater increase in the intervention group compared with the controls for SWLS (b=0.13, 95% CI 0.03-0.23; *P*=.001), FS (b=0.19, 95% CI 0.08-0.30; *P*=.001), positive emotions (b=0.43, 95% CI 0.27-0.60; *P*<.001), and these factors combined into a global well-being measure (b=0.28, CI 0.16-0.39; *P*<.001). Effect sizes (Cohen d) for the well-being outcomes ranged from 0.30 to 0.49. In addition, a significant decrease in anxiety and depressive symptoms was observed (b=–0.17, 95% CI –0.30 to –0.04; *P*=.001) with an effect size (Cohen d) of –0.20.

Conclusions: The findings suggest that the web-based 5waysA intervention could serve as an effective approach for enhancing well-being and mental health within the general population. This study offers individuals, policy makers, and local stakeholders an accessible and potentially cost-effective well-being intervention that could be easily implemented.

Trial Registration: ClinicalTrials.gov NCT04784871; https://clinicaltrials.gov/study/NCT04784871

(JMIR Ment Health 2024;11:e49050) doi:10.2196/49050

KEYWORDS

well-being; mental health promotion; intervention; web based; low cost; broad outreach; framework; web-based intervention; randomized controlled trial; effectiveness

Introduction

Background

The burden on society is significant due to compromised well-being and a sharp increase in mental health issues [1]. Longitudinal studies indicate that more than 80% of the population could meet the diagnostic criteria for a mental disorder at some point in their lifetime [2,3]. The early onset, widespread prevalence, presence of multiple health issues, and prolonged work absences as a result of associated disabilities significantly contribute to the substantial burden of mental health problems and low well-being [4]. Well-being is not simply the absence of mental illness [5]; rather, it encompasses positive elements such as joy and life satisfaction. However, the notably high negative correlation between well-being and mental illness implies a significant connection between them. Thus, gains in levels of well-being can possibly decrease the risk of future mental illness, and vice versa [6,7]. Remarkably, mental illness and compromised well-being can also impact our physical health [8], influencing cardiovascular and immune system functioning [9,10], hastening physical aging [11], and raising mortality rates [12]. Therefore, promoting well-being may enhance not only mental health [1] but also physical health. Evidence-based interventions aimed at promoting well-being, with wide-reaching accessibility and implemented before the onset of mental illness, have been shown to be cost-effective [13,14] and are strongly recommended. In this regard, internet-based interventions emerge as a promising and scalable approach for both treating [15] and preventing [16] mental disorders, as well as enhancing well-being [15,17-19].

Strategies for promoting well-being should target different societal levels and address various risk factors [20]. Poverty and social exclusion are notable risk factors for both compromised well-being and mental health issues, highlighting the necessity of adopting a public health approach and implementing universal interventions, such as political and structural measures [20]. However, the significance of structural strategies aimed at the general population is still not adequately recognized [21], and their implementation within the population is surprisingly limited [22]. In addition to universal interventions, there is a critical need for more targeted interventions that are easily accessible and can be seamlessly integrated into individuals' daily lives. Recent years have demonstrated that pandemics and related crises can profoundly impact health and well-being on a widespread scale. As a result, the significance of scalable interventions using adaptable delivery methods through technology has become increasingly apparent [23,24].

Well-Being and Mental Health

Numerous studies using various research designs have consistently demonstrated a close and often prospective association between well-being, mental health, and overall health outcomes, including longevity [8,12,25,26]. The concept and components of well-being have been extensively deliberated upon since ancient times [27]. Contemporary conceptualizations frequently incorporate 3 dimensions of well-being, encompassing how individuals (1) emotionally experience life, (2) cognitively evaluate life, and (3) find meaning and fulfillment in life (eudaimonia) [28,29]. Therefore, well-being entails both feeling good and functioning effectively, including the ability to cope with negative or distressing emotions. In summary, well-being and its components are multidimensional and closely intertwined with mental health, as defined by the World Health Organization (WHO) as "a state of well-being in which individuals realize their own abilities, can cope with the normal stresses of life, can work productively and fruitfully, and are able to make contributions to their communities" [30]. In addition, well-being is closely associated with fundamental psychological needs [31,32].

Well-being is frequently compromised when individuals endure prolonged periods of elevated negative emotions [33]. Similarly, psychopathology is acknowledged as a risk factor for diminished well-being. However, a dual model of mental health underscores that well-being and mental health are distinct, partially independent dimensions. Hence, experiencing high levels of well-being (such as flourishing) does not necessarily imply the absence of mental illness, and conversely, having a mental illness does not mean that joy, meaning, and positive emotions are absent from life [5]. However, the significant correlations between psychopathological symptoms (eg, depression and anxiety) and well-being [6,7] underscore the importance of addressing both well-being and mental health in interventions.

The connections between well-being and both physical and mental health appear to, at least partially, stem from the beneficial and protective effects of positive emotions on health and healthy behaviors [26,34]. Consequently, interventions aimed at promoting well-being and related psychological interventions frequently incorporate the enhancement of positive affect as a target outcome among others.

The Effectiveness of Interventions

Numerous meta-analyses indicate that various types of psychological interventions can enhance various aspects of well-being while concurrently reducing psychological symptoms, with average Cohen *d* effect sizes ranging between 0.29 and 0.62 [35-38]. Furthermore, lifestyle-oriented interventions that prioritize health behaviors such as physical activity and maintaining a balanced diet have shown promise in promoting well-being and positive mental health [39-41].

Psychological interventions can be costly and resource-intensive, particularly when they require individual face-to-face follow-up. In a health care system under economic strain, there is a need for more efficient solutions. Numerous studies and meta-analyses have shown that web-based interventions are effective alternatives [16,19,24,42-45]. In a recent meta-analysis focusing on positive psychology interventions in nonclinical populations, comparisons were made between interventions using technology-assisted and traditional (ie, face-to-face) methods. The findings indicated that face-to-face interventions were somewhat more effective in enhancing well-being compared with technology-assisted ones (Cohen d=0.32 vs 0.17, respectively) [23]. Indeed, research corroborates the effectiveness of stand-alone (ie, without human support) internet and mobile-based psychological interventions in treating various types of psychological issues and improving mental health [16,19,24]. Moreover, web-based psychological interventions have the potential to reach a larger number of individuals in need and are frequently less resource-intensive and costly [15,18]. The ongoing research on the web-based intervention Five Ways to Wellbeing for All (5waysA) seeks to evaluate the effectiveness of a stand-alone, cost-effective, internet-based intervention aimed at promoting well-being with a wide reach.

The Five Ways to Wellbeing Framework and The Five Ways to All Intervention

Overview

The Five Ways to Wellbeing (5ways) originated from a UK government initiative aimed at identifying 5 action domains crucial for promoting well-being and mental health. These action domains were to be evidence based and applicable to individuals of all ages [46]. The resulting 5 action domains are connecting (emphasizing the need for social relations, belongingness, and inclusion), being active (highlighting movement), taking notice (focusing on awareness), keep learning (emphasizing growth, development, and mastery), and giving (encouraging prosocial behavior and adding value). The following sections elaborate on the 5 action domains of the Five ways framework.

Connecting

Decades of psychological research have underscored the fundamental human need for attachment [47] and belonging [31,48,49], demonstrating that social relations are indispensable for well-being [50-52]. In a recent study, researchers used network analysis to investigate the connection between well-being and various environmental factors in a sample of 31,000 Norwegian adults. A strong association was observed between perceiving social relations as supportive and rewarding and higher levels of well-being, as well as fewer symptoms of depression and anxiety [53]. In addition, social relations appear to serve as a protective factor against stressors and are linked to mental resilience, even after accounting for genetic factors [54].

Be Active

Studies have demonstrated that activity and movement play a crucial role in well-being and can be enhanced through regular practice [41,55,56]. Physical activity has been shown to be effective in alleviating symptoms of depression, anxiety, and psychological distress, both in the general population and in groups with compromised health [57]. Indeed, a recent study using Mendelian randomization provides further evidence supporting a causal relationship between physical activity and the preventative effect on depression [39].

Taking Notice

Meta-analyses indicate that engaging in mindfulness practices, which involve being present and aware in the moment, is linked to various benefits for well-being [58-60]. These benefits include changes in brain regions associated with mood modulation [61], suggesting that mindfulness may have a positive impact on mental health. Furthermore, research suggests that awareness, as a skill, can be cultivated and strengthened through practice [32]. In a systematic review and meta-analysis examining the impact of various psychological interventions on well-being, mental, and physical health, mindfulness training and other psychological interventions had the most significant effect on both clinical and nonclinical populations [38].

Learning

Humans possess an inherent drive to develop competence, foster curiosity, strive for mastery in diverse life skills, and seek autonomy [62]. Learning and exposure to novelty are linked to the anticipation and sustained engagement of reward circuits in the brain, which typically have positive effects on well-being and serve as a buffer against stress [63,64]. Moreover, introducing novelty into familiar experiences in life appears to be an effective method for enhancing well-being [65,66].

Giving

Engaging in acts of kindness typically leads to increased happiness among individuals [67]. Furthermore, being generous and participating in informal giving are associated with various eudaimonic well-being benefits [68] and are considered essential for fulfilling our basic need for relatedness [31]. In addition, the need for mattering, which involves feeling valued and adding value, is emphasized as crucial for a fulfilling life [69]. A recent review of 15 preregistered studies on prosocial spending concluded that such acts not only benefit others but also increase one's own happiness [70].

Since its launch in 2008, the Five Ways to Wellbeing concept has gained traction in both policy and practice. It has been incorporated into national well-being surveys, school curricula, and local procurement decisions [71]. In 2011, the National Mental Health Development Unit and the National Association of Health Authorities and Trusts Confederation (UK) conducted a survey to assess the extent to which the Five Ways to Wellbeing had been used since its inception. The survey identified 76 instances of the Five Ways to Wellbeing framework being applied across various settings, organizations, and initiatives. However, it also underscored the necessity for rigorous evaluations [71]. Despite the evidence-based nature and widespread adoption of the 5 activity domains, there is a noticeable absence of high-quality preregistered studies demonstrating the effectiveness of these activities when combined. Furthermore, to the best of our knowledge, no one has empirically examined the Five Ways to Wellbeing as a web-based intervention. Research on the Five Ways to Wellbeing framework appears to be limited to a few studies. One previous well-powered cross-sectional study conducted in New Zealand reported positive associations between the 5ways activities and well-being [72]. Similarly, other cross-sectional surveys across different nations have documented a positive relationship [73]. In addition, 2 experimental studies have incorporated the 5ways concept into their research, albeit focusing on different angles [74] and targeting different groups (ie, in-patients) [75] compared with our study.

The web-based 5waysA intervention tested in this study is a multicomponent approach aimed at the nonclinical population. It offers individuals knowledge (well-being literacy), exercises, activities, and experiences tailored to build resources and fulfill basic psychological needs, thereby enhancing well-being. This 10-week intervention is resource effective and delivered via the web, incorporating 2 webinars and subsequent SMS text message reminders to support participants. Adopting a high-quality research design, we used a randomized controlled trial with waitlist controls, encompassing both an active and a passive control group.

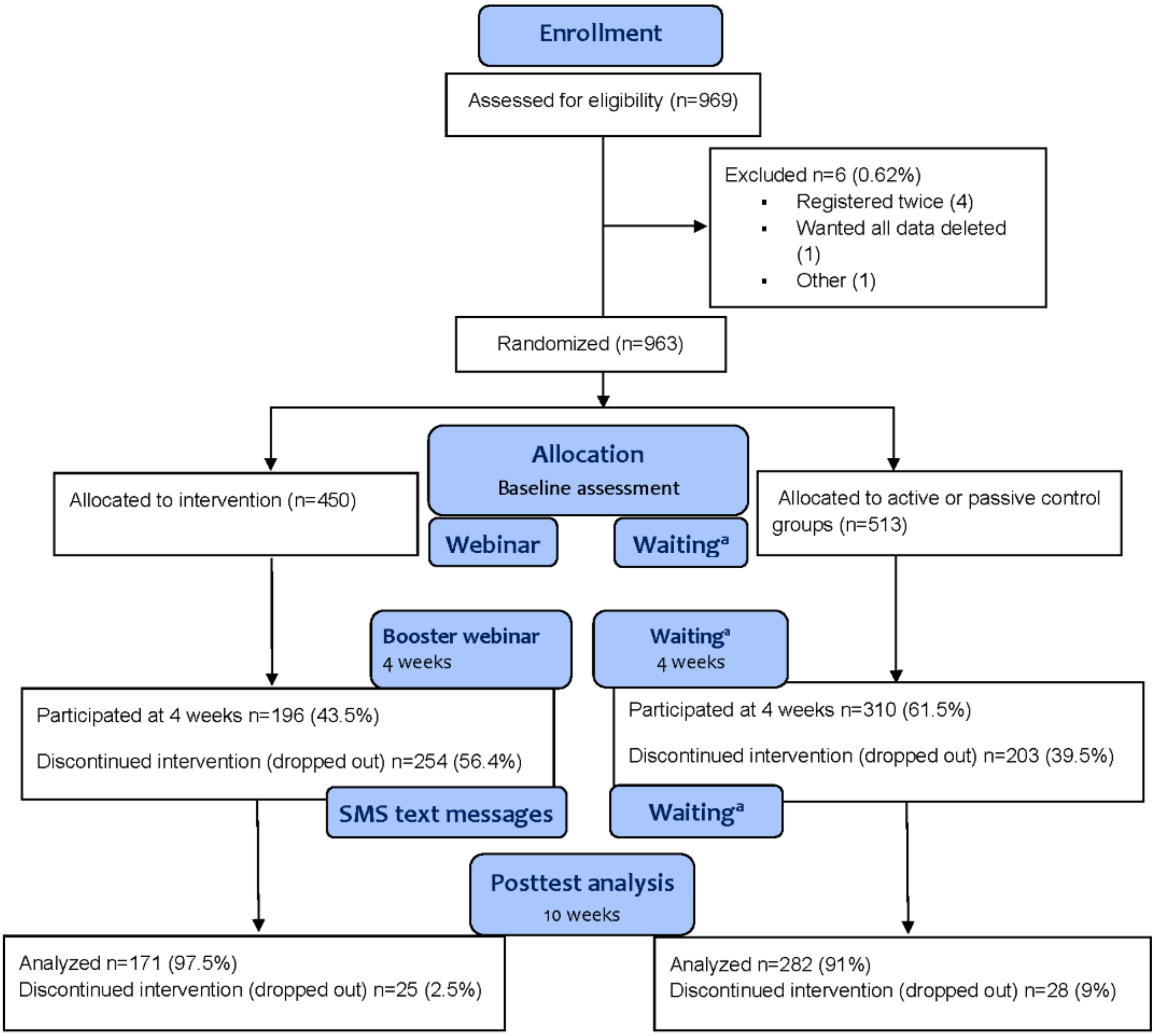
Methods

Design

This study adheres to a 2-armed randomized controlled trial design and follows the CONSORT (Consolidated Standards of Reporting Trials) guidelines and checklist (Multimedia Appendix 1 [76]) for transparent reporting of clinical trials [77,78]. Initially intending to recruit 1500 participants, recruitment was halted upon reaching a total of 969 participants to align with the project's time frame. Inclusion criteria were broad, with the only requirement being participants to be 18 years or older, while there were no specific exclusion criteria. All participants were recruited in 3 rounds from across Norway between February 2021 and November 2021. Our primary recruiting channels were the official social media platforms (ie, Facebook and Instagram) of local municipalities, the Norwegian Institute of Public Health, the Norwegian unit of the WHO European Healthy Cities Network, and a women's magazine (Kamille). The overall recruitment message invited individuals to participate in a study entitled "Hverdagsglede for alle" (Everyday Joy for All) aimed at learning about a health-promoting concept and related exercises to enhance well-being, health, and joy in their everyday lives. Participants were provided with information about the study and the randomization process. We explicitly outlined that 2 groups (the "active" group and the "passive control" group) would have to wait 3-5 months for the intervention, while 1 group (the intervention group) would receive the intervention promptly. Moreover, we openly communicated that 1 of the waiting list groups would receive a small registration exercise while awaiting the intervention. All participants were informed that they were automatically entered into a lottery for 5 gift cards (each worth €50 [US $54]). Data collection was conducted using "Nettskjema," a Norwegian standard tool for designing and administering online surveys developed by the University Information Technology Center (USIT) at the University of Oslo, Norway. This software is specifically designed for collecting highly sensitive research data, is integrated with the Services for Sensitive Data (TSD), and is known for its user-friendly interface [79].

The randomization was computer generated within the "Nettskjema" solution. As shown in Figure 1, 2 control conditions, namely, the "passive" and "active" control groups, were included to investigate potential differences across control conditions [80] and mitigate the risk of artificially inflating effect sizes [81,82]. The "passive" control group received no recommendations or exercises while waiting. The "active" control group was instructed to use an activity log containing a list of 10 everyday activities. They were asked to indicate whether they had completed these activities or not during the previous week. Participants in the active control group repeated this process for 10 weeks while waiting for the intervention.

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flow diagram. [a]Half of the control group wrote an activity log while waiting.



Power

The a priori power analysis was conducted using G*Power3 (The G*Power Team), assuming a small effect size (f=0.10) for the intervention, with an α level of .05. Even with a high dropout rate of 50%, our power of 0.999 would still be sufficient to detect a group × time interaction.

Participant Eligibility

All participants received an online letter of information about the study, which also outlined the requirements of The Regional Committees for Health Research Ethics (REK). Written consent was obtained from all participants in all groups through “Nettskjema,” along with 2-factor authentication (Multimedia Appendix 2). All sensitive personal information was managed in compliance with General Data Protection Regulation (GDPR) requirements. Data were stored and analyzed using the TSD at the University of Oslo.

Ethics Approval

The trial obtained ethical approval (reference number: 62155) from The Regional Committees for Health Research Ethics (REK).

The 5waysA Intervention

The 5waysA program is structured as a 10-week intervention comprising a 2-hour webinar, a 1-hour booster webinar (4 weeks later), and 10 SMS text message reminders (Table 1). In summary, the webinars offer a thorough introduction to how stressors in life impact well-being and health. In addition, participants are educated on how the 5ways activities can serve as a remedy for unhealthy stress, elevate positive emotions, and improve overall well-being. Following the 2 webinars, participants receive 2 SMS text messages weekly for 5 weeks. Each SMS text message contains a brief review text, a reminder of the 5ways activities, and a list of example exercises for practicing the activities. Moreover, they provide a time schedule, allowing participants to select a suitable time for practicing the activities (eg, “On my way to work,” “during the weekends when I’m at home”).

Table 1. Intervention schedule.

Measurement schedule	Intervention activity
Baseline assessment	
0-5 weeks before the intervention	Email with the questionnaire and a link to the main webinar
1	Main webinar (2 hours)
3	Email with a link to the booster webinar
4	Booster webinar (1 hour)
4 weeks' assessment	
5	Email with the questionnaire
5	SMS text messages 1-2: Take notice + exercise
6	SMS text messages 3-4: Be active + exercises
7	SMS text messages 5-6: Keep learning + exercises
8	SMS text messages 7-8: Connect + exercises
9	SMS text messages 9-10: Give + exercises
Posttest assessment	
10-12	Email with a link to the questionnaire

Procedure

The webinars were delivered as live lectures and facilitated by a medical doctor. Participants who were unable to attend received a recorded version of the webinars. Web-based questionnaires were distributed by the project manager, and participants were assessed at 3 time points: (1) baseline measurement at the time of enrollment, (2) measurement after the booster webinar 4 weeks later, and (3) measurement after completing the full 10-week intervention. The control groups were assessed at the same time points as the intervention group. While most participants completed the baseline assessment at least 5 weeks before the intervention and within 2 weeks of the end of the intervention (posttest), a small subset experienced delays or misunderstandings, resulting in completion outside this time frame. Nonetheless, they were still included in the study. On average, it took participants 10 minutes to complete the questionnaires.

Outcome Measures

To capture key aspects of well-being and mental health problems, we used various measures. Life satisfaction was assessed using the Satisfaction With Life Scale (SWLS), a well-established 5-item scale designed to gauge overall satisfaction with life [83]. The SWLS includes statements such as "I am satisfied with my life," with responses provided on a 7-point scale ranging from 1 (strongly disagree) to 7 (strongly agree). In our sample, Cronbach α for the SWLS was .90, indicating high internal consistency. Besides, we used the 8-item Flourishing Scale (FS), which measures eudaimonic aspects of well-being [84]. This scale has excellent psychometric properties and includes statements such as "I lead a purposeful and meaningful life," scored on the same 7-point scale as the SWLS. The estimated Cronbach α for the FS was .87, demonstrating strong internal reliability. In addition to the SWLS and FS, we used a set of questions from the National Quality of Life Survey to assess basic positive emotions (ie, happiness, engagement, calmness, and curiosity/interest) experienced over the past week. Participants rated each emotion on an 11-point scale ranging from 0 (not at all) to 10 (very much) [85]. The Cronbach α for the 4 positive emotions in our sample was .79, indicating good internal consistency. Notably, this set of emotional items is included in several large-scale national and regional surveys in Norway from 2019, with a total sample size exceeding 400,000. In addition to the individual scales, we aggregated the SWLS, FS, and positive emotion measures into a global well-being measure, which encompasses 3 crucial factors (cognitive, eudaimonic, and positive affect) in subjective well-being [86]. Furthermore, mental health was assessed using the Hopkins Symptom Checklist-8 (HSCL-8), an 8-item scale commonly used to gauge anxiety and depressive symptoms. The HSCL-8 includes questions such as whether participants feel "nervousness and shakiness inside" or "down or blue". We used a 5-point response scale ranging from 1 (not being bothered at all) to 5 (bothering me a lot) [87]. To ensure consistency across measures, 1 item concerning hopelessness about the future was excluded from the HSCL-8 scale, as it was partially covered by other measures. The Cronbach α for the resulting 7-item HSCL (HSCL-7) was calculated to be .90, indicating high internal consistency. Before analysis, each of the 5 outcome measures (SWLS, FS, positive emotion, global well-being, and HSCL-8) was standardized with respect to the first measurement point. Specifically, we subtracted the mean at baseline and divided by the SD at baseline.

Analytic Strategy

We used linear mixed (multilevel) models to investigate changes during the intervention period, using data from all assessments. Mixed modeling offers flexibility in handling missingness resulting from nonparticipation and dropout, as well as accounting for clustering caused by observations nested within individuals [88]. The time variable was coded as 0.0, 0.5, and 1.0 for baseline, booster, and posttest, respectively. The intervention group exhibited a greater increase in well-being and a decrease in mental health problems over time, as evidenced by a significant group × time interaction in the mixed models (α<.05). This interaction was interpreted as evidence for an intervention effect. Effect sizes were reported using Cohen *d* for all measures. In the primary analysis, we used an intention-to-treat approach, which included all participants who had completed more than 1 questionnaire within the intervention period. To assess the robustness of the results, we conducted additional analyses using only complete cases, which included participants who responded at all 3 time points (Multimedia Appendix 3). Furthermore, to explore the association between baseline levels and subsequent development, we fitted a model with a 3-way interaction between baseline level, time, and group (Multimedia Appendix 4).

Results

Characteristics of Participants

Table 2 presents the self-reported characteristics of the study participants at baseline. The majority of participants identified themselves as women (287/450, 63.8%), with a smaller proportion identifying as men (3 participants identified themselves as nonbinary). In addition, 393/450 (87.3%) participants reported having higher education (>high school). Chi-square analyses indicated no significant differences in age (*P*=.30), gender (*P*=.19), or education levels (*P*=.28) between the intervention and control groups. Across all groups, there was a moderate tendency for participants who dropped out at the posttest assessment to have lower SWLS scores at baseline.

Table 2. Participants characteristics (N=963).

Characteristics	Intervention group (n=450), n (%)	Control group (n=513), n (%)
Age (years)		
18-39	161 (35.8)	175 (34.1)
40-59	246 (54.7)	281 (54.8)
≥60	43 (9.6)	57 (11.1)
Gender		
Female	287 (63.8)	328 (63.9)
Male	160 (35.6)	185 (36.1)
Nonbinary	3 (0.7)	0 (0)
Education		
Primary school	6 (1.3)	6 (1.2)
High school	51 (11.3)	54 (10.5)
≤4 years of college or university	93 (20.7)	131 (25.5)
>4 years of college or university	300 (66.7)	322 (62.7)

Outcomes

While there was a small trend toward more positive development in the active control group compared with the passive control group, these differences were not statistically significant for any of the outcome variables (SWLS, *P*=.97; flourishing, *P*=.44; positive emotions, *P*=.76; global well-being, *P*=.70; and HSCL, *P*=.15). Therefore, in the main analysis, the 2 control conditions were combined to increase statistical power and simplify interpretation. As reported in Table 3, multilevel analysis revealed a significantly greater increase from baseline to posttest in the intervention group compared with the control group for various outcome measures. Specifically, for the SWLS, the increase was significant (b=0.13, 95% CI 0.03-0.23; *P*=.001). Similarly, for the FS, positive emotions, and global well-being, the increases were also significant (FS: b=0.19, 95% CI 0.08-0.30; *P*=.001; positive emotions: b=0.43, 95% CI 0.27-0.60; *P*<.001; and global well-being: b=0.28, 95% CI 0.16-0.39; *P*<.001). The effect sizes for the intervention effect, calculated as Cohen *d*, were as follows: 0.30 (95% CI 0.11-0.49) for the SWLS, 0.32 (95% CI 0.12-0.51) for the FS, 0.49 (95% CI 0.29-0.68) for positive emotions, and 0.48 (95% CI 0.29-0.68) for global well-being. All of these effect sizes were statistically significant at *P*<.05. Furthermore, the multilevel analysis for HSCL-7 revealed a greater decrease from baseline to posttest in the intervention group compared with the control group (b=–0.17, 95% CI –0.30 to –0.04). This finding underscores the effectiveness of the intervention in reducing mental health problems. The significance level for the decrease in the HSCL-7 score from baseline to posttest was *P*=.001, with a Cohen *d* of –0.20 (95% CI –0.40 to –0.01). The results for all the outcomes are graphically presented in Figure 2. The attrition rate from baseline to posttest was 53.2% (516/969). Across all groups, there was a moderate tendency for participants who dropped out at the posttest assessment to have lower SWLS scores at baseline. To further investigate the robustness of the data, we reran the analyses with only the participants who had completed all 3 measurements (baseline, 4 weeks, and posttest). The results from the complete case analysis remained significant for all outcome measures, indicating the robustness of the findings (see Tables S6-S10 in Multimedia Appendix 3 for more details). Although we found no significant 3-way interaction between baseline levels, time, and group for any of the well-being outcomes, a significant interaction was observed for symptoms of depression and anxiety. This suggests that individuals with higher symptoms at baseline experienced a greater decrease in symptoms following the intervention compared with those with low symptom levels (see Multimedia Appendix 4 for more details).

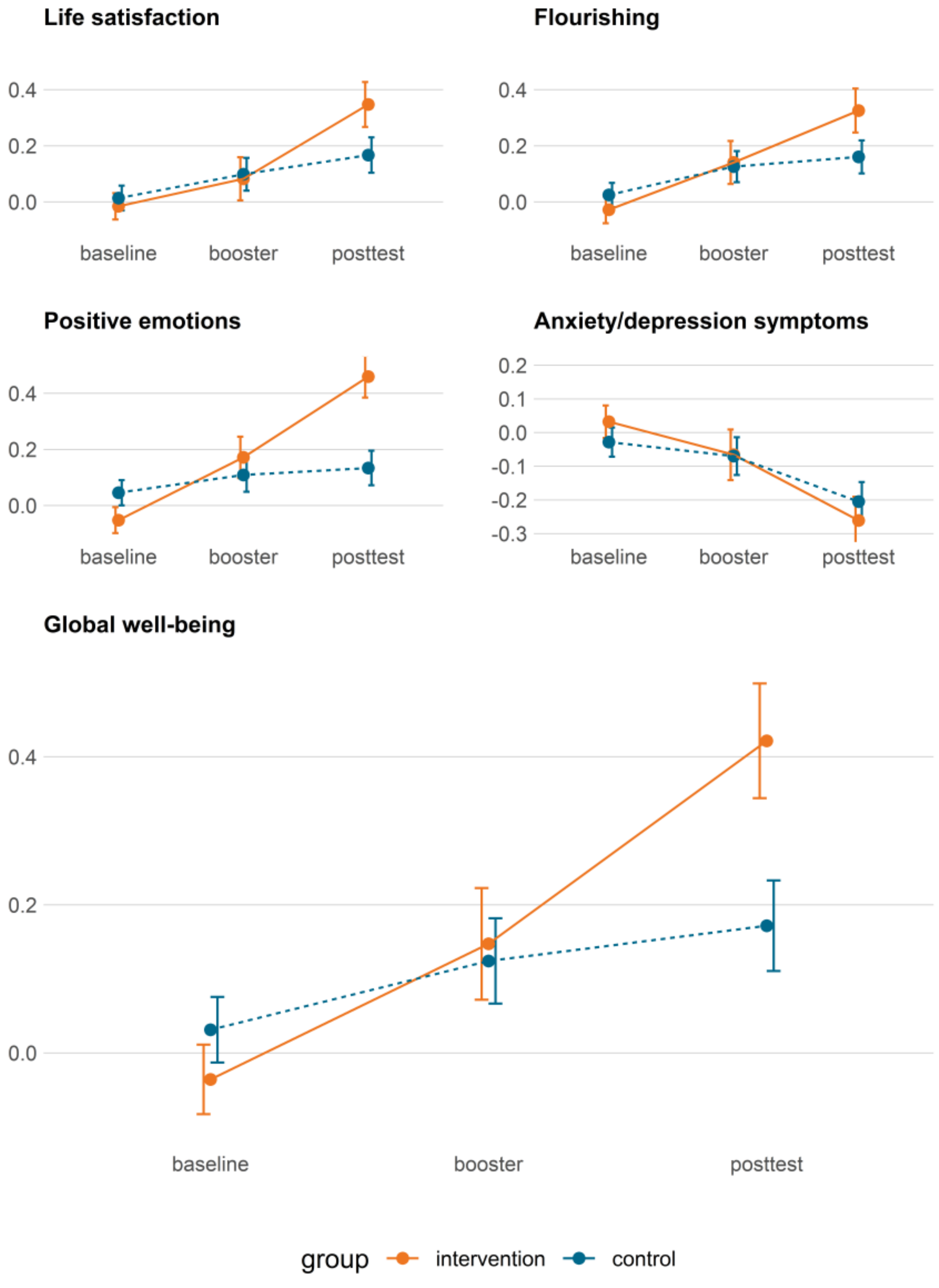
Table 3. Results from the multilevel regression model of the 5 outcome measures with the effect size Cohen d for the intervention effect (*P*<.05).

Outcome measure	Intervention group		Control group		Intervention effect	
	n	Mean (SD)	n	Mean (SD)	Estimates (95% CI)[a]	Cohen *d* (95% CI)[b]
Life satisfaction					0.13 (0.03 to 0.23)	0.30 (0.11 to 0.50)
Baseline	450	4.23 (1.28)	513	4.27 (1.27)		
Booster	196	4.36 (1.37)	310	4.38 (1.31)		
Posttest	171	4.69 (1.34)	282	4.46 (1.35)		
Flourishing					0.19 (0.08 to 0.30)	0.32 (0.12 to 0.51)
Baseline	450	5.38 (1.16)	513	5.44 (1.12)		
Booster	196	5.58 (1.22)	310	5.56 (1.11)		
Posttest	171	5.79 (1.17)	282	5.6 (1.13)		
Positive emotions					0.43 (0.27 to 0.60)	0.49 (0.29 to 0.68)
Baseline	450	26.16 (6.71)	513	26.83 (6.97)		
Booster	196	27.69 (7.07)	310	27.26 (7.18)		
Posttest	171	29.67 (6.7)	282	27.43 (7.10)		
Global well-being					0.28 (0.16 to 0.39)	0.48 (0.29 to 0.68)
Baseline	450	–0.04 (1)	513	0.03 (1)		
Booster	196	0.15 (1.06)	310	0.12 (1.01)		
Posttest	171	0.42 (1.01)	282	0.17 (1.03)		
Anxiety/depression symptoms					–0.17 (–0.30 to –0.04)	–0.20 (–0.40 to –0.01)
Baseline	450	2.22 (0.89)	513	2.17 (0.85)		
Booster	196	2.14 (0.92)	310	2.13 (0.86)		
Posttest	171	1.97 (0.80)	282	2.02 (0.85)		

[a]Estimates for the intervention effect from baseline to posttest

[b]Cohen *d* for the intervention effect from baseline to posttest.

Figure 2. Mean scores for the intervention group and the control group for life satisfaction, flourishing, positive emotions, the 3 well-being outcomes combined (global well-being), and depression and anxiety symptoms at baseline, 4 weeks into the intervention (booster), and at posttest 10 weeks after baseline. The bars indicate SEs.



Adherence

To assess adherence, participants were asked whether they watched the webinars in their entirety or partially, either live or recorded. Moreover, they were asked whether they engaged in the exercises sent to them via SMS text messages. Of the 171 participants in the intervention group, adherence was assessed in 134 participants, while 37 participants were not asked about this because of a delay in incorporating the relevant questions in the questionnaires. As we could not identify attendees or register them through the webinar platform, nor could we directly measure whether they received the SMS text messages, self-reporting became the sole method for assessing adherence. This realization emerged after the initial 37 participants had provided their responses, which is why they were excluded from this assessment. At the posttest, 117/134 (87.3%) of the participants reported having watched the entire main webinar, or parts of it, either live or recorded. In addition, 105/134 (78.3%) reported having watched the entire or parts of the booster webinar, either live or recorded. Furthermore, 126/134 (94.0%) reported engaging in some or all of the exercises sent to them in the weekly SMS text messages.

Discussion

Principal Findings

The findings from this study suggest that 5waysA is an effective web-based intervention for promoting well-being and mental health. This aligns with previous research indicating the effectiveness of internet-based interventions [16,19,24]. While the effect sizes may be small classified according to Cohen's *d*, they are still noteworthy and comparable to psychological intervention studies that detect small to moderate effects [35,38]. Moreover, the effect sizes observed in this study are comparable to those reported for psychotherapy or antidepressant medication in treating depression [89-91]. Indeed, meta-analyses have highlighted the importance of control conditions in intervention studies [81,82]. However, in this study, the distinction between passive and active control conditions did not significantly impact the observed effects of the 5waysA intervention. This underscores the effectiveness of the intervention itself rather than the specific type of control condition used. These positive outcomes align with existing research demonstrating that psychological interventions, regardless of their theoretical framework or delivery method, can effectively improve well-being and reduce symptoms of anxiety and depression [35-38]. This study provides additional support for the effectiveness of the Five Ways to Wellbeing framework in enhancing well-being and mental health among the general population. Previous examinations of the framework have primarily occurred in nonintervention settings [74] or with hospitalized patients [75], rather than as a web-based intervention aimed at the broader population.

The significance of our study lies in its ability to demonstrate effects across all core dimensions of subjective well-being—cognitive evaluations, emotional experiences, and eudaimonic aspects. This breadth of impact suggests relevance for individuals facing various struggles or experiencing a state of languishing within the multidimensional construct of well-being. Moreover, our findings indicate that participants with a higher number of symptoms associated with depression and anxiety were more responsive to the intervention than those with few symptoms, suggesting that the 5waysA intervention may also be beneficial for clinical or subclinical populations. The alignment between the 5 action domains in the Five Ways to Wellbeing framework and the basic psychological needs in self-determination theory, such as autonomy, relatedness, and competence, as well as the core dimensions of the training-based framework by Dahl et al [32], including awareness, connection, insight, and purpose, is noteworthy. These perspectives share the common goal of guiding individuals toward actions that yield observable and meaningful positive consequences for their overall well-being, health, and capacity to thrive and flourish.

The potential to guide individuals toward actions and promote cognitive and behavioral coping strategies is often considered a common mechanism of change across psychological interventions. In web-based interventions, this can be facilitated through reminders sent via emails, apps, or SMS text messages, promoting user self-reliance in a cost-effective and easily implemented manner. The strong emphasis on self-management and self-empowerment may be one of several mechanisms that contribute to making internet-based interventions comparable to on-site treatments and interventions. However, the active mechanisms of change in web-based interventions are still not fully understood.

Web-based interventions show promise in their ability to reach a broad audience. A recent review and meta-analysis [15] suggested that internet-based interventions for treating mental health problems are likely cost-effective, although the heterogeneity of studies makes generalizability challenging. In addition, prevention is known to be more cost-effective than treatment. The paradox of prevention [13] suggests that universal, broad health and well-being–promoting interventions aimed at the general population can be highly beneficial. Even modest improvements in many individuals can have a powerful impact at the population level. The 5waysA intervention, with its web-based delivery mode, has the potential to reach many people efficiently. It can be implemented through primary care, workplaces, municipal websites, high schools, universities, and other institutions.

Enhanced well-being can potentially impact both mental and physical health. The 5waysA intervention, as a transdiagnostic, multidimensional intervention, may represent a cost-effective measure capable of promoting overall well-being and mental health. Its low administration costs and web-based format make it suitable for an economically pressured health care service. This solution is also suitable for individuals who live in remote areas without easy access to health care services, are physically disabled, have busy schedules, or simply prefer web-based options (eg, young people). In addition, because this study was conducted during the COVID-19 outbreak, which had a significant impact on well-being, we anticipate that the intervention could also prove effective in future pandemics and lockdowns.

Strengths and Limitations

This study possesses several strengths, including its relatively large sample size, utilization of various control conditions, incorporation of validated outcome measures, and the use of items/scales enabling comparison with national surveys conducted concurrently. However, it also exhibits some limitations. As anticipated with web-based interventions [92], the attrition rate was relatively high (516/969, 53.2%). In terms of demographic variables, there were no systematic differences observed regarding the participants lost between groups. However, between the baseline and booster webinar, there was a higher dropout rate observed in the intervention group compared with the control group. The reason behind the increased dropout rate in the intervention group during this period remains unclear to us. The higher dropout rate in the intervention group during this period may be attributed to elevated expectations, such as potential disappointment following the webinar attendance, or reluctance to engage with the intervention if the webinar was missed, while the control group remained expectant. Future studies should delve deeper into strategies for preventing attrition in web-based interventions, exploring the intricate interplay among technological challenges, levels of support, and reminders, alongside individual user demographics [93,94].

The control condition in this study comprises both an active and a passive group, thus demonstrating a high-quality design. However, future studies should contemplate comparing the 5waysA with other web-based interventions within the multicomponent positive psychology interventions spectrum (eg, Chilver and Gatt [42]) to mitigate uncertainties surrounding its true effectiveness.

In addition, an important limitation of this study pertains to the short follow-up time. Future research should also explore potential mediating mechanisms. The representativeness of the sample is suboptimal, with 393/450 (87.3%) participants reporting a higher level of education, whereas the population mean is 36.9% [95]. However, the participants' mean baseline SWLS (sum) score of 21.25 was notably lower than the population mean of 25.21 [53], suggesting that our sample population might be at risk of transitioning toward compromised mental health or a state of languishing. Moreover, our results appear to indicate that they constitute a subset of the population that responds positively to well-being–promoting interventions. Their improved well-being may subsequently enhance their ability to manage work-life balance and caregiving responsibilities and shield them from detrimental stress. Indeed, future studies should evaluate the effectiveness of the 5waysA intervention in a more diverse sample, encompassing marginalized populations and specific illness groups. It might also be suitable to incorporate the Five Ways to Wellbeing concept into a population health approach, where interventions are less focused on individual behavior change and more integrated into local structures as societal measures and initiatives. This approach could ensure equal accessibility of the concept to various societal groups. Future research should assess the effectiveness of these types of Five Ways to Wellbeing measures.

Conclusions

Our results indicate that the web-based 5waysA intervention is an effective and promising intervention for promoting well-being among the general population. This study offers policy makers and local stakeholders a scalable, potentially cost-effective, and efficacious health promotion intervention with the potential for widespread impact. The 5waysA intervention targets the individual level, but its benefits can have meaning for individuals, groups, and societies as well.

Acknowledgments

First and foremost, we thank all the participants in this project. We also thank the user group who gave us feedback all the way from creating the intervention to the process of disseminating the results. In addition, we thank all the municipalities that helped us with the recruitment process, and especially the municipality of Bærum for letting us run a pilot study in 2020. Finally, we express our gratitude to the Norwegian Council of Mental Health for their support and collaboration throughout the entirety of the project. This project is funded by the Foundation Dam of Norway (grant 2021/FO347431). The University of Oslo applied for the funding through The Norwegian Council of Mental Health. The Dam Foundation of Norway has no role in designing the study, collecting data, analysis or interpretation of data, writing, or deciding to submit it for publication. This likewise applies to the Norwegian Council of Mental Health, except for its assistance in the recruitment of participants.

Data Availability

In accordance with ethical approval from The Regional Committees for Health Research Ethics (REK), we are currently not permitted to share data.

Conflicts of Interest

None declared.

Multimedia Appendix 1
CONSORT-EHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1195 KB - mental_v11i1e49050_app1.pdf]

Multimedia Appendix 2
Information letter, written consent , and questionnaires used in the study (in Norwegian).
[DOCX File , 14 KB - mental_v11i1e49050_app2.docx]

Multimedia Appendix 3
Results of the mixed models 1-5.
[DOCX File , 40 KB - mental_v11i1e49050_app3.docx]

Multimedia Appendix 4
Results of the mixed models 1-6.
[DOCX File , 32 KB - mental_v11i1e49050_app4.docx]

References

1. Patel V, Saxena S, Lund C, Thornicroft G, Baingana F, Bolton P, et al. The Lancet Commission on global mental health and sustainable development. Lancet 2018 Oct 27;392(10157):1553-1598. [doi: 10.1016/S0140-6736(18)31612-X] [Medline: 30314863]
2. Moffitt TE, Caspi A, Taylor A, Kokaua J, Milne BJ, Polanczyk G, et al. How common are common mental disorders? Evidence that lifetime prevalence rates are doubled by prospective retrospective ascertainment. Psychol. Med 2009 Sep 01;40(6):899-909. [doi: 10.1017/s0033291709991036]
3. Caspi A, Houts RM, Ambler A, Danese A, Elliott ML, Hariri A, et al. Longitudinal assessment of mental health disorders and comorbidities across 4 decades among participants in the Dunedin Birth Cohort Study. JAMA Netw Open 2020 Apr 01;3(4):e203221 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.3221] [Medline: 32315069]
4. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Psychiatry 2022 Feb;9(2):137-150 [FREE Full text] [doi: 10.1016/S2215-0366(21)00395-3] [Medline: 35026139]
5. Keyes C. The mental health continuum: from languishing to flourishing in life. Journal of Health and Social Behavior 2002 Jun;43(2):207-222 [FREE Full text] [doi: 10.2307/3090197]
6. Keyes CLM, Dhingra SS, Simoes EJ. Change in level of positive mental health as a predictor of future risk of mental illness. Am J Public Health 2010 Dec;100(12):2366-2371. [doi: 10.2105/ajph.2010.192245]
7. Lamers SM, Westerhof GJ, Glas CA, Bohlmeijer ET. The bidirectional relation between positive mental health and psychopathology in a longitudinal representative panel study. The Journal of Positive Psychology 2015 May 27;10(6):553-560. [doi: 10.1080/17439760.2015.1015156]
8. Firth J, Siddiqi N, Koyanagi A, Siskind D, Rosenbaum S, Galletly C, et al. The Lancet Psychiatry Commission: a blueprint for protecting physical health in people with mental illness. Lancet Psychiatry 2019 Aug;6(8):675-712. [doi: 10.1016/S2215-0366(19)30132-4] [Medline: 31324560]
9. Furman D, Campisi J, Verdin E, Carrera-Bastos P, Targ S, Franceschi C, et al. Chronic inflammation in the etiology of disease across the life span. Nat Med 2019 Dec;25(12):1822-1832 [FREE Full text] [doi: 10.1038/s41591-019-0675-0] [Medline: 31806905]
10. Guidi J, Lucente M, Sonino N, Fava GA. Allostatic load and its impact on health: a systematic review. Psychother Psychosom 2021;90(1):11-27 [FREE Full text] [doi: 10.1159/000510696] [Medline: 32799204]
11. Bourassa KJ, Moffitt TE, Ambler A, Hariri AR, Harrington H, Houts RM, et al. Association of treatable health conditions during adolescence with accelerated aging at midlife. JAMA Pediatr 2022 Apr 01;176(4):392-399 [FREE Full text] [doi: 10.1001/jamapediatrics.2021.6417] [Medline: 35188538]
12. Plana-Ripoll O, Pedersen CB, Agerbo E, Holtz Y, Erlangsen A, Canudas-Romo V, et al. A comprehensive analysis of mortality-related health metrics associated with mental disorders: a nationwide, register-based cohort study. The Lancet 2019 Nov;394(10211):1827-1835. [doi: 10.1016/s0140-6736(19)32316-5]
13. Rose G. Strategy of prevention: lessons from cardiovascular disease. Br Med J (Clin Res Ed) 1981 Jun 06;282(6279):1847-1851 [FREE Full text] [doi: 10.1136/bmj.282.6279.1847] [Medline: 6786649]
14. McDaid D, Park A, Wahlbeck K. The economic case for the prevention of mental illness. Annu Rev Public Health 2019 Apr 01;40(1):373-389. [doi: 10.1146/annurev-publhealth-040617-013629] [Medline: 30601725]
15. Rohrbach PJ, Dingemans AE, Evers C, Van Furth EF, Spinhoven P, Aardoom JJ, et al. Cost-effectiveness of internet interventions compared with treatment as usual for people with mental disorders: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2023 Jan 05;25:e38204 [FREE Full text] [doi: 10.2196/38204] [Medline: 36602854]
16. Sander L, Rausch L, Baumeister H. Effectiveness of internet-based interventions for the prevention of mental disorders: a systematic review and meta-analysis. JMIR Ment Health 2016 Aug 17;3(3):e38 [FREE Full text] [doi: 10.2196/mental.6061] [Medline: 27535468]
17. Renfrew ME, Morton DP, Morton JK, Hinze JS, Beamish PJ, Przybylko G, et al. A web- and mobile app-based mental health promotion intervention comparing email, short message service, and videoconferencing support for a healthy cohort: randomized comparative study. J Med Internet Res 2020 Jan 06;22(1):e15592 [FREE Full text] [doi: 10.2196/15592] [Medline: 31904578]
18. Ebert DD, Van Daele T, Nordgreen T, Karekla M, Compare A, Zarbo C, et al. Internet- and mobile-based psychological interventions: applications, efficacy, and potential for improving mental health. Eur Psychol 2018 Apr;23(2):167-187. [doi: 10.1027/1016-9040/a000318]
19. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019 Jul 22;21(7):e12869 [FREE Full text] [doi: 10.2196/12869] [Medline: 31333198]

20. Arango C, Díaz-Caneja CM, McGorry PD, Rapoport J, Sommer IE, Vorstman JA, et al. Preventive strategies for mental health. The Lancet Psychiatry 2018 Jul;5(7):591-604. [doi: 10.1016/s2215-0366(18)30057-9]
21. Enns J, Holmqvist M, Wener P, Halas G, Rothney J, Schultz A, et al. Mapping interventions that promote mental health in the general population: a scoping review of reviews. Prev Med 2016 Jun;87:70-80. [doi: 10.1016/j.ypmed.2016.02.022] [Medline: 26896634]
22. Campion J, Javed A, Lund C, Sartorius N, Saxena S, Marmot M, et al. Public mental health: required actions to address implementation failure in the context of COVID-19. Lancet Psychiatry 2022 Feb;9(2):169-182 [FREE Full text] [doi: 10.1016/S2215-0366(21)00199-1] [Medline: 35065723]
23. Koydemir S, Sökmez AB, Schütz A. A meta-analysis of the effectiveness of randomized controlled positive psychological interventions on subjective and psychological well-being. Applied Research Quality Life 2020 Jan 02;16(3):1145-1185. [doi: 10.1007/s11482-019-09788-z]
24. Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. World Psychiatry 2019 Oct 09;18(3):325-336. [doi: 10.1002/wps.20673] [Medline: 31496095]
25. Steptoe A, Deaton A, Stone AA. Subjective wellbeing, health, and ageing. Lancet 2015 Feb 14;385(9968):640-648 [FREE Full text] [doi: 10.1016/S0140-6736(13)61489-0] [Medline: 25468152]
26. Chida Y, Steptoe A. Positive psychological well-being and mortality: a quantitative review of prospective observational studies. Psychosom Med 2008 Sep;70(7):741-756. [doi: 10.1097/PSY.0b013e31818105ba] [Medline: 18725425]
27. Dodge R, Daly A, Huyton J, Sanders L. The challenge of defining wellbeing. Intnl J Wellbeing 2012 Aug 28;2(3):222-235. [doi: 10.5502/ijw.v2i3.4]
28. Diener E, Oishi S, Tay L. Advances in subjective well-being research. Nat Hum Behav 2018 Apr 12;2(4):253-260. [doi: 10.1038/s41562-018-0307-6] [Medline: 30936533]
29. Martela F, Ryan RM. Clarifying eudaimonia and psychological functioning to complement evaluative and experiential well-being: why basic psychological needs should be measured in national accounts of well-being. Perspect Psychol Sci 2023 Sep;18(5):1121-1135 [FREE Full text] [doi: 10.1177/17456916221141099] [Medline: 36626353]
30. World Health Organization (WHO). Mental health WHO. WHO. Geneva, Switzerland: World Health Organization; 2022 Jun 17. URL: https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response [accessed 2024-04-22]
31. Ryan R, Deci E. Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness. New York, NY: The Guilford Press; 2017.
32. Dahl CJ, Wilson-Mendenhall CD, Davidson RJ. The plasticity of well-being: a training-based framework for the cultivation of human flourishing. Proc Natl Acad Sci U S A 2020 Dec 22;117(51):32197-32206 [FREE Full text] [doi: 10.1073/pnas.2014859117] [Medline: 33288719]
33. Huppert FA. Psychological well‐being: evidence regarding its causes and consequences. Applied Psych Health Well 2009 Jun 05;1(2):137-164. [doi: 10.1111/j.1758-0854.2009.01008.x]
34. Pressman SD, Jenkins BN, Moskowitz JT. Positive affect and health: what do we know and where next should we go? Annu Rev Psychol 2019 Jan 04;70(1):627-650 [FREE Full text] [doi: 10.1146/annurev-psych-010418-102955] [Medline: 30260746]
35. Carr A, Cullen K, Keeney C, Canning C, Mooney O, Chinseallaigh E, et al. Effectiveness of positive psychology interventions: a systematic review and meta-analysis. The Journal of Positive Psychology 2020 Sep 10;16(6):749-769. [doi: 10.1080/17439760.2020.1818807]
36. Hendriks T, Schotanus-Dijkstra M, Hassankhan A, de Jong J, Bohlmeijer E. The efficacy of multi-component positive psychology interventions: a systematic review and meta-analysis of randomized controlled trials. J Happiness Stud 2019 Feb 11;21(1):357-390. [doi: 10.1007/s10902-019-00082-1]
37. Sin NL, Lyubomirsky S. Enhancing well-being and alleviating depressive symptoms with positive psychology interventions: a practice-friendly meta-analysis. J Clin Psychol 2009 May;65(5):467-487. [doi: 10.1002/jclp.20593] [Medline: 19301241]
38. van Agteren J, Iasiello M, Lo L, Bartholomaeus J, Kopsaftis Z, Carey M, et al. A systematic review and meta-analysis of psychological interventions to improve mental wellbeing. Nat Hum Behav 2021 May 19;5(5):631-652. [doi: 10.1038/s41562-021-01093-w] [Medline: 33875837]
39. Choi KW, Chen C, Stein MB, Klimentidis YC, Wang M, Koenen KC, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. Assessment of bidirectional relationships between physical activity and depression among adults: a 2-sample Mendelian randomization study. JAMA Psychiatry 2019 Apr 01;76(4):399-408. [doi: 10.1001/jamapsychiatry.2018.4175] [Medline: 30673066]
40. Maderuelo-Fernandez JA, Recio-Rodríguez JI, Patino-Alonso MC, Pérez-Arechaederra D, Rodriguez-Sanchez E, Gomez-Marcos MA, et al. Effectiveness of interventions applicable to primary health care settings to promote Mediterranean diet or healthy eating adherence in adults: a systematic review. Prev Med 2015 Jul;76 Suppl:S39-S55. [doi: 10.1016/j.ypmed.2014.12.011] [Medline: 25524613]

41. Firth J, Solmi M, Wootton RE, Vancampfort D, Schuch FB, Hoare E, et al. A meta-review of "lifestyle psychiatry": the role of exercise, smoking, diet and sleep in the prevention and treatment of mental disorders. World Psychiatry 2020 Oct 15;19(3):360-380 [FREE Full text] [doi: 10.1002/wps.20773] [Medline: 32931092]
42. Chilver MR, Gatt JM. Six-week online multi-component positive psychology intervention improves subjective wellbeing in young adults. J Happiness Stud 2022 Sep 05;23(3):1267-1288 [FREE Full text] [doi: 10.1007/s10902-021-00449-3] [Medline: 34512122]
43. Drozd F, Mork L, Nielsen B, Raeder S, Bjørkli CA. Better days – a randomized controlled trial of an internet-based positive psychology intervention. The Journal of Positive Psychology 2014 Apr 22;9(5):377-388. [doi: 10.1080/17439760.2014.910822]
44. Renfrew ME, Morton DP, Morton JK, Przybylko G. The influence of human support on the effectiveness of digital mental health promotion interventions for the general population. Front Psychol 2021 Aug 19;12:716106 [FREE Full text] [doi: 10.3389/fpsyg.2021.716106] [Medline: 34489818]
45. Schotanus-Dijkstra M, Drossaert CH, Pieterse ME, Boon B, Walburg JA, Bohlmeijer ET. An early intervention to promote well-being and flourishing and reduce anxiety and depression: a randomized controlled trial. Internet Interv 2017 Sep;9:15-24 [FREE Full text] [doi: 10.1016/j.invent.2017.04.002] [Medline: 30135833]
46. Aked J, Marks N, Cordon C, Thompson S. Five ways to well-being: communicating the evidence. New Economics Foundation. 2012 Aug 01. URL: https://neweconomics.org/2008/10/five-ways-to-wellbeing [accessed 2024-04-25]
47. Bowlby J. Attachment and Loss. Volume 1 Attachment. New York, NY: Basic Books; 1969.
48. Baumeister R, Leary M. The need to belong: desire for interpersonal attachments as a fundamental human motivation. Psychological Bulletin 1995;117(3):497-529. [doi: 10.1037//0033-2909.117.3.497]
49. Allen K, Gray D, Baumeister R, Leary M. The need to belong: a deep dive into the origins, implications, and future of a foundational construct. Educ Psychol Rev 2022;34(2):1133-1156 [FREE Full text] [doi: 10.1007/s10648-021-09633-6] [Medline: 34483627]
50. Helliwell JF, Aknin LB. Expanding the social science of happiness. Nat Hum Behav 2018 Apr;2(4):248-252. [doi: 10.1038/s41562-018-0308-5] [Medline: 30936535]
51. Jacques-Hamilton R, Sun J, Smillie LD. Costs and benefits of acting extraverted: a randomized controlled trial. J Exp Psychol Gen 2019 Sep;148(9):1538-1556. [doi: 10.1037/xge0000516] [Medline: 30489119]
52. Schroeder J, Lyons D, Epley N. Hello, stranger? Pleasant conversations are preceded by concerns about starting one. J Exp Psychol Gen 2022 May;151(5):1141-1153. [doi: 10.1037/xge0001118] [Medline: 34618536]
53. Bjørndal LD, Ebrahimi OV, Lan X, Nes RB, Røysamb E. Mental health and environmental factors in adults: a population-based network analysis. Am Psychol 2024 Apr;79(3):368-383. [doi: 10.1037/amp0001208] [Medline: 37439756]
54. Hofgaard LS, Nes RB, Røysamb E. Introducing two types of psychological resilience with partly unique genetic and environmental sources. Sci Rep 2021 Apr 21;11(1):8624 [FREE Full text] [doi: 10.1038/s41598-021-87581-5] [Medline: 33883571]
55. Sabe M, Chen C, Sentissi O, Deenik J, Vancampfort D, Firth J, et al. Thirty years of research on physical activity, mental health, and wellbeing: a scientometric analysis of hotspots and trends. Front Public Health 2022 Aug 9;10:943435 [FREE Full text] [doi: 10.3389/fpubh.2022.943435] [Medline: 36016904]
56. Lee I, Shiroma EJ, Lobelo F, Puska P, Blair SN, Katzmarzyk PT, Lancet Physical Activity Series Working Group. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. Lancet 2012 Jul 21;380(9838):219-229 [FREE Full text] [doi: 10.1016/S0140-6736(12)61031-9] [Medline: 22818936]
57. Singh B, Olds T, Curtis R, Dumuid D, Virgara R, Watson A, et al. Effectiveness of physical activity interventions for improving depression, anxiety and distress: an overview of systematic reviews. Br J Sports Med 2023 Sep;57(18):1203-1209 [FREE Full text] [doi: 10.1136/bjsports-2022-106195] [Medline: 36796860]
58. Querstret D, Morison L, Dickinson S, Cropley M, John M. Mindfulness-based stress reduction and mindfulness-based cognitive therapy for psychological health and well-being in nonclinical samples: a systematic review and meta-analysis. International Journal of Stress Management 2020 Nov;27(4):394-411. [doi: 10.1037/str0000165]
59. Gu J, Strauss C, Bond R, Cavanagh K. How do mindfulness-based cognitive therapy and mindfulness-based stress reduction improve mental health and wellbeing? A systematic review and meta-analysis of mediation studies. Clin Psychol Rev 2015 Apr;37:1-12. [doi: 10.1016/j.cpr.2015.01.006] [Medline: 25689576]
60. Chen B, Yang T, Xiao L, Xu C, Zhu C. Effects of mobile mindfulness meditation on the mental health of university students: systematic review and meta-analysis. J Med Internet Res 2023 Jan 03;25:e39128 [FREE Full text] [doi: 10.2196/39128] [Medline: 36596239]
61. Singleton O, Hölzel BK, Vangel M, Brach N, Carmody J, Lazar SW. Change in brainstem gray matter concentration following a mindfulness-based intervention is correlated with improvement in psychological well-being. Front Hum Neurosci 2014;8:33 [FREE Full text] [doi: 10.3389/fnhum.2014.00033] [Medline: 24600370]
62. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 2000 Jan;55(1):68-78. [doi: 10.1037//0003-066x.55.1.68] [Medline: 11392867]

63. Heller AS, van Reekum CM, Schaefer SM, Lapate RC, Radler BT, Ryff CD, et al. Sustained striatal activity predicts eudaimonic well-being and cortisol output. Psychol Sci 2013 Nov 01;24(11):2191-2200 [FREE Full text] [doi: 10.1177/0956797613490744] [Medline: 24058063]
64. Heller AS, Shi TC, Ezie CEC, Reneau TR, Baez LM, Gibbons CJ, et al. Association between real-world experiential diversity and positive affect relates to hippocampal-striatal functional connectivity. Nat Neurosci 2020 Jul 18;23(7):800-804. [doi: 10.1038/s41593-020-0636-4] [Medline: 32424287]
65. West C, Mogilner C, DeVoe SE. Happiness from treating the weekend like a vacation. Social Psychological and Personality Science 2020 Jun 15;12(3):346-356. [doi: 10.1177/1948550620916080]
66. O'Brien E, Smith RW. Unconventional consumption methods and enjoying things consumed: recapturing the "first-time" experience. Pers Soc Psychol Bull 2019 Jan;45(1):67-80. [doi: 10.1177/0146167218779823] [Medline: 29911504]
67. Curry OS, Rowland LA, Van Lissa CJ, Zlotowitz S, McAlaney J, Whitehouse H. Happy to help? A systematic review and meta-analysis of the effects of performing acts of kindness on the well-being of the actor. Journal of Experimental Social Psychology 2018 May;76:320-329. [doi: 10.1016/j.jesp.2018.02.014]
68. Hui BPH, Ng JCK, Berzaghi E, Cunningham-Amos LA, Kogan A. Rewards of kindness? A meta-analysis of the link between prosociality and well-being. Psychol Bull 2020 Dec;146(12):1084-1116. [doi: 10.1037/bul0000298] [Medline: 32881540]
69. Prilleltensky I. Mattering at the intersection of psychology, philosophy, and politics. Am J Community Psychol 2020 Mar;65(1-2):16-34. [doi: 10.1002/ajcp.12368] [Medline: 31407358]
70. Aknin LB, Dunn EW, Whillans AV. The emotional rewards of prosocial spending are robust and replicable in large samples. Curr Dir Psychol Sci 2022 Nov 09;31(6):536-545. [doi: 10.1177/09637214221121100]
71. Aked J, Thompson S. Five ways to well-being: new applications, new ways of thinking. New Economics Foundation. 2011. URL: https://neweconomics.org/2011/07/five-ways-well-new-applications-new-ways-thinking [accessed 2024-04-22]
72. Mackay L, Egli V, Booker L, Prendergast K. New Zealand’s engagement with the Five Ways to Wellbeing: evidence from a large cross-sectional survey. Kōtuitui: New Zealand Journal of Social Sciences Online 2019 Apr 28;14(2):230-244. [doi: 10.1080/1177083x.2019.1603165]
73. Harrison E, Quick A, Abdallah S, Ruggeri K, Garcia Garzon G, Maguire Á, et al. Looking through the wellbeing kaleidoscope: lessons from the European social survey. New Economics Foundation. 2016. URL: https://neweconomics.org/2016/05/looking-wellbeing-kaleidoscope [accessed 2024-04-22]
74. Coren E, Phillips J, Moore J, Brownett T, Whitfield L. An examination of the impacts of volunteering and community contribution at a community festival through the lens of the five ways to wellbeing. Int J Com WB 2021 Nov 29;5(1):137-156 [FREE Full text] [doi: 10.1007/s42413-021-00154-2] [Medline: 34870098]
75. Ng SSW, Leung TKS, Cheng EKN, Chan FSM, Chan JYH, Poon DF, et al. Efficacy of ‘Five Ways to Well-being Program’ in promotion of mental wellbeing for persons admitted to acute psychiatric service. J Psychosoc Rehabil Ment Health 2015 Dec 9;2(2):143-151. [doi: 10.1007/s40737-015-0039-0]
76. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
77. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, CONSORT. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. Int J Surg 2012;10(1):28-55 [FREE Full text] [doi: 10.1016/j.ijsu.2011.10.001] [Medline: 22036893]
78. Montgomery P, Grant S, Hopewell SG, Macdonald G, Moher D, Michie S, et al. Protocol for CONSORT-SPI: an extension for social and psychological interventions. Implement Sci 2013 Sep 02;8(99):99-97 [FREE Full text] [doi: 10.1186/1748-5908-8-99] [Medline: 24004579]
79. Nettskjema. University of Oslo. Oslo, Norway: University of Oslo; 2018. URL: https://www.uio.no/english/services/it/adm-services/nettskjema/about-nettskjema.html [accessed 2024-04-22]
80. Mohr DC, Spring B, Freedland KE, Beckner V, Arean P, Hollon SD, et al. The selection and design of control conditions for randomized controlled trials of psychological interventions. Psychother Psychosom 2009;78(5):275-284. [doi: 10.1159/000228248] [Medline: 19602916]
81. Gold SM, Enck P, Hasselmann H, Friede T, Hegerl U, Mohr DC, et al. Control conditions for randomised trials of behavioural interventions in psychiatry: a decision framework. Lancet Psychiatry 2017 Sep;4(9):725-732. [doi: 10.1016/S2215-0366(17)30153-0] [Medline: 28396067]
82. Cuijpers P, Cristea IA, Karyotaki E, Reijnders M, Huibers MJ. How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence. World Psychiatry 2016 Oct 22;15(3):245-258 [FREE Full text] [doi: 10.1002/wps.20346] [Medline: 27717254]
83. Pavot W, Diener E. The Satisfaction With Life Scale and the emerging construct of life satisfaction. The Journal of Positive Psychology 2008 Apr;3(2):137-152. [doi: 10.1080/17439760701756946]
84. Diener E, Wirtz D, Tov W, Kim-Prieto C, Choi D, Oishi S, et al. New well-being measures: short scales to assess flourishing and positive and negative feelings. Soc Indic Res 2009 May 28;97(2):143-156. [doi: 10.1007/s11205-009-9493-y]

85. Nes R, Nilsen T, Hauge L, Eilertsen M, Gustavson K, Aarø L. Fra Nord til Sør: Livskvalitet i Norge 2019. [Quality of Life in Norway. From North to South]. In: Folkehelseinstituttet. Oslo, Norway: Folkehelseinstituttet [Norwegian Institute of Public Health]; 2020.
86. Bjørndal LD, Nes R, Czajkowski N, Røysamb E. The structure of well-being: a single underlying factor with genetic and environmental influences. Qual Life Res 2023 Oct;32(10):2805-2816 [FREE Full text] [doi: 10.1007/s11136-023-03437-7] [Medline: 37209357]
87. Tambs K, Moum T. How well can a few questionnaire items indicate anxiety and depression? Acta Psychiatr Scand 1993 May;87(5):364-367. [doi: 10.1111/j.1600-0447.1993.tb03388.x] [Medline: 8517178]
88. Hox J. In: Moerbeek M, van de Schoot R, editors. Multilevel Analysis? Techniques and Applications (3rd Edition). New York, NY: Routledge; 2017.
89. Cipriani A, Furukawa TA, Salanti G, Chaimani A, Atkinson LZ, Ogawa Y, et al. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. Lancet 2018 Apr 07;391(10128):1357-1366 [FREE Full text] [doi: 10.1016/S0140-6736(17)32802-7] [Medline: 29477251]
90. Cuijpers P, van Straten A, Bohlmeijer E, Hollon SD, Andersson G. The effects of psychotherapy for adult depression are overestimated: a meta-analysis of study quality and effect size. Psychol Med 2010 Feb;40(2):211-223. [doi: 10.1017/S0033291709006114] [Medline: 19490745]
91. Cuijpers P. Four decades of outcome research on psychotherapies for adult depression: an overview of a series of meta-analyses. Canadian Psychology/Psychologie canadienne 2017 Feb;58(1):7-19. [doi: 10.1037/cap0000096]
92. Christensen H, Griffiths KM, Farrer L. Adherence in internet interventions for anxiety and depression. J Med Internet Res 2009 Apr 24;11(2):e13 [FREE Full text] [doi: 10.2196/jmir.1194] [Medline: 19403466]
93. Ryan C, Bergin M, Wells JS. Theoretical perspectives of adherence to web-based interventions: a scoping review. Int J Behav Med 2018 Feb;25(1):17-29. [doi: 10.1007/s12529-017-9678-8] [Medline: 28730402]
94. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
95. Statistics Norway. Educational attainment of the population. Oslo, Norway: Statistics Norway; 2023. URL: https://www.ssb.no/en/utdanning/utdanningsniva/statistikk/befolkningens-utdanningsniva [accessed 2024-04-22]

Abbreviations

5waysA: Five Ways to Wellbeing for All
CONSORT: Consolidated Standards of Reporting Trials
FS: Flourishing Scale
GDPR: General Data Protection Regulation
HSCL-7: 7-item Hopkins Symptom Checklist
HSCL-8: 8-item Hopkins Symptom Checklist
REK: The Regional Committees for Health Research Ethics
SWLS: Satisfaction With Life Scale
TSD: Services for Sensitive Data
USIT: University Information Technology Center
WHO: World Health Organization

Edited by J Torous; submitted 21.05.23; peer-reviewed by R Virgara, F Huppert, J Moskowitz, L Bolier; comments to author 07.09.23; revised version received 15.10.23; accepted 09.03.24; published 20.05.24.

Please cite as:
Prydz MB, Czajkowski NO, Eilertsen M, Røysamb E, Nes RB
A Web-Based Intervention Using "Five Ways to Wellbeing" to Promote Well-Being and Mental Health: Randomized Controlled Trial
JMIR Ment Health 2024;11:e49050
URL: https://mental.jmir.org/2024/1/e49050
doi:10.2196/49050
PMID:38767958

Original Paper

Empowering Mental Health Monitoring Using a Macro-Micro Personalization Framework for Multimodal-Multitask Learning: Descriptive Study

Meishu Song[1], MSc; Zijiang Yang[1], MSc; Andreas Triantafyllopoulos[2], MSc; Zixing Zhang[3], PhD; Zhe Nan[4], MSc; Muxuan Tang[1], MSc; Hiroki Takeuchi[1], PhD; Toru Nakamura[5], PhD; Akifumi Kishi[1], PhD; Tetsuro Ishizawa[1], MD, PhD; Kazuhiro Yoshiuchi[1], MD, PhD; Björn Schuller[6], PhD; Yoshiharu Yamamoto[1], PhD

[1]Educational Physiology Laboratory, Graduate School of Education, The University of Tokyo, Tokyo, Japan
[2]Technical University of Munich, Munich, Germany
[3]Hunan University, Changsha, China
[4]Google, San Carlos, CA, United States
[5]Osaka University, Osaka, Japan
[6]Imperial College London, London, United Kingdom

Corresponding Author:
Yoshiharu Yamamoto, PhD
Educational Physiology Laboratory
Graduate School of Education
The University of Tokyo
7 Chome-3-1 Hongo
Bunkyo City
Tokyo, 1130033
Japan
Phone: 81 09035784363
Email: yamamoto@p.u-tokyo.ac.jp

Abstract

Background: The field of mental health technology presently has significant gaps that need addressing, particularly in the domain of daily monitoring and personalized assessments. Current noninvasive devices such as wristbands and smartphones are capable of collecting a wide range of data, which has not yet been fully used for mental health monitoring.

Objective: This study aims to introduce a novel dataset for personalized daily mental health monitoring and a new macro-micro framework. This framework is designed to use multimodal and multitask learning strategies for improved personalization and prediction of emotional states in individuals.

Methods: Data were collected from 298 individuals using wristbands and smartphones, capturing physiological signals, speech data, and self-annotated emotional states. The proposed framework combines macro-level emotion transformer embeddings with micro-level personalization layers specific to each user. It also introduces a Dynamic Restrained Uncertainty Weighting method to effectively integrate various data types for a balanced representation of emotional states. Several fusion techniques, personalization strategies, and multitask learning approaches were explored.

Results: The proposed framework was evaluated using the concordance correlation coefficient, resulting in a score of 0.503. This result demonstrates the framework's efficacy in predicting emotional states.

Conclusions: The study concludes that the proposed multimodal and multitask learning framework, which leverages transformer-based techniques and dynamic task weighting strategies, is superior for the personalized monitoring of mental health. The study indicates the potential of transforming daily mental health monitoring into a more personalized app, opening up new avenues for technology-based mental health interventions.

(JMIR Ment Health 2024;11:e59512) doi:10.2196/59512

KEYWORDS

multimodal; multitask; daily mental health; mental health; monitoring; macro; micro; framework; personalization; strategies; prediction; emotional state; wristbands; smartphone; mobile phones; physiological; signals; speech data;

Introduction

Background

Mental health, recognized as a critical component of overall well-being, has garnered increasing attention and concern. The World Health Organization [1] defines mental health as a state of well-being where individuals realize their potential, cope with normal life stresses, work productively, and contribute to their community. However, mental health issues continue to present a significant burden globally, affecting individuals' quality of life and posing challenges. In response to these challenges, the concept of "daily mental health monitoring" has emerged as a critical area of research and application [2,3]. This concept refers to the regular, continuous observation and assessment of an individual's emotional states, using a variety of methods and tools to capture data in real time [4]. Such monitoring aims to provide a comprehensive understanding of an individual's mental health, facilitating early detection of patterns, changes, or emerging issues. Consequently, accurate monitoring and understanding of daily mental health have become imperative for timely interventions and sustained mental well-being.

However, the field of daily mental health monitoring remains surprisingly underdeveloped, particularly regarding real-life applications [4]. The challenges faced by existing datasets and methods are not merely academic concerns but represent significant barriers to the effective and widespread adoption of mental health monitoring in everyday life. These challenges include the following.

- Real-world representation where a significant portion of existing datasets lack data derived from real-world settings, instead relying on artificial or laboratory conditions.
- Lack of self-annotation. many datasets do not use self-annotation [5-8], relying instead on experts' observation or clinical interpretation. This approach often fails to capture the subjective experience of the individual, crucial for a person-centered understanding and monitoring of mental health [4,9]. In addition, clinical assessments typically occur at discrete time points, potentially missing the dynamic, moment-to-moment fluctuations in mental states that individuals experience in their daily lives [10].
- Challenges in accessibility of monitoring data. Many studies use electroencephalography (EEG) [11,12] while providing valuable insights into brain activity and emotional states, requiring specialized equipment and expertise, making it impractical for daily monitoring. Similarly, facial expression data [13,14] capture often necessitates continuous video monitoring, posing substantial privacy and practicality challenges for everyday use.
- Limited modalities and single-model approach. Most available research focuses on a single modality [15,16], this overlooks the inherently multimodal nature of human emotional expression and mental states, reducing the systems' reliability.

To address the aforementioned challenges, our study adopts an innovative methodology aimed at forging more accurate, and efficacious tools for mental health monitoring. The contributions of our research are manifold, highlighted by the following key developments.

- Introducing a novel dataset collected from 298 individuals using noninvasive, everyday devices including wristband-type devices and smartphones, our dataset captures physiological signals: zero crossing mode (ZCM), proportional integration mode (PIM), and speech data. Participants provided self-annotated emotional states over 2 weeks, creating a rich, multimodal resource for understanding daily mental health dynamics.
- Developing a macro-micro framework for personalized daily mental health. Our framework develops a multimodal and multitask learning (MTL) strategy, innovatively built global emotion embeddings with individual personalization embedding.

In our research, the decision to focus on physiological signals and speech data, while excluding modalities such as facial expressions and EEG, was driven by several key considerations: Physiological signals have been shown to have a significant association with mental health and well-being. These signals, such as heart rate, skin conductance, and activity levels, can provide valuable insights into an individual's emotional and psychological state [17]. The relationship between physiological signals and mental health is complex and multifaceted. For example, changes in heart rate variability (HRV) have been linked to stress, anxiety, and depression [18]. Reduced HRV has been observed in individuals with mental health disorders, suggesting that it may serve as a potential biomarker for mental well-being [19].

In our research, we apply the wrist-worn device used in our study which is equipped with a highly sensitive piezoelectric accelerometer that can detect even the most subtle wrist movements, with a resolution as fine as 0.01 G/rad/s. This allows for the capture of a wide range of daily activities and movements that may be relevant to mental health assessment [20]. The device uses 2 key modes for processing the accelerometer data: ZCM and PIM [21].

In addition to physiological signals, our study also incorporates speech data as a key modality for assessing mental health. Speech provides a rich source of information about an individual's emotional state, cognitive functioning, and overall well-being [22]. There are several reasons why speech is a valuable tool for mental health assessment. First, speech carries emotional information through various features such as tone, pitch, and intonation. Changes in these features can reflect an individual's emotional state, such as increased monotonicity in speech being associated with depression [23]. Second, speech patterns and characteristics can provide insights into an individual's cognitive processes. For example, changes in speech fluency, coherence, and word choice have been linked to cognitive impairments and mental health conditions [24]. In addition, speech data can be collected noninvasively using readily available devices such as smartphones or voice recorders. This makes it a convenient and accessible modality for mental health assessment, especially in remote or telehealth settings [25].

Ecological momentary assessment (EMA) is a key methodological approach used in our study. EMA involves the repeated sampling of individuals' current behaviors and experiences in real time, in their natural environments [26]. While EMA offers several advantages, such as reducing recall bias and capturing the dynamics of mental states in real-world contexts [10], it also has limitations. These include potential reactivity (ie, the act of self-reporting influencing the very experiences being reported) and compliance issues [27]. In this study, we aim to mitigate these limitations through careful design and participant training, which will be discussed in the Methods section.

Furthermore, our study introduces the Dynamic Restrained Uncertainty Weighting (DRUW) fusion method, a novel approach for integrating multimodal data in the context of mental health monitoring. The DRUW fusion method adaptively weights the contribution of each modality based on its uncertainty and distinct characteristics, ensuring a balanced representation of the fused data. This method builds upon the principles of uncertainty weighting [28] and extends them to the multimodal fusion context. The key novelty of the DRUW fusion method lies in its ability to dynamically adjust the weighting of each modality based on the inherent uncertainty and complementary nature of the physiological signals and speech data [29].

By collecting and analyzing data on emotional states, speech characteristics, and physiological patterns, our study aims to contribute to the development of more effective, personalized, and accessible mental health interventions. The data collected in our study can contribute to better mental health outcomes in several ways.

Early Detection and Intervention

By correlating objective measures with subjective emotional states, we can develop tools for early detection of mental health issues, enabling timely interventions [30].

Personalized Treatment and Monitoring

Insights from our study can inform personalized treatment plans and monitoring strategies, tailoring interventions to individual needs [31].

Remote Monitoring and Telemedicine

Our use of wearable devices and speech analysis can contribute to remote monitoring tools, crucial for mental health support, especially in light of recent global events [32].

Reducing Stigma and Increasing Access

By demonstrating objective measures for mental health assessment, we can potentially reduce stigma and increase access to care, particularly for underserved populations [33].

Related Work

This section delves into various aspects of mental health monitoring research domain, including mental health data, EMA, personalization, and their real-world implications.

Recent research efforts, particularly in mental health detection and monitoring, have gained significant momentum. Key studies like the systematic review by Hickey et al [34] and Long et al [35] have critically evaluated the use of smart devices and wearable technologies. These investigations underline the capability of these devices in detecting stress, anxiety, and depression through physiological measures such as HRV, electrodermal activity, and EEG data. However, they also identify a notable gap in the availability of commercial depression-detecting devices, emphasizing the need for integrating multimodal data to enhance both accuracy and predictive power.

Recent advancements in multimodal data analysis have shown promising results in mental health diagnosis [36]. For instance, a study by Xu et al [36] proposed a measurement method for mental health based on dynamic multimodal feature recognition. This approach integrates various data sources, including physiological signals, speech patterns, and behavioral indicators, to provide a more comprehensive assessment of an individual's mental state. Similarly, Huckins et al [37] developed a multimodal machine learning approach that combines smartphone sensing data with self-reported mental health scores to predict changes in depression and anxiety among college students.

Building on this, the role of mental health datasets becomes crucial in understanding the complex and varied nature of mental health conditions across different populations. The comprehensive analysis of datasets, such as those examined during the COVID-19 pandemic [38], offers deep insights into the mental health effects of global crises on specific demographics, like the Bangladeshi population. These datasets are instrumental not only in assessing the prevalence and severity of mental health conditions across various groups but also in supporting longitudinal studies vital for tracking changes over time.

Recent studies have also focused on improving data collection methods for mental health monitoring. For example, Morshed et al [39] introduced a novel approach using passive sensing and machine learning to predict mood instability in bipolar disorder. This method leverages smartphone usage patterns and environmental data to provide continuous, unobtrusive monitoring of mental health states.

In this study, we apply EMA [40], which represents a method for recording participants' behavior, psychological state, and physical symptoms in real-time and at multiple time points. The primary advantage of EMA lies in its ability to minimize the biases often associated with retrospective recall in self-report data. Traditional self-report measures, which ask participants to remember and report past feelings, behaviors, or symptoms, can be influenced by memory distortions and subjective interpretations of past events, Thus, it reduces the likelihood of recall errors and increases the accuracy and reliability of the data collected.

Further, personalization in mental health monitoring systems is increasingly important [4]. Innovations in digital phenotyping [41] exemplify this trend. This is further advanced by groundbreaking approaches like those proposed by Gerczuk et al [42], using zero-shot personalization strategies for large speech foundation models in mood recognition.

The application of artificial intelligence and deep learning techniques in mental health monitoring has seen significant growth. A comprehensive review by Su et al [43] highlights the potential of deep learning models in analyzing multimodal data for mental health assessment. These advanced techniques allow for more nuanced interpretation of complex, high-dimensional data, potentially leading to more accurate and personalized mental health interventions.

In summary, the related work shows the dynamic nature of mental health detection and monitoring. However, bridging the gap between technological capabilities and personalized mental health care presents numerous challenges. The integration of multimodal data, advancements in data collection methods, and the application of sophisticated artificial intelligence techniques represent promising avenues for overcoming these challenges and improving mental health monitoring and diagnosis.

Methods

Methodology and Data Collection

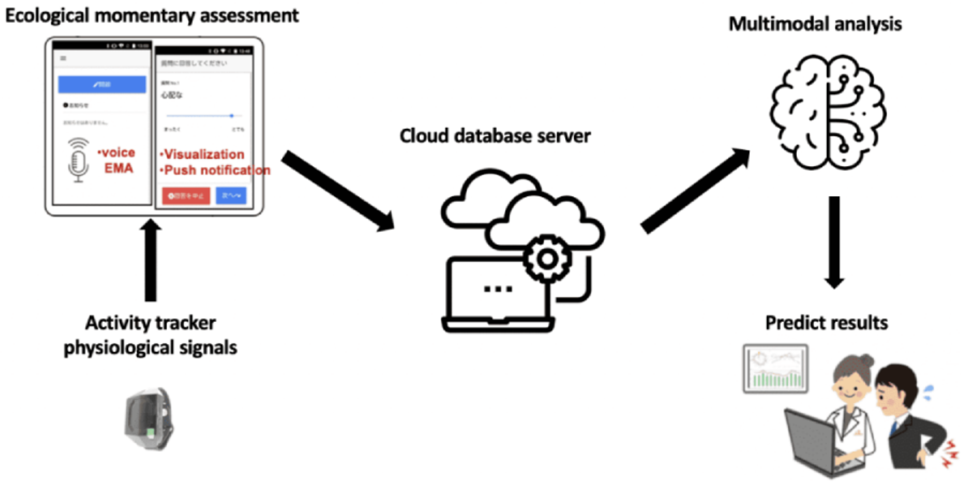
Overview

The study followed a 2-week data collection protocol, during which participants wore wrist-worn devices and used a smartphone app to record their speech and self-report their emotional states. The collected data included physiological signals from the wrist-worn devices and speech recordings from the smartphone app. To collect data, we developed a platform called Mental Healthcare Internet of Things (MHIT) system.

The MHIT System

The MHIT system, as shown in Figure 1, is a cloud-based platform specifically crafted to gather and analyze data from Internet of Things devices. This state-of-the-art system combines the collection of physical activity signals with speech data. The MHIT system is comprised of 2 key components: a cloud server (MHIT server) and a smartphone app (MHIT app).

Figure 1. The MHIT system designed for data collection. MHIT: Mental Healthcare Internet of Things.



Participants

A convenience sample of 298 Japanese office workers participated in our study. They were recruited by sending digital flyers. Those who agreed to participate in the study were asked to open a URL link on the flyer and complete a web-based registration form. Subsequently, the in-house wrist-worn device (ScienceNet device; ScienceNet Inc), survey progression guide, and informed consent form were mailed to them. Participants were instructed on the aim and procedure of the study through the survey progression guide. In addition, by scanning a QR code on the guide, they were able to watch a tutorial video to learn the use of the device and the MHIT app. After completing the informed consent form, they commenced the study. This recruitment process was conducted fully on the web, which contribute to achieve data collection from workers in different residential locations.

Annotation and EMA

This study used an EMA paradigm to capture participants' emotional states in real time, thus avoiding potential distortions of retrospective recall in self-report data. The EMA protocol involved the following.

Sampling Scheme

Participants were prompted to report their emotional states by using the MHIT app at randomly selected times within –10 minutes to +10 minutes of predetermined times (11 AM, 3 PM, and 7 PM). In addition, they were instructed to voluntarily complete the same EMA questionnaires when they woke up and went to bed.

Data Collection

At each EMA, participants self-reported the intensity of 9 different expressed emotions on a (0:100) visual-analogue scale (slider) displayed on the MHIT app. These 9 emotions correspond to the items of the Depression and Anxiety Mood Scale (DAMS) [44]. To prevent response bias and predetermination, the order of the DAMS items was randomized for each evaluation.

Speech Data

Before the mental state evaluation, participants recorded their voices by speaking, for example, "the current date and time are September 5, 2022, at 10:23 PM" on the MHIT app. The reasoning behind this is to keep the content emotionally neutral. Participants also recorded activities, and the actual time was recorded by the system.

Physiological Signals

The instrument is fitted with a sensitive piezoelectric accelerometer that detects minute wrist accelerations (as fine as 0.01 G/rad/s), capturing even the most subtle daily movements. The ZCM within the device tallies the instances the accelerometer's signal traverses the 0 mark over a predefined duration, known as the epoch time. Conversely, the PIM assesses the integral of the root-mean-square for the triaxial accelerometer signals. For the purposes of this investigation, we have configured the epoch interval at 1 minute, aggregating 60 data points (representing 1 hour) prior to each participant's DAMS entry within their routine activities. To ensure the integrity of our dataset, we have meticulously curated instances that comprise both ZCM and PIM recordings, each consisting of 60 data points.

Annotation Scheme

The DAMS serves as a self-reported measure of an individual's emotional state, providing a subjective assessment of their mental well-being. This scale, which encompasses nine distinct emotions—vigorous, gloomy, concerned, happy, unpleasant, anxious, cheerful, depressed, and worried—is used for comprehensively assessing mental health experiences pertinent to depression and anxiety. DAMS's effectiveness in measuring depressive and anxious moods is particularly notable, as it uses a variety of descriptors, including adjectives, adjectival verbs, and phrases, to delineate depressive, anxious, and positive moods with high discriminant validity [44]. Moreover, its psychometric soundness has been established through methods such as parallel testing and test-retest evaluations [44], confirming its high convergent, discriminant validity, and reliability. The scale's sensitivity to mood fluctuations is evidenced by the variance in scores observed between normal and stressful periods, underscoring its use in detecting mood changes. These features of DAMS, combined with its thorough statistical analysis across 9 emotional labels, confirmed it is a comprehensive choice for our study.

Previous research has shown that speech characteristics and patterns can reflect an individual's emotional state. For example, depression has been associated with changes in prosody, such as reduced pitch variability and slower speaking rate [22]. Similarly, anxiety has been linked to increased vocal tension and higher fundamental frequency [45]. By analyzing speech features such as pitch, intonation, and speaking rate, we can potentially identify objective markers that correlate with the subjective emotional states reported through DAMS.

Physiological data, collected through wrist-worn accelerometers, can also provide an indirect measure of an individual's emotional state. Studies have demonstrated that mood disorders, such as depression and anxiety, can influence an individual's activity levels and patterns [46]. Depression, for instance, has been associated with reduced physical activity and increased sedentary behavior [47]. By examining the activity data captured by the accelerometers, we can explore potential correlations between the objective measures of physical activity and the subjective emotional states assessed by DAMS.

While speech and physical activity data do not directly measure the emotional states captured by DAMS, they can offer complementary and objective insights into an individual's mental well-being. By combining these different modalities—self-reported mood, objective speech characteristics, and objective physiological signal patterns—we aim to develop a more comprehensive understanding of an individual's mental health status.

Self-Annotation

Self-annotation is a cornerstone in daily mental health monitoring for important reasons such as capturing subjective emotional experience and ecological validity. Using the MHIT app, participants self-reported their emotional states 5 times daily over 2 weeks, using the 9 emotional states outlined in DAMS.

Capturing Subjective Emotional Experiences

Emotions are inherently subjective, and self-annotation allows individuals to express their emotional states based on personal experiences. This method ensures an authentic portrayal of their mental state, which is crucial for accurate mental health assessment.

Ecological Validity

By self-reporting in real-time within their usual environments, participants provide data that more accurately reflect their day-to-day emotional experiences, enhancing the ecological validity of our study.

Data Preprocessing

In our research, the preprocessing of collected data was a critical step for both speech and physical activity. This process involved several stages, each tailored to the specific nature of the data being processed.

Preprocessing of Speech Data

For speech data, audio files were standardized in terms of their sampling rate and format for subsequent analysis.

Data Cleansing

Any recordings that were unsuccessful or contained data anomalies were removed. This step was crucial to ensure the integrity and quality of the speech dataset.

Voice Activity Detection

We used algorithms to detect and eliminate silences in voice recordings. This focus on active speech segments helped in isolating meaningful data.

Denoising

Background noise within the recordings was reduced using digital signal processing techniques. While the specific method may vary depending on the characteristics of the noise and the recording environment, common approaches include spectral subtraction, Wiener filtering, or more advanced techniques such as deep learning-based noise suppression algorithms [48].

Preprocessing of Physical Activity Data

The preprocessing of physical activity data includes the following approaches.

Signal Cleaning

Similar to speech data, physical activity data were cleaned to remove any erroneous signals.

Signal Standardization

The raw data from the physical activity sensors were standardized to ensure consistency across different participants.

Normalization Process

The intensity ratings of the emotional states reported by participants were normalized to a uniform scale ranging from 0 to 1.

Multimodal Multitask Analysis

In transitioning from data collection to the analysis of daily mental health in our study, we shift our focus toward developing a robust multimodal multitask analysis framework. The initial step involves defining the analytical task, which in our case is predicting various mental health indicators as outlined by DAMS. To effectively achieve our goals in daily mental health monitoring, we need to address three pivotal questions.

- How to fuse different modalities: specifically, how do we integrate physical activities and speech data?
- How to achieve personalization: what strategies can we use to tailor the analysis to individual participants?
- How to balance different emotional states: how can we ensure that our analysis provides a balanced view of various emotional states?

To respond to these questions, our approach involves the introduction of a comprehensive framework architecture. We plan to detail each component of this framework, starting with multimodal fusion, then moving on to personalization, and concluding with multitask balancing. This sequence is carefully chosen by allowing multimodal fusion to initially integrate and align different data types (physiological signals and speech) after feature extraction, creating a whole picture of data for further analysis. Personalization subsequently adapts this integrated data to individual differences, ensuring the model accurately represents each participant's unique mental health profile. The final stage, multitask balancing, refines the network to efficiently manage multiple analytical tasks.

Framework

The proposed framework commences with a robust feature extraction phase. For physiological signals, PIM and ZCM are input into individual 2-layer feed-forward neural networks (FFNNs). Concurrently, speech signals are preprocessed through a specialized *wav2vec-l-emo* model [49], which is a pretrained model. These speech features are then similarly processed by a 2-layer FFNN. This standardization of feature dimensions across modalities primes the data for integration.

The fusion of these data streams is executed through a DRUW fusion block, effectively merging the standardized features from physiological signals and speech. This fusion process not only integrates the data but also applies DRUW fusion.

Upon fusion, the data advances into a specific emotional FFNN and a transformer layer, means, the combined features into a global emotional space. This space is not user-specific; rather, it serves as a shared domain, namely, macro space.

Personalization is introduced at the micro stage. Here, the framework uses additional FFNN layers tailored to individual users, enabling the selection of embedding elements pertinent to their unique emotional profiles. This adaptation leverages DRUW loss.

DRUW Multimodal Fusion

A fundamental question in our study of mental health monitoring is "how to fuse different modalities," specifically the integration of physical activity and speech data. To answer this, we have developed the DRUW fusion method.

The DRUW fusion formula can be represented as follows.

where and are the uncertainty parameters for the physical activity data and the speech data , respectively. The term acts as a constraint, similar to in the DRUW loss, the weighting of each modality in the fusion process is regulated, ensuring that neither modality is disproportionately represented in the fused data and maintaining a balanced integration. This adjustment, based on the uncertainty and distinct characteristics, ensures that each modality contributes appropriately to the combined dataset. Meanwhile, the complementary nature of these data types is capitalized upon by the DRUW fusion method. Physical activity data provides objective, quantifiable measures of movement and physiological responses, while speech data offers subjective insights into emotional states and mental well-being. Furthermore, a key advantage of the DRUW fusion method is its straightforward implementation.

Macro-Micro Personalization

A key aspect of our multimodal analysis framework is the implementation of macro-micro personalization.

To quantitatively define the macro-micro personalization approach, we can formulate the integration of the macro emotional space with the micro personalization layer. This can be represented as follows.

where represents the personalized output for the participant, denotes the embeddings or features extracted from the macro emotional space, represents the embeddings or features specific to the participant, and is a weighting factor that determines the balance between the influence of the macro and micro layers. It can be a fixed value or adaptively determined based on factors such as the diversity of the dataset or the specificity of the micro data.

Macro Emotional Space

Initially, we establish a macro emotional space that serves as a common ground for all participants. This space is built using FFNN-transformer embeddings, capturing generalized emotional patterns and trends observed across the entire participant pool. It reflects the shared aspects of emotional experiences and is crucial for understanding the broader context of mental health states.

Micro Personalization Space

After macro space, a micro layer is designed for each participant. This layer allows for the customization of the model based on microspecific data. It adapts the general insights from the macro space to the nuances of each participant's emotional profile.

DRUW Loss for Multitask

MTL is a crucial component in our study, particularly relevant to the diverse nature of mental health monitoring. MTL is a form of learning that involves training a model on several related tasks simultaneously. This approach is underpinned by the principle that "transfer should always be useful"; essentially, any pair of tasks should share some commonalities in their underlying distributions [50].

In addressing the critical challenge of "how to balance different emotional states" in our study, we use the DRUW loss [29], a solution developed in our previous work. This approach is particularly crucial in the context of multitask learning, where balancing the contribution of each task—especially when dealing with a spectrum of emotional states—is key to the overall model's performance. This method allows for the adaptive balancing of tasks, taking into account the varying degrees of complexity and uncertainty inherent in each task. The equation of the DRUW loss function is as follows.

Where and are uncertainty parameters corresponding to different tasks in our model, with and representing the respective task-specific loss functions. The inclusion of serves as a constraint, regulating the sum of these weights to prevent trivial solutions and maintain the balance among tasks.

Evaluation

Experimental Setup

To construct a reliable evaluation scheme, the dataset is partitioned into training, development, and test sets based on time-dependent criteria, as outlined in Table 1. Given that the dataset is collected over a 2-week period, we allocate the first 70% of the data from each participant to the training set. The subsequent 15% forms the development set, and the remaining 15% constitutes the test set. This partitioning strategy ensures that the evaluation is robust and reflects the temporal dynamics of the data.

Table 1. Data partitioning.

Set	Data (%)	Samples
Training	70	4340
Development	15	931
Test	15	929

Evaluation Metrics

In this study, the concordance correlation coefficient (CCC) is used as a key evaluation metric. CCC considers both the scale and location shifts between the predicted and actual data, providing a comprehensive measure of the model's predictive performance. In our context, this metric is crucial for assessing the accuracy of our model in reflecting the true emotional experiences of participants.

The CCC is defined as follows.

Where represents the Pearson correlation coefficient between the predicted and actual values, and are the SDs of the predicted and actual values, respectively, and and are their means.

Benchmark Model

To effectively evaluate our macro-micro framework, it is crucial to establish a benchmark for comparison. This benchmark model consists of the following key components: pure concatenation for modality fusion, 2-layer FFNN for presentation, and equal weight strategy for MTL. This benchmark model, with its straightforward concatenation, basic personalization, and uniform task weighting, provides a solid foundation for comparison.

Pure Concatenation for Modality Fusion

This approach linearly combines features from both physical activity and speech data without weighting or transformation.

2-Layer FFNN for Personalization

We designed FFNN to adapt the concatenated features to get personalized embeddings.

Equal Weight Strategy for MTL

In handling multiple tasks, we applied an equal weight strategy across all tasks.

Comparison Methods

Multimodal Fusion Techniques

In our exploration of multimodal fusion techniques, we first investigated the use of separate transformer embeddings for each modality. This approach aimed to capture unique features within physical activity and speech data independently before combining them. The results indicate that while this method was effective in isolating modality-specific characteristics, it also necessitated sophisticated alignment strategies during the fusion stage.

We also applied attention mechanisms, including solo attention [51] and postconcatenation attention. These techniques allowed the model to dynamically focus on the most informative features from each modality. The solo attention mechanism, applied before concatenation, proved particularly effective in enhancing the model's sensitivity to contextually relevant multimodal cues.

A more straightforward approach, the pure weighted method, involved assigning fixed weights to each modality during fusion. Despite its simplicity, this method displayed limitations in adaptability, especially in scenarios where the relative importance of each modality varied.

Besides, max fusion [52] was used to capture the most significant features across modalities by taking the maximum value across feature dimensions. This method was found to be particularly useful in scenarios where the dominant features in the data were more predictive of the outcome.

We tested gated fusion [53], which was designed for dynamic control over the contribution of each modality based on the data's contextual information. This adaptability resulted in improved performance, especially in complex scenarios where the relevance of each modality changed.

Personalization Strategies

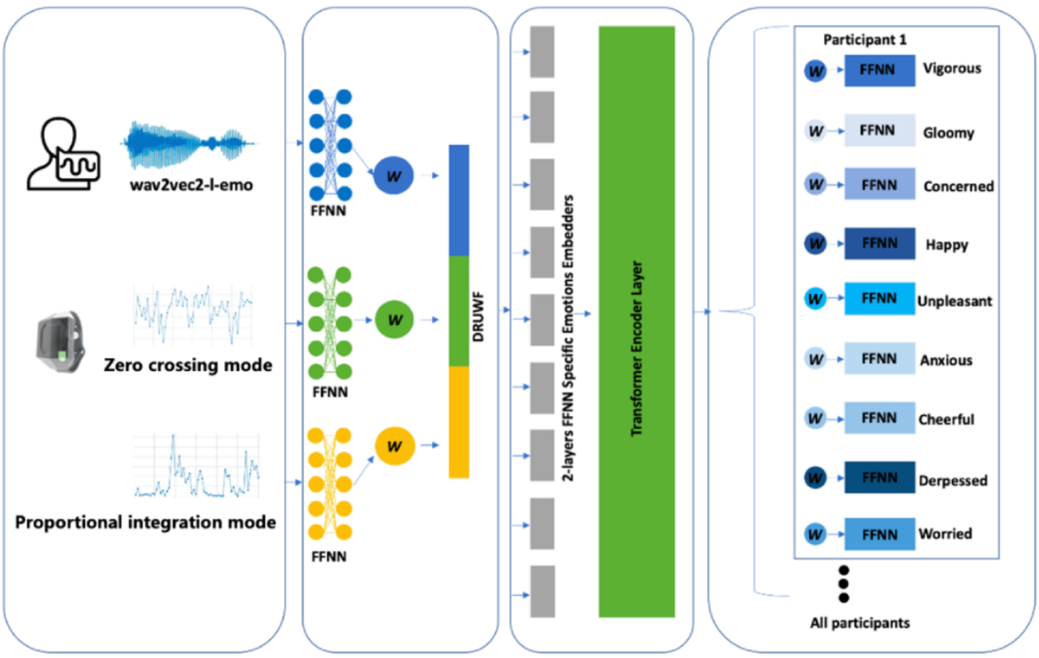
The FFNN was used as a baseline for personalization, adapting concatenated multimodal features to individual profiles. Comparatively, we used a transformer model without separate emotional FFNN for personalization.

We also applied an adapter [54] to compare. The adapter method involved integrating small, trainable modules into a pretrained model. This approach facilitated efficient and effective personalization without the need for extensive retraining.

MTL Approaches

Using a single embedding to produce outputs for multiple tasks with equal weighting provided a baseline for multitask performance. We also compared the pure weighted approach, assigning fixed weights to each task, which showed some improvements over the equal weight strategy but still lacked the dynamic adaptability required for more complex multitask scenarios. The architecture of the proposed model is given in Figure 2.

Figure 2. Macro-micro personalization framework for multimodal-multitask learning in mental health monitoring. W: weight; DRUW: Dynamic Restrained Uncertainty Weighting; DRUWF: Dynamic Restrained Uncertainty Weighting Fusion; FFNN: feed-forward neural network. The framework integrates physiological signals and speech data through feature extraction, DRUWF, macro layers with transformer encoder, and micro personalized layers.



Model Training Procedure

All models in our study were trained over 100 epochs using stochastic gradient descent with an initial learning rate set at 0.001, using a Nesterov momentum of 0.9 to enhance convergence. The learning rate was adaptively reduced by a factor of 0.9 if no improvement was observed on the development set after 5 consecutive epochs, ensuring efficient optimization. The training was conducted with a batch size of 16, balancing computational resources and model performance. Additionally, a weight decay of 0.0001 was applied to prevent overfitting. The final model configuration selected for evaluation on the test set was the one yielding the best performance on the development set, ensuring the reliability and robustness of our results. Before producing the final output, we used a sigmoid function to ensure that the predicted values ranged from 0 to 1. This adjustment was necessary because our labels had been normalized to a scale of 0 to 1.

Ethical Considerations

This study was approved by the Ethics Committee of the University of Tokyo (21-353). The study participants provided written informed consent.

Results

Overview

Our results first show CCC for various emotional dimensions using different single-modal data types (physical activity and audio), with and without personalization in Table 2. Our analysis also investigated the efficacy of various multimodal fusion techniques and their capacity for personalization in assessing different emotional dimensions in Table 3. The multimodal fusion approaches examined included basic, max fusion, gated fusion, attention fusion, solo-attention fusion, cross-modal attention, and a number of proposed methods. Each of these techniques was also analyzed in conjunction with various multitasking frameworks such as basic, transformer, adapter, equal, multioutputs, and our proposed method. In comparison between the 2 tables, multimodal results generally outperform single-modal ones.

The results indicate a differential impact on the CCC across emotional states and fusion methods. For instance, the max fusion approach yielded a CCC of 0.451 for vigorous, which was a notable improvement over the basic approach's 0.415. However, this method seemed less effective for gloomy, with a CCC of 0.356. In contrast, the gated fusion technique exhibited a more consistent performance across different emotional states, with CCCs ranging from 0.277 for anxious to 0.537 for worried.

Of particular interest were the results from the proposed methods, which showed promising CCC values across several emotional states. The highest recorded CCC of the proposed methods was for worried, with a value of 0.581. Conversely, gloomy showed the lowest CCC at 0.377 using the same proposed methods.

Overall, the mean CCC values across all emotional states suggested that the proposed methods combined with our proposed multitasking framework outperformed the other techniques, achieving a mean CCC of 0.503. This mean value was computed by averaging CCCs across all the emotional states for each method. Notably, the proposed methods consistently yielded CCC values above the overall mean, underscoring their potential for enhancing emotion recognition tasks in multimodal settings.

In conclusion, our results underscore the importance of choosing the appropriate fusion and multitasking methods to maximize the agreement between predicted and actual emotional states. The proposed methods, when tailored for individual emotional dimensions, demonstrate significant promise for personalization, which is a critical aspect of effective emotional state prediction.

Table 2. CCC[a] for various emotional dimensions using different single-modal data types (physical activity and audio), with and without personalization. The emotional dimensions covered are vigorous, gloomy, concerned, happy, unpleasant, anxious, cheerful, depressed and worried.

Single model and personalization	Vigorous	Gloomy	Concerned	Happy	Unpleasant	Anxious	Cheerful	Depressed	Worried	Mean (SD)[b]
ZCM[c]										
No	0.287	0.187	0.325	0.341	0.364	0.179	0.224	0.282	0.334	0.281 (0.069)
Yes	0.407	0.426	0.499	0.485	0.487	0.349	0.322	0.454	0.535	0.441 (0.071)
PIM[d]										
No	0.130	0.142	0.368	0.251	0.374	0.244	0.006	0.302	0.291	0.225 (0.121)
Yes	0.145	0.315	0.499	0.368	0.489	0.272	0.051	0.493	0.537	0.341 (0.172)
Speech										
No	0.287	0.187	0.325	0.341	0.364	0.179	0.224	0.282	0.334	0.281 (0.069)
Yes	0.407	0.426	0.499	0.485	0.487	0.349	0.322	0.454	0.535	0.441 (0.071)

[a]CCC: concordance correlation coefficient.

[b]The “Mean” column represents the average CCC across the emotional dimensions.

[c]ZCM: zero crossing mode.

[d]PIM: proportional integration mode.

Table 3. CCC[a] for various emotional dimensions, measured using participant-dependent partitions on our dataset. The table also demonstrates the effectiveness of different fusion methods and indicates whether personalization was used. The emotional dimensions covered are vigorous, gloomy, concerned, happy, unpleasant, anxious, cheerful, depressed, and worried.

Multimodal fusion	Personalization	Multitask	Vigorous	Gloomy	Concerned	Happy	Unpleasant	Anxious	Cheerful	Depressed	Worried	Mean (SD)[b]
Basic	Basic	Basic	0.415	0.404	0.469	0.452	0.499	0.212	0.304	0.472	0.514	0.416 (0.099)
Max fusion	Basic	Basic	0.451	0.356	0.464	0.465	0.501	0.356	0.325	0.421	0.509	0.428 (0.067)
Gated fusion	Basic	Basic	0.417	0.417	0.495	0.473	0.497	0.277	0.317	0.446	0.537	0.431 (0.086)
Attention fusion	Basic	Basic	0.414	0.281	0.524	0.413	0.554	0.386	0.317	0.447	0.574	0.435 (0.102)
Solo-attention fusion	Basic	Basic	0.403	0.337	0.488	0.467	0.513	0.394	0.298	0.470	0.540	0.434 (0.082)
Cross-modal attention	Basic	Basic	0.436	0.443	0.483	0.455	0.497	0.268	0.323	0.463	0.527	0.447 (0.084)
Proposed	Basic	Basic	0.393	0.401	0.504	0.435	0.488	0.469	0.322	0.465	0.556	0.449 (0.069)
Basic	Transformer	Basic	0.434	0.330	0.547	0.470	0.548	0.538	0.325	0.421	0.578	0.466 (0.09)
Basic	Adapter	Basic	0.457	0.329	0.554	0.463	0.549	0.543	0.320	0.432	0.586	0.472 (0.098)
Basic	Proposed	Basic	0.460	0.401	0.554	0.462	0.544	0.557	0.351	0.441	0.585	0.484 (0.080)
Basic	Basic	Equal	0.420	0.321	0.521	0.436	0.511	0.446	0.298	0.381	0.563	0.431 (0.090)
Basic	Basic	Multioutput	0.407	0.399	0.493	0.470	0.526	0.469	0.324	0.467	0.531	0.454 (0.067)
Basic	Basic	Proposed	0.434	0.330	0.547	0.470	0.548	0.538	0.325	0.422	0.579	0.466 (0.095)
Proposed	Proposed	Basic	0.414	0.389	0.564	0.522	0.556	0.581	0.364	0.419	0.589	0.489 (0.091)
Basic	Proposed	Proposed	0.454	0.450	0.549	0.519	0.554	0.581	0.358	0.424	0.583	0.497 (0.079)
Proposed	Basic	Proposed	0.449	0.377	0.525	0.466	0.554	0.573	0.296	0.437	0.543	0.469 (0.091)
Proposed	Proposed	Proposed	0.464	0.464	0.549	0.538	0.554	0.581	0.373	0.420	0.581	0.503 (0.075)

[a]CCC: concordance correlation coefficient.

[b]The "Mean" column represents the average CCC across the emotional dimensions.

Statistical Validation

We also conducted a statistical analysis to complement the CCC results from our deep learning model, crucial for validating the model's reliability and generalizability across different datasets and conditions. Our mixed linear model analysis [55], presented in Table 4, reveals 2 critical insights. First, the highly significant within-individual associations (Q) across 9 emotional scales underscore the model's capability to capture nuanced emotional responses, indicating its robust predictive power. Second, the observation of group and residual variances highlights the variability that the model does not account for, signaling areas that require further refinement. This unexplained variability invites a deeper investigation into potential factors, such as the model's sensitivity to specific data characteristics or the necessity for incorporating a more diverse range of training data. Understanding these elements can guide targeted improvements in the model's architecture and training process, ultimately enhancing its accuracy and applicability in personalized mental health monitoring.

Table 4. Mixed linear model regression results for emotional dimensions[a].

Emotional dimension	Intercept	Regression coefficient	SE	*z* score	*P* value>|*z*|	95% CI	Group variance	Residual variance
Vigorous	0.155	0.549	0.076	7.176	<.001	0.399-0.698	0.039	0.057
Gloomy	0.087	0.607	0.064	9.512	<.001	0.482-0.733	0.023	0.067
Concerned	0.153	0.500	0.073	6.813	<.001	0.356-0.644	0.043	0.057
Happy	0.236	0.446	0.077	5.771	<.001	0.295-0.598	0.030	0.064
Unpleasant	0.071	0.630	0.066	9.534	<.001	0.501-0.760	0.028	0.061
Anxious	0.127	0.577	0.071	8.177	<.001	0.439-0.716	0.035	0.061
Cheerful	0.210	0.477	0.074	6.475	<.001	0.333-0.621	0.022	0.072
Depressed	0.044	0.647	0.060	10.758	<.001	0.529-0.765	0.021	0.062
Worried	0.138	0.539	0.076	7.097	<.001	0.390-0.688	0.048	0.052

[a]The table summarizes the intercept, regression coefficients, standard errors, *z* scores, *P* values, CIs, group variances, and residual variances for each emotion studied. The emotional dimensions covered are vigorous, gloomy, concerned, happy, unpleasant, anxious, cheerful, depressed, and worried.

Discussion

Overview

This study introduces a novel dataset and a macro-micro framework for personalized daily mental health monitoring, leveraging multimodal and MTL strategies. The results demonstrate the efficacy of our approach in predicting emotional states, with a mean CCC of 0.503 across 9 emotional dimensions.

The proposed macro-micro framework, which combines macro-level emotion transformer embeddings with micro-level personalization layers, shows superior performance compared to traditional approaches. This suggests that incorporating both general emotional patterns and individual-specific adaptations is crucial for accurate mental health monitoring. The effectiveness of our DRUW fusion method in integrating multimodal data further underscores the importance of adaptive weighting strategies in handling diverse data types.

Our findings align with previous studies highlighting the potential of multimodal approaches in mental health monitoring. However, our work extends beyond existing research by incorporating personalization at both macro and micro levels, addressing a critical gap in current mental health technology.

The high significance of within-individual associations across emotional scales, as revealed by our mixed linear model analysis, validates the model's capability to capture nuanced emotional responses. This has important implications for the development of personalized mental health interventions, as it suggests that our model can detect subtle changes in an individual's emotional state over time.

However, the observed group and residual variances in our statistical analysis indicate that there is still unexplained variability in emotional states. This highlights a limitation of our model and suggests that additional factors, not captured in our framework, may influence daily emotional states. These could include external stressors, social interactions, or physiological factors not measured in our study.

Another limitation is the reliance on self-reported emotional states, which, while valuable for capturing subjective experiences, may be subject to reporting biases. Future research could explore the integration of objective measures of emotional state, such as facial expression analysis or additional physiological markers, to complement self-reports.

Looking ahead, several avenues for future research emerge from our findings. First, expanding the dataset to include a more diverse range of participants and longer monitoring periods could enhance the generalizability of our model. Second, investigating the incorporation of additional modalities, such as sleep patterns or social media activity, could provide a more comprehensive picture of mental health. Finally, exploring the application of our framework in clinical settings could help bridge the gap between research and practical mental health interventions.

Conclusions

In conclusion, this study introduces a groundbreaking dataset and a macro-micro framework that significantly advances personalized daily mental health monitoring. By leveraging multimodal and MTL strategies, we have demonstrated a robust model capable of predicting emotional states. The statistical analysis further validates the model's reliability, highlighting its potential for wider application in the mental health domain.

Moving forward, our focus will be on expanding the dataset, incorporating additional modalities, and refining our model to address these variances, with the ultimate goal of making daily mental health monitoring a more accessible, nonintrusive, and personalized practice.

Acknowledgments

The authors acknowledge support by the Japan Science and Technology Agency (JST) Mirai Program (grant 21473074), JST FORESTO (grant JPMJFR2156) Program, and the JST MOONSHOT Program (grant JPMJMS229B).

Conflicts of Interest

None declared.

References

1. Rosenfield S, Mouzon D. Gender and mental health. In: Aneshensel CS, Phelan JC, Bierman A, editors. Handbook of the Sociology of Mental Health. Dordrecht, Netherlands: Springer; 2013:277-296.
2. Hou W, Lai FT, Ben-Ezra M, Goodwin R. Regularizing daily routines for mental health during and after the COVID-19 pandemic. J Glob Health 2020;10(2):020315 [FREE Full text] [doi: 10.7189/jogh.10.020315] [Medline: 33110518]
3. Laird KT, Tanner-Smith EE, Russell AC, Hollon SD, Walker LS. Comparative efficacy of psychological therapies for improving mental health and daily functioning in irritable bowel syndrome: a systematic review and meta-analysis. Clin Psychol Rev 2017;51:142-152 [FREE Full text] [doi: 10.1016/j.cpr.2016.11.001] [Medline: 27870997]
4. Song M, Triantafyllopoulos A, Yang Z, Takeuchi H, Nakamura T, Kishi A, et al. Daily mental health monitoring from speech: a real-world Japanese dataset and multitask learning analysis. 2023 Presented at: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); June 10, 2023; Rhodes, Greece p. 1-5. [doi: 10.1109/icassp49357.2023.10096884]
5. Tripathi S, Beigi H. Multi-modal emotion recognition on IEMOCAP dataset using deep learning. arXiv Preprint posted online on April 16, 2018. [doi: 10.48550/arXiv.1804.05788]
6. Palotti J, Narula G, Raheem L, Bay H. Analysis of emotion annotation strength improves generalization in speech emotion recognition models. 2023 Presented at: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); June 17, 2023; Vancouver, BC, Canada p. 5828-5836. [doi: 10.1109/cvprw59228.2023.00619]
7. Fan W, Xu X, Xing X, Chen W, Huang D. LSSED: a large-scale dataset and benchmark for speech emotion recognition. 2021 Presented at: ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); June 06, 2021; Toronto, ON, Canada p. 641-645. [doi: 10.1109/icassp39728.2021.9414542]
8. Bertero D, Fung P. A first look into a convolutional neural network for speech emotion detection. 2017 Presented at: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); March 05, 2017; New Orleans, LA p. 5115-5119. [doi: 10.1109/icassp.2017.7953131]
9. Truong KP, van Leeuwen DA, de Jong FMG. Speech-based recognition of self-reported and observed emotion in a dimensional space. Speech Commun 2012;54(9):1049-1063. [doi: 10.1016/j.specom.2012.04.006]
10. Trull TJ, Ebner-Priemer U. Ambulatory assessment. Annu Rev Clin Psychol 2013;9:151-176 [FREE Full text] [doi: 10.1146/annurev-clinpsy-050212-185510] [Medline: 23157450]
11. Suhaimi NS, Mountstephens J, Teo J. EEG-based emotion recognition: a state-of-the-art review of current trends and opportunities. Comput Intell Neurosci 2020;2020:8875426 [FREE Full text] [doi: 10.1155/2020/8875426] [Medline: 33014031]
12. Jenke R, Peer A, Buss M. Feature extraction and selection for emotion recognition from EEG. IEEE Trans Affective Comput 2014;5(3):327-339. [doi: 10.1109/taffc.2014.2339834]
13. Ko BC. A brief review of facial emotion recognition based on visual information. Sensors 2018;18(2):401 [FREE Full text] [doi: 10.3390/s18020401] [Medline: 29385749]
14. Canal FZ, Müller TR, Matias JC, Scotton GG, de Sa Junior AR, Pozzebon E, et al. A survey on facial emotion recognition techniques: a state-of-the-art literature review. Inf Sci 2022;582:593-617. [doi: 10.1016/j.ins.2021.10.005]
15. El Ayadi M, Kamel MS, Karray F. Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit 2011;44(3):572-587. [doi: 10.1016/j.patcog.2010.09.020]
16. Wani TM, Gunawan TS, Qadri SAA, Kartiwi M, Ambikairajah E. A comprehensive review of speech emotion recognition systems. IEEE Access 2021;9:47795-47814. [doi: 10.1109/access.2021.3068045]
17. Massaro S, Pecchia L. Heart rate variability (HRV) analysis: a methodology for organizational neuroscience. Organizational Res Methods 2016;22(1):354-393. [doi: 10.1177/1094428116681072]
18. Kemp AH, Quintana DS, Gray MA, Felmingham KL, Brown K, Gatt JM. Impact of depression and antidepressant treatment on heart rate variability: a review and meta-analysis. Biol Psychiatry 2010;67(11):1067-1074. [doi: 10.1016/j.biopsych.2009.12.012] [Medline: 20138254]

19. Beauchaine TP, Thayer JF. Heart rate variability as a transdiagnostic biomarker of psychopathology. Int J Psychophysiol 2015;98(2):338-350. [doi: 10.1016/j.ijpsycho.2015.08.004] [Medline: 26272488]
20. Kim J, Nakamura T, Kikuchi H, Yoshiuchi K, Sasaki T, Yamamoto Y. Covariation of depressive mood and spontaneous physical activity in major depressive disorder: toward continuous monitoring of depressive mood. IEEE J Biomed Health Inform 2015;19(4):1347-1355. [doi: 10.1109/JBHI.2015.2440764] [Medline: 26054079]
21. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak CP. The role of actigraphy in the study of sleep and circadian rhythms. Sleep 2003;26(3):342-392. [doi: 10.1093/sleep/26.3.342] [Medline: 12749557]
22. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. Speech Commun 2015;71:10-49. [doi: 10.1016/j.specom.2015.03.004]
23. Mundt JC, Vogel AP, Feltner DE, Lenderking WR. Vocal acoustic biomarkers of depression severity and treatment response. Biol Psychiatry 2012;72(7):580-587 [FREE Full text] [doi: 10.1016/j.biopsych.2012.03.015] [Medline: 22541039]
24. Cohen AS, Elvevåg B. Automated computerized analysis of speech in psychiatric disorders. Curr Opin Psychiatry 2014;27(3):203-209 [FREE Full text] [doi: 10.1097/YCO.0000000000000056] [Medline: 24613984]
25. Torous J, Kiang MV, Lorme J, Onnela JP. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR Ment Health 2016;3(2):e16 [FREE Full text] [doi: 10.2196/mental.5165] [Medline: 27150677]
26. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [doi: 10.1146/annurev.clinpsy.3.022806.091415] [Medline: 18509902]
27. Fahrenberg J, Myrtek M, Pawlik K, Perrez M. Ambulatory assessment—monitoring behavior in daily life settings. Eur J Psychol Assess 2007;23(4):206-213. [doi: 10.1027/1015-5759.23.4.206]
28. Kendall M, Stuart A. The Advance Theory of Statistics. London: Charles Griffin and Company, Ltd.; 1943.
29. Song M, Yang Z, Triantafyllopoulos A, Jing X, Karas V, Xie J, et al. Dynamic restrained uncertainty weighting loss for multitask learning of vocal expression. arXiv Preprint posted online on June 22, 2022. [doi: 10.48550/arXiv.2206.11049]
30. Dagum P. Digital biomarkers of cognitive function. NPJ Digit Med 2018;1:10 [FREE Full text] [doi: 10.1038/s41746-018-0018-4] [Medline: 31304295]
31. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA 2017;318(13):1215-1216. [doi: 10.1001/jama.2017.11295] [Medline: 28973224]
32. Torous J, Jän Myrick K, Rauseo-Ricupero N, Firth J. Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow. JMIR Ment Health 2020;7(3):e18848 [FREE Full text] [doi: 10.2196/18848] [Medline: 32213476]
33. Mohr DC, Riper H, Schueller SM. A solution-focused research approach to achieve an implementable revolution in digital mental health. JAMA Psychiatry 2018;75(2):113-114. [doi: 10.1001/jamapsychiatry.2017.3838] [Medline: 29238805]
34. Hickey BA, Chalmers T, Newton P, Lin CT, Sibbritt D, McLachlan CS, et al. Smart devices and wearable technologies to detect and monitor mental health conditions and stress: a systematic review. Sensors 2021;21(10):3461 [FREE Full text] [doi: 10.3390/s21103461] [Medline: 34065620]
35. Long N, Lei Y, Peng L, Xu P, Mao P. A scoping review on monitoring mental health using smart wearable devices. Math Biosci Eng 2022;19(8):7899-7919 [FREE Full text] [doi: 10.3934/mbe.2022369] [Medline: 35801449]
36. Xu H, Wu X, Liu X. A measurement method for mental health based on dynamic multimodal feature recognition. Front Public Health 2022;10:990235 [FREE Full text] [doi: 10.3389/fpubh.2022.990235] [Medline: 36620271]
37. Huckins JF, daSilva AW, Wang W, Hedlund E, Rogers C, Nepal SK, et al. Mental health and behavior of college students during the early phases of the COVID-19 pandemic: longitudinal smartphone and ecological momentary assessment study. J Med Internet Res 2020;22(6):e20185 [FREE Full text] [doi: 10.2196/20185] [Medline: 32519963]
38. Das R, Hasan MR, Daria S, Islam MR. Impact of COVID-19 pandemic on mental health among general Bangladeshi population: a cross-sectional study. BMJ Open 2021;11(4):e045727 [FREE Full text] [doi: 10.1136/bmjopen-2020-045727] [Medline: 33837107]
39. Morshed MB, Saha K, Li R, D'Mello SK, De Choudhury M, Abowd GD, et al. Prediction of mood instability with passive sensing. 2019 Presented at: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies; September 09, 2019; New York, NY p. 1-21. [doi: 10.1145/3351233]
40. Takeuchi H, Suwa K, Kishi A, Nakamura T, Yoshiuchi K, Yamamoto Y. The effects of objective push-type sleep feedback on habitual sleep behavior and momentary symptoms in daily life: mHealth intervention trial using a health care internet of things system. JMIR Mhealth Uhealth 2022;10(10):e39150 [FREE Full text] [doi: 10.2196/39150] [Medline: 36201383]
41. Melcher J, Hays R, Torous J. Digital phenotyping for mental health of college students: a clinical review. Evid Based Ment Health 2020;23(4):161-166 [FREE Full text] [doi: 10.1136/ebmental-2020-300180] [Medline: 32998937]
42. Gerczuk M, Triantafyllopoulos A, Amiriparian S, Kathan A, Bauer J, Berking M, et al. Zero-shot personalization of speech foundation models for depressed mood monitoring. Patterns 2023;4(11):100873 [FREE Full text] [doi: 10.1016/j.patter.2023.100873] [Medline: 38035199]
43. Su C, Xu Z, Pathak J, Wang F. Deep learning in mental health outcome research: a scoping review. Transl Psychiatry 2020;10(1):116 [FREE Full text] [doi: 10.1038/s41398-020-0780-3] [Medline: 32532967]

44. Fukui I. The Depression and Anxiety Mood Scale (DAMS): scale development and validation. Jpn J Behav Ther 1997;23:83-93 [FREE Full text] [doi: 10.24468/jjbt.23.2_83]
45. Satt A, Rozenberg S, Hoory R. Efficient emotion recognition from speech using deep learning on spectrograms. 2017 Presented at: Interspeech 2017; August 10, 2017; Stockholm, Sweden p. 1089-1093. [doi: 10.21437/interspeech.2017-200]
46. Burton C, McKinstry B, Tătar AS, Serrano-Blanco A, Pagliari C, Wolters M. Activity monitoring in patients with depression: a systematic review. J Affect Disord 2013;145(1):21-28 [FREE Full text] [doi: 10.1016/j.jad.2012.07.001] [Medline: 22868056]
47. Helgadóttir B, Forsell Y, Ekblom Ö. Physical activity patterns of people affected by depressive and anxiety disorders as measured by accelerometers: a cross-sectional study. PLoS One 2015;10(1):e0115894 [FREE Full text] [doi: 10.1371/journal.pone.0115894] [Medline: 25585123]
48. Loizou PC. Speech Enhancement: Theory and Practice. 2nd Edition. Boca Raton, FL: CRC Press; 2013.
49. Wagner J, Triantafyllopoulos A, Wierstorf H, Schmitt M, Burkhardt F, Eyben F, et al. Dawn of the transformer era in speech emotion recognition: closing the valence gap. IEEE Trans Pattern Anal Mach Intell 2023;45(9):10745-10759. [doi: 10.1109/TPAMI.2023.3263585] [Medline: 37015129]
50. Zhang Y, Wei Y, Yang Q. Learning to multitask. 2018 Presented at: Advances in Neural Information Processing Systems; November 21, 2001; Cambridge, MA p. 5771-5782. [doi: 10.7551/mitpress/11474.003.0014]
51. Song M, Yang Z, Triantafyllopoulos A, Nakamura T, Zhang Y, Ren Z, et al. Crossmodal transformer on multi-physical signals for personalised daily mental health prediction. 2023 Presented at: IEEE ICDM 2023—23rd IEEE International Conference on Data Mining; January 11, 2023; Shanghai, China p. 1299-1305. [doi: 10.1109/icdmw60847.2023.00167]
52. Kuncheva LI. A theoretical study on six classifier fusion strategies. IEEE Trans Pattern Anal Machine Intell 2002;24(2):281-286. [doi: 10.1109/34.982906]
53. Ren W, Ma L, Zhang J, Pan J, Cao X, Liu W, et al. Gated fusion network for single image dehazing. 2018 Presented at: CVPR 2018; June 05, 2018; Salt Lake City, UT p. 3253-3261. [doi: 10.1109/cvpr.2018.00343]
54. Dong W, Yan D, Lin Z, Wang P. Efficient adaptation of large vision transformer via adapter re-composing. 2023 Presented at: 2023 Conference on Neural Information Processing Systems (NeurIPS 2023); December 10, 2023; New Orleans, LA URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/a4ca07aa108036f80cbb5b82285fd4b1-Abstract-Conference.html
55. Kalaian HA, Raudenbush SW. A multivariate mixed linear model for meta-analysis. Psychol Methods 1996;1(3):227-235. [doi: 10.1037//1082-989x.1.3.227]

Abbreviations

CCC: concordance correlation coefficient
DAMS: Depression and Anxiety Mood Scale
DRUW: Dynamic Restrained Uncertainty Weighting
EEG: electroencephalography
EMA: ecological momentary assessment
FFNN: feed-forward neural network
HRV: heart rate variability
MHIT: Mental Healthcare Internet of Things
MTL: multitask learning
PIM: proportional integration mode
ZCM: zero crossing mode.

Edited by J Torous; submitted 14.04.24; peer-reviewed by M Bagewadi Ellur, L Henry, X Wu; comments to author 22.05.24; revised version received 16.07.24; accepted 03.08.24; published 18.10.24.

Please cite as:
Song M, Yang Z, Triantafyllopoulos A, Zhang Z, Nan Z, Tang M, Takeuchi H, Nakamura T, Kishi A, Ishizawa T, Yoshiuchi K, Schuller B, Yamamoto Y
Empowering Mental Health Monitoring Using a Macro-Micro Personalization Framework for Multimodal-Multitask Learning: Descriptive Study
JMIR Ment Health 2024;11:e59512
URL: https://mental.jmir.org/2024/1/e59512
doi:10.2196/59512
PMID:39422993

Original Paper

Experiences of Patients With Mental Health Issues Having Web-Based Access to Their Records: National Patient Survey

Jonas Moll[1], PhD; Gunilla Myreteg[2*], PhD; Hanife Rexhepi[3*], PhD

[1]Centre for Empirical Research on Information Systems, Örebro University School of Business, Örebro University, Örebro, Sweden

[2]Department of Business Studies, Uppsala University, Uppsala, Sweden

[3]School of Informatics, University of Skövde, Skövde, Sweden

[*]these authors contributed equally

Corresponding Author:
Jonas Moll, PhD
Centre for Empirical Research on Information Systems
Örebro University School of Business
Örebro University
Nova Building, 4th floor
Fakultetsgatan 1
Örebro, 70182
Sweden
Phone: 46 19 303878
Email: jonas.moll@oru.se

Abstract

Background: Sharing mental health notes through patient accessible electronic health records (PAEHRs) is controversial. Many psychiatric organizations and regions in Sweden have resisted the implementation, as clinicians worry about possible harms when patients are reading their notes. Despite the documented benefits of PAEHRs, there is still a lack of knowledge regarding whether patients with mental health issues could reap similar benefits of reading their notes as other patient groups.

Objective: The aim of the study is to examine the use, attitudes, and experiences of patients with mental health issues by reading their notes in the PAEHR and, moreover, whether their experiences differ from other patient groups, and if so, how.

Methods: A national patient survey was conducted with answers from 2587 patients from different patient groups. In total, 504 respondents (19.5%) indicated that they experienced a mental health disease. Answers from this patient group were compared to the answers from all other respondents. Survey questions related to attitudes, information usage, and effects on contacts with care were selected for analysis. Mann-Whitney *U* tests were used to detect groupwise differences.

Results: Patients with mental health issues use PAEHRs for checking that they have received the right care (mean_mental health 2.83, SD_mental health 1.39; mean_others 2.62, SD_others 1.37; *P*=.002) or suspected inaccuracies (mean_mental health 2.55, SD_mental health 1.34; mean_others 2.31, SD_others 1.30; *P*=.001), blocking access for professionals in other specialties (mean_mental health 3.43, SD_mental health 1.46; mean_others 3.04, SD_others 1.42; *P*<.001), and checking which care professionals have accessed their record (mean_mental health 4.28, SD_mental health 1.14; mean_others 4.05, SD_others 1.25; *P*<.001) to a significantly higher degree than other patients. On the other hand, the results show that a significantly lower proportion of patients with mental health issues (mean_mental health 3.38, SD_mental health 1.21; mean_others 3.52, SD_others 1.18; *P*=.02) believe that PAEHRs help them in shared decision-making compared to other patient groups.

Conclusions: Patients with mental health issues who took part in the survey, as a group, express some minor differences in both the use of the PAEHR and their experiences regarding its usefulness, as compared to other patients, as a group. This patient group shows a slightly higher interest in 2 types of use: checking for accuracy of care in the record and blocking access to mental health notes for professionals from other parts of the health care system. Compared to other patient groups, these patients are less likely to experience that the PAEHR is a support in shared decision-making. The study indicates that the benefits of PAEHR on a general level are the same for this patient group as for other patients. The study does not support clinicians' worry about possible harm to this patient group. Further research is however needed.

(JMIR Ment Health 2024;11:e48008) doi:10.2196/48008

KEYWORDS

patient accessible electronic health record; patient portal; patient experiences; mental health; eHealth; national survey; digital mental health; digital health

Introduction

Patient accessible electronic health records (PAEHRs) aim to promote patients' engagement with their care by giving patients direct access to their electronic health records (EHRs) through a national patient portal. Patients in around 20 countries worldwide, including Estonia, the Nordic countries, Australia, the United States, Canada, and England, are now offered web-based access to at least some of their EHRs. In Sweden, the PAEHR called "Journalen" was launched in 2012, when the region of Uppsala offered all citizens 18 years and older of age access to their EHRs through the national patient portal 1177 Vårdguiden. In 2015, Journalen was launched as the national system in Sweden for web-based access to clinical notes, and at the end of 2018, all regions had implemented Journalen. The PAEHR offers the patient access to his or her medical notes, prescribed medications, laboratory results, diagnosis, maternity care records, referrals, and vaccinations. Since health care in Sweden is governed by 21 autonomous regions with their own regulations, there are some regional differences concerning what type of information a patient can access and how soon (immediately or after 2 weeks). All regions offer patients access to visit notes from somatic care and test results.

Despite the documented benefits of PAEHRs [1-3], clinicians have raised concerns that patients could become confused or anxious by what they read [4]. The web-based access to mental health notes is especially controversial. The clinicians' main argument is that in mental health, the information concerns sensitive topics that can have negative consequences for patients when they access their notes. In the study by Peck et al [5], several clinicians approved of the possibility to exclude patients from access, when they were considered too vulnerable. Different survey studies related to PAEHR and mental health suggest that clinicians worry about possible harms, and many health care professionals anticipate that patients will become confused, get angry, or decompensate when reading their notes [6,7]. Other studies report that patients with mental health can benefit from accessing their notes. Some reported benefits are increased feelings of engagement [8-10], feeling of control over their health, trusting their providers, taking better care of themselves, remembering their care plan, understanding better the rationale for medications, and being more likely to take their medications as prescribed [11,12]. A small number of studies have however found negative consequences for patients with mental health issues because of reading the mental health notes, such as feeling judged, worried, or offended [5,12-14]. A majority of studies [5,9,15] suggest that outpatient patients with mental health issues value reading their notes, that psychiatrists do not experience increased work burden or perceive negative outcomes, and that respectful, accurate mental health notes may enhance patient trust.

In Sweden, region Skåne made mental health notes accessible to adult patients from 2015. Since then, more regions have followed, and as of today, 17 of 21 regions share mental health notes through the PAEHR Journalen [16]. There are no other differences between patients with mental health issues and other patient groups regarding what information types you have access to and when. Thus, patients with mental health issues have the same access to notes from somatic care, test results, and other information types accessible in their region as all other patients. Attitudes among physicians were studied (along with other professional affiliations) before and after the implementation of PAEHR in region Skåne [7]. That study reported that some physicians were more careful with what they documented in the record, as a result of not knowing how the patient might interpret and use the information. Similar results were reported by Dobscha et al [6] and Denneson et al [17], who found that clinicians were less detailed and changed their tone of the notes when they knew that the patient might choose to read the notes. Respondents in the study by Petersson and Erlingsdóttir [7] also indicated a fear of increasing tension between clinicians and the patient, which could manifest itself in threats and acts of violence. In their follow-up study, the professionals rather expressed that there were no changes in patient involvement after the implementation of PAEHRs [18].

Additionally, Denneson et al [17] reported that clinicians expressed concern that access to mental health notes "could damage the therapeutic relationship by exposing a disconnect between the patients' in-person experience with their clinicians and the documentation they read in their notes." Mental health notes can, they argue, reveal aspects of the therapeutic process—such as clinical formulations and subjective impressions—which clinicians frequently do not communicate to their patients. Thus, a patient reading his or her notes could cause the patient to misinterpret the clinician's notation, which could have negative effects on the patient, such as having feelings of being judged or stigmatized [17]. Moreover, patients with mental health issues are also generally considered a vulnerable patient group, which begs the question whether patients with mental health conditions can reap similar benefits of accessing their PAEHRs as other patient groups [19]. This review shows the topic to be controversial, and in practice, many psychiatric organizations resist implementation.

Empirical research is scarce, especially in Sweden. To our knowledge, there is no comparative research on how PAEHRs are perceived among patients with mental health issues as compared to other patient groups. Moreover, Petersson and Erlingsdóttir [18] make a call for empirical research regarding the perspectives of patients with mental health issues toward PAEHRs. The aim of this paper is to examine the use, attitudes, and experiences of patients with mental health issues by reading their notes in the PAEHR and if their experiences differ from other patient groups. More practical knowledge is needed in this area as input to the ongoing debate regarding the possible benefits for patients in accessing their mental health notes through PAEHRs.

Methods

Ethical Considerations

The survey, which focused on attitudes toward and experiences of using Journalen, was approved by the regional ethical review board in Uppsala, Sweden (EPN 2017/045). The respondents were informed about the voluntary participation and the aim of the study as well as presented with standard consent that needed to be accepted before the survey could be started. No data were stored unless the respondent chose to submit the answers at the end of the survey. The data were anonymized by representing each respondent with a number. No incentives were offered for participation.

Study Design

This paper is based on data from an open anonymous self-completion digital national patient survey distributed to users of the Swedish PAEHR system, Journalen, through a link on the login page. Thus, all citizens who logged in to the service during the period that the questionnaire was accessible (June to October 2016) were potential respondents to the voluntary survey.

The survey included 24 questions with a combination of Likert-scale items, multiple-choice items, and free-text alternatives. The questions covered the following themes: attitudes and reactions, access to and usage of information, effects on contact with health care, information content, security and privacy, personal health information, and demographics.

The theme “personal health information” included a question about which diagnosis group the respondents identified himself or herself to belong to. The respondents could choose between the alternatives of cancer, mental health, diabetes, high blood pressure, and others. The diagnoses of cancer, diabetes, and high blood pressure were specified as survey alternatives as they are the most common chronic conditions in Sweden. The alternative to mental health was included in order to address the ongoing debate in Sweden regarding whether psychiatric records should be made available and whether this patient group can benefit from accessing their medical record. Of the respondents, 504 people chose to identify themselves as belonging to the group of patients with mental health issues. This constitutes 19.5% (n=504) of the respondents who answered the survey. Globally, about 1 in every 8 people live with a mental disorder—most commonly, an anxiety or depressive disorder [20]. In Sweden, an official national health survey reported that 16% of respondents experienced severe mental difficulties, but that as many as 71% of respondents experienced feelings of anxiety or worry [21]. It is thus not remarkable or questionable that as many as 19.5% (n=504) of respondents in this study identified themselves to belong to this group of patients.

The full national survey was analyzed and presented by Moll et al [22]. In this study, 7 Likert-scale questions, including several items, related to attitudes, information usage, and contacts with care were selected for further analysis in relation to patients with mental health issues. Questions related to general attitudes and information usage were also picked out in order to shed light on any differences regarding how patients view and use the possibilities that Journalen gives. Finally, questions related to information accuracy, contact with care, and involvement in the care process were selected, since they reflect issues that mental health professionals have raised, as reported by previous studies. This set of selected questions is motivated by the need to develop knowledge that addresses the controversial question regarding access of patients with mental health issues to their records, and the set of questions relate closely to the concerns that were raised by health care professionals. The paper focuses on the answers of patients with mental health issues as a group and compares those to the answers from the other respondents as constituting another group.

During the time that the survey was distributed on the login page of Journalen, only 2 regions (Skåne and Kronoberg) had opened up web-based access to mental health notes. One consequence of this is that some of the patients with mental health issues who answered the survey could not yet access the mental health notes, while others could. Patients who lived in other Swedish regions could still access other types of health information (eg, test results and notes from primary care visits). During the survey period, 154 patients (30.6% of the mental health respondents) belonged to Skåne or Kronoberg and could thus access their mental health notes.

Analysis

Apart from descriptive analysis, Mann-Whitney *U* tests were used for detecting groupwise differences in answers on the 5-point Likert-scale questions between the group of patients with mental health issues and the group of all other respondents. The same test was used for detecting groupwise differences between mental health respondents from the regions Skåne and Kronoberg, who could read the mental health notes at the time the survey was open, and all other mental health respondents. This extra comparison was performed to investigate if the survey answers were affected by the fact that some of the mental health respondents could not actually access their mental health information but only information related to somatic care. Prior to the analysis, the Likert-scale options strongly agree, agree, neutral, disagree, and strongly disagree were converted to a numerical scale (1=strongly disagree and 5=strongly agree). No free-text questions were analyzed in this study. The SPSS software (version 25; IBM Corp) was used for all calculations.

Results

Result Presentation

In the tables in the following subsections, the response options “strongly agree” and “agree” have been combined for readability. For the same reason, “strongly disagree” and “disagree” were also combined. In Multimedia Appendix 1, the results for all response options are provided for completeness. As mentioned, some patients with mental health issues could access their mental health notes in Journalen at the time of the study, while others could not. Mann-Whitney *U* tests were used to check if there were any significant differences between patients with mental health issues who could and could not access their mental health notes. Significant differences were only found for 2 of the survey questions in the study (regarding access to all types of record entries and access to log list). Hence, the vast majority of the results presented here were not affected by the respondents being able to access their mental health notes.

Demographic Information

Demographic information about the respondents is provided in Table 1 together with a comparison against demographic data from the other group of respondents. Chi-square tests were used to check for significant associations between the compared variables. The group of patients with mental health issues was a bit younger, which was expected based on current statistics [23]. Moreover, there was a female dominance that was however a bit stronger than the statistics would suggest. Additionally, the level of education is lower for the mental health group. No differences were found regarding previous work experience in health care.

Table 1. Demographic information for respondents who identified themselves as patients with mental health issues.

Demographic	Mental health, n (%)	Others, n (%)	*P* value
Age[a] (years)			<.001
18-25	75 (15)	96 (4.9)	
26-35	145 (29)	269 (13.8)	
36-45	109 (21.8)	264 (13.5)	
46-55	89 (17.8)	366 (18.8)	
56-65	55 (11)	427 (21.9)	
>66	27 (5.4)	528 (27.1)	
Sex[b]			<.001
Female	387 (78.3)	1242 (63.9)	
Male	100 (20.3)	698 (35.9)	
Others	7 (1.4)	3 (0.2)	
Works or has worked in health care[c]			.11
Yes	226 (45.4)	804 (41.4)	
No	272 (54.6)	1139 (58.6)	
Education[d]			.004
Research education	11 (2.2)	64 (3.3)	
Higher education ≥3 years	168 (33.8)	776 (39.7)	
Higher education <3 years	94 (18.9)	372 (19)	
High school ≥3 years	112 (22.5)	296 (15.1)	
High school <3 years	56 (11.3)	192 (9.8)	
Less than high school	32 (6.5)	127 (6.5)	
No formal education	11 (2.2)	55 (2.8)	
Others	13 (2.6)	72 (3.7)	

[a]Patients with mental health issues: n=500; other patients: n=1950.
[b]Patients with mental health issues: n=494; other patients: n=1943.
[c]Patients with mental health issues: n=498; other patients: n=1943.
[d]Patients with mental health issues: n=497; other patients: n=1954.

General Attitudes

Both patients with mental health issues and all other patients are generally positive toward the Swedish PAEHR system, Journalen (Table 2). The vast majority of the respondents, no matter which of the 2 groups they belong to, believe that access to Journalen is good for them (Q3b), and that web-based access to medical records is generally a good reform (Q3a). The Mann-Whitney *U* tests showed significant differences between the 2 groups for questions Q3a (mean_mental health 4.73, SD_mental health 0.68; mean_others 4.80, SD_others 0.60; *P*=.01) and Q3b (mean_mental health 4.78, SD_mental health 0.64; mean_others 4.86, SD_others 0.51; *P*=.005), indicating a slightly less positive attitude among patients with mental health issues. The difference between the 2 groups is, however, very small, pointing to that the results are not that significant.

Regarding the content within the PAEHR Journalen, the respondents were asked to rate how accurate they believe the content is. This question was split into 2 aspects: whether the information that is found in the record is correct (Q15a) and whether sufficient information was recorded (Q15b; Table 3). Both groups, patients with mental health issues and all other patients, gave a fairly high rating regarding the correctness of information (mean_mental health 3.98, SD_mental health 0.99; mean_others 4.22, SD_others 0.91; *P*<.001) and a lower score regarding the completeness of information (mean_mental health 3.32, SD_mental health 1.34; mean_others 2.85, SD_others 1.40; *P*<.001). The differences between the groups were statistically significant in both cases, indicating that patients with mental health issues were more inclined to think that information was complete but less convinced that the information was correct, as compared to all other patients. This being said, the average rating among patients with mental health issues for the question about completeness was fairly low (close to neutral), while it was higher (close to "agree") for the question about correctness.

Table 2. The results regarding general attitudes toward Journalen from patients with mental health issues and all other patients.

Question	Mean_mental health (SD)	Mean_others (SD)	*P* value
I believe that access to medical records online is generally a good reform	4.73 (0.68)	4.80 (0.60)	.01
I believe that access to "*Journalen*" is good for me	4.78 (0.64)	4.86 (0.51)	.005

Table 3. The results regarding information accuracy in Journalen from patients with mental health issues and all other patients.

Question	Mean_mental health (SD)	Mean_others (SD)	*P* value
The content in the record reflects the information I think that health care has about me	3.98 (0.99)	4.22 (0.91)	<.001
There is information about me that is missing in the record which I think should be there and that the staff should know	3.32 (1.34)	2.85 (1.40)	<.001

Accessing Patient Information

Regarding the respondents' answers to why a patient uses Journalen (Q4a-h), the results show that the most common reasons for accessing Journalen, among patients with mental health issues, are to get an overview of the medical history and treatment (Q4b), to follow-up what has been said during a health care visit (Q4e), and to become more involved in the care (Q4h). The same holds true for all other survey respondents. The least common reason for access, for both groups, was to get an overview of relatives' medical history and treatment (Q4c). Results for all these questions are presented in Table 4.

There were 4 reasons for access, where the analysis showed significant differences between patients with mental health issues and the other respondents (Q4a,b,d, and f; also in Table 4; detailed results for all items related to this question (Q4) are presented in Multimedia Appendix 1). Of these 4, to get an overview of medical history and treatment, Q4b got the highest mean score, where patients with mental health issues gave slightly lower ratings (mean_mental health 4.58, SD_mental health 0.78) than other patients (mean_others 4.65, SD_others 0.80; *P*=.001). Still, the results show that patients with mental health issues see this as one of the most important reasons for accessing the PAEHR. On the other hand, compared to the other respondents, patients with mental health issues gave slightly higher, significant, ratings for the following reasons to use: general interest (mean_mental health 3.86, SD_mental health 1.19; mean_others 3.66, SD_others 1.29; *P*=.002), insecurity of whether the care is right (mean_mental health 2.83, SD_mental health 1.39; mean_others 2.62, SD_others 1.37; *P*=.002), and suspicion of inaccuracies (mean_mental health 2.55, SD_mental health 1.34; mean_others 2.31, SD_others 1.30; *P*<.001). The differences are, however, not very large, and none of these 3 were marked as one of the most common reasons for access by any of the groups.

Respondents were also asked to rate items of their importance, in connection to being able to access Journalen (Q5). The respondents in the mental health group rated the following benefits of being able to access Journalen the highest: it makes me feel informed (Q5e), it makes me feel safe (Q5d), and it improves communication between medical staff and me (Q5a). The other respondents' ratings gave similar results aside from that they rated Q5c (it improves the understanding of the condition) as one of the top 3 benefits instead of Q5d. It is also of interest to note that both groups of respondents gave a very low rating to the item it has no relevance (Q5j), indicating that the respondents generally see clear benefits from accessing Journalen. See Table 5 for detailed results.

Table 4. Respondents' answers to the question "Why do you use Journalen?" by patients with mental health issues and all other patients[a].

Question	Mean_mental health (SD)	Mean_others (SD)	*P* value
Mostly general interest	3.86 (1.19)	3.66 (1.29)	.002[b]
To get an overview of my medical history and treatment	4.58 (0.78)[c]	4.65 (0.80)[c]	.001[b]
To get an overview of my relatives' medical history and treatment	2.06 (1.45)	2.21 (1.52)	.05
Because I am not sure if I got the right care	2.83 (1.39)	2.62 (1.37)	.002[b]
To follow up what has been said during a health care visit	4.47 (0.91)[c]	4.39 (0.99)[c]	.10
Because I suspect inaccuracies	2.55 (1.34)	2.31 (1.30)	<.001[b]
To prepare for my health care visit	3.40 (1.35)	3.51 (1.33)	.11
To become more involved in my care	4.21 (1.10)[c]	4.28 (1.02)[c]	.39

[a]The Mann-Whitney *U* test was used for statistical analysis.

[b]Significant *P* values.

[c]The most highly ranked options by mental health respondents and other respondents.

Table 5. Respondents' answers to the question "How important is it for you to be able to access patient information?" by patients with mental health issues and all other patients. Some items related to Q5 are only shown in Multimedia Appendix 1[a].

Question	Mean_mental health (SD)	Mean_others (SD)	*P* value
It improves communication between medical staff and me	4.21 (0.96)[b]	4.36 (0.92)[b]	<.001[c]
It improves the understanding of the condition	4.18 (1.02)	4.26 (0.98)[b]	.12
It makes me feel safe	4.24 (0.99)[b]	4.22 (0.97)	.44
It makes me feel informed	4.59 (0.76)[b]	4.62 (0.72)[b]	.42
It leads to that I can take care of my health better	3.40 (1.16)	3.54 (1.13)	.02[c]
It leads to that I can take care of my relatives health better	2.39 (1.38)	2.52 (1.36)	.04[c]
It is essential that I am able to actively participate in decisions about my or my relatives' health	3.31 (1.44)	3.47 (1.41)	.03[c]
It has no relevance	1.49 (0.96)	1.50 (0.93)	.52

[a]The Mann-Whitney *U* test was used for statistical analysis.

[b]The most highly ranked options by mental health respondents and other respondents.

[c]Significant *P* values.

Some significant differences were however found between the groups (also in Table 5). Patients with mental health issues rated the following benefits of Journalen as of lower importance than did the other patients: improves communication with health care (*P*<.001), enables better self-care (*P*=.02), improves the possibility to take better care of relatives (*P*=.04), and enables the essential possibility to participate in health decisions (*P*=.03). None of the stated items in Q5 were rated of higher importance by patients with mental health issues compared to other patients, with a significant difference between the groups. Among the items where significant differences could be identified, only item Q5a about improved communication was one of the most highly rated benefits of accessing Journalen.

Moreover, respondents were asked how important different information types and functions in Journalen are to them (Q17a-r). The 3 functions or information types that the group of patients with mental health issues rated to be of highest importance to have access to were results of tests (Q17d), being able to read all types of record entries (Q17g), and overview of all health care contacts (Q17e). This rating corresponds well to the ratings from the group of other patients. Both respondent groups gave the lowest ratings to the importance of being able to communicate electronically with other patients (Q17o).

There were 18 items included in this survey question. For most of the items, no significant difference could be found between the 2 groups of respondents. The 6 items for which significant differences *could* be identified are presented in Table 6. Patients with mental health issues gave higher ratings to the importance of the following information types or functions: psychiatry records (*P*<.001), all types of medical notes or record entries (*P*=.02), blocking professionals from access to certain information (*P*<.001), and access to the log list (*P*<.001). Of these, the differences were largest regarding the importance of having access to the psychiatry record (mean_mental health

4.47, SD_mental health 1.09 and mean_others 3.65, SD_others 1.38) and the possibility to block other professionals from having access to all patient information (mean_mental health 3.43, SD_mental health 1.46 and mean_others 3.04, SD_others 1.42), which was expected due to the special needs of this group. Furthermore, patients with mental health issues gave slightly lower ratings regarding the importance of the following information types or functions: referral tracking (*P*=.02) and test results (*P*=.01). It is also noticeable that both studied groups of patients gave generally high ratings to most of the information types and functions included in question Q17.

For 2 of the information types or functions listed in the survey (access to all record entries and access to the log list), the responses from patients with mental health issues who could or could not access their mental health records differed significantly. In the case of access to all types of record entries, those who could access their mental health notes gave significantly higher ratings (*P*=.049), and in the case of the log list, this group of respondents gave significantly lower ratings (*P*<.001).

Table 6. Respondents' answers to the question "How important is it for you to have access to the following information which is wholly or partly based on information contained in "Journalen"?" by patients with mental health issues and all other patients. Some items related to Q17 are only shown in Multimedia Appendix 1[a].

Question	Mean_mental health (SD)	Mean_others (SD)	*P* value
Referral (content and how it is handled in care)	4.50 (0.88)	4.62 (0.74)[b]	.02[c]
Results of tests	4.69 (0.76)[b]	4.78 (0.61)[b]	.01[c]
Overview of all health care contacts	4.59 (0.79)[b]	4.61 (0.77)	.60
Being able to read record entries from psychiatry	4.47 (1.09)	3.65 (1.38)	<.001[c]
Being able to read all types of record entries	4.68 (0.83)[b]	4.64 (0.79)[b]	.02[c]
Ability to communicate electronically with other patients	2.15 (1.36)	2.02 (1.24)	.14
Ability to block certain medical records from access by other medical staff	3.43 (1.46)	3.04 (1.42)	<.001[c]
See which care units and staff groups have been inside "*Journalen*" (see log data)	4.28 (1.14)	4.05 (1.25)	<.001[c]

[a]The Mann-Whitney *U* test was used for statistical analysis.

[b]The most highly ranked options by mental health respondents and other respondents.

[c]Significant *P* values.

Relationship With Health Care and Patient Involvement

Two of the questions in the survey (Q7 and Q16) covered aspects of the patient's relationship with health care and his or her involvement in the care process. Regarding possible changes in the patient's relationship to health care (in general) after using Journalen (Q7a) and to health care professionals (more specific) due to communication about Journalen (Q7b and c) and its content (Q7d), no significant differences could be found between respondents in the 2 groups (Table 7). Here, it is again important to remember that not all of the respondents who answered in the role of patients with mental health issues, as of the date of the survey, had access to his or her psychiatry record. The results however show that all patients, regardless of group, experience at least a moderate positive effect on the relationship with health care. Furthermore, patients (regardless of group) and health care professionals generally do not talk about the possibility for the patient to use Journalen nor do they discuss its content.

Regarding patient involvement in the care process, some significant differences were found between the 2 groups (Table 8). Patients with mental health issues gave significantly lower ratings when it came to Journalen's potential to support communication with medical staff (*P*=.02) and its potential to enable shared decision-making (*P*=.02). No significant differences were found regarding support for following prescription of treatment (*P*=.053) or support for self-care (*P*=.69). Overall, the results indicate that Journalen had at least a moderate positive effect in the involvement in the care process for patients with mental health issues, and the same holds true for the other respondents as well.

Table 7. Respondents' answers to the question "To what extent do you agree with the following statements regarding your relationship with health care?" by patients with mental health issues and all other patients[a].

Question	Mean_mental health (SD)	Mean_others (SD)	*P* value
To take part of the patient information via "*Journalen*" has affected the relationship with health care system positively	3.80 (1.11)	3.88 (1.07)	.13
Medical staff has informed me about the possibility to read "*Journalen*"	1.82 (1.29)	1.87 (1.24)	.17
Medical staff has encouraged me to use the "*Journalen*"	1.70 (1.12)	1.72 (1.09)	.40
I discuss the content of "*Journalen*" with medical staff	2.53 (1.49)	2.52 (1.41)	.95

[a]The Mann-Whitney *U* test was used for statistical analysis.

Table 8. Respondents' answers to the question "How important is "Journalen" to make you feel that you are involved in your own care?" by patients with mental health issues and all other patients[a].

Question	Mean_mental health (SD)	Mean_others (SD)	*P* value
Information in "*Journalen*" has helped me in communication with medical staff	3.57 (1.19)	3.71 (1.13)	.02
Information in "*Journalen*" had a positive impact on the ability to work together with medical staff making decisions about care and treatment	3.38 (1.21)	3.52 (1.18)	.02
Information in "*Journalen*" had a positive impact on the ability to follow the prescription of treatment	3.71 (1.22)	3.83 (1.16)	.05
Information in "*Journalen*" had a positive impact on the ability to take own steps to improve health	3.56 (1.22)	3.60 (1.18)	.69

[a]The Mann-Whitney *U* test was used for statistical analysis.

Discussion

Principal Findings

The aim of this study was to, through a national patient survey, investigate the experiences of patients with mental health issues with the Swedish PAEHR Journalen, as well as possible differences between patients with mental health issues and other patients, related to experiences with and attitudes toward the eHealth service. The paper contributes most and foremost to a much-needed knowledge about the effects of Journalen for a specific patient group—patients with mental health issues—several years after the launch of the service. Several important conclusions about aspects that patients with mental health issues value with regard to Journalen were identified in this study, and some interesting differences between the groups of patients with mental health issues and all other patients were brought to light in the comparative analysis. The results also reveal that, in most cases, patients with mental health issues see the same values in Journalen as other patients.

First, and on an overall level, it is clear from the results of the survey that respondents in the mental health group, as well as all other respondents, were positive toward being able to access personal health information in Journalen, and that there are no big differences between patients with mental health issues and other patients. These results are in accordance with earlier research [9,12]. These results are important, as health care professionals have raised concerns that patients with mental health issues in particular would become confused and agitated from reading their PAEHR [6,7]. Moreover, the results reveal that the group of patients with mental health issues is somewhat more critical toward the accuracy of the content compared to other respondents. A possible explanation for why more patients with mental health issues find inaccuracies could be that mental health conditions are more subjective and difficult to quantify and therefore may give rise to disagreements in how they should be described and documented. Since patients with mental health issues find more inaccuracies in the record, and since current research [6,17] has reported that some clinicians change the way they document as a result of patients reading their notes, it is of utmost importance that we open up a discussion regarding how notes could or should be written, and if, how, and when the patients should be involved.

Second, regarding the reasons for using the service as well as what information types were considered to be important to patients, there were no big differences between the 2 respondent groups. The reasons for use that were rated highest among mental health respondents are using Journalen for receiving an overview of one's treatments, following up on what was said during a health care visit, and becoming more involved in the care process. This should be seen as an important result, since it shows that the reasons for implementing Journalen in the first place are as relevant for patients with mental health issues as they are for all other patient groups represented in the survey. Nevertheless, this patient group has been treated differently during the implementation process in that, for example, mental health notes were excluded in all regions during the first 3 years and were only accessible in 2 regions 6 years after launch. Results from the comparison between mental health respondents who could access their mental health notes and those who could not, show, interestingly enough, that there are no significant differences for any of the results related to the reasons for using Journalen. Hence, this study does not show any indication that access to these particular notes makes a difference in the reasons for using Journalen or the attitudes that patients with mental health issues have toward it.

There were some significant differences indicating specific needs that are related to patients with mental health issues, but most of these differences are still small. Hence, even though there were statistically significant differences between the groups, most of the results do not support that there, in practice, would be any big differences between patients with mental health issues and other patients. Patients with mental health issues gave a somewhat higher rating for possibilities of reviewing which of the health care professionals have read the content in the EHR (access to the log list) and for blocking professionals from other health specialties from accessing all information. They also reported somewhat higher feelings of insecurity regarding having received the right care.

The results showed that some patients might use Journalen because of insecurity about receiving the right care. This can be related to some of the concerns raised by mental health professionals. Both Petersson and Erlingsdóttir [7], from the Swedish perspective, and Dobscha et al [6], from the US perspective, reported the general concern that patients would request changes in the health record both due to found inaccuracies and notes that can be considered sensitive and that the patient might not agree with. These studies, from the perspective of health care professionals, report on health care concerns regarding the consequences of patients' access to mental health notes. This survey's results show that patients from this group indeed use Journalen to check whether they have received the right care. The perceived importance of the blocking feature in the system as well as the log list of who has accessed the record points toward an insecurity in who can access the information. This could possibly be due to the sensitive nature of the information related to mental health.

Third, when it comes to communication with health care, no big differences between patients with mental health issues and all other patients could be observed. Respondents with mental health conditions, as well as all other respondents, were generally positive regarding the effects on communication with health care, which is in line with existing research [22], but they gave lower ratings when it comes to communicating with health care professionals about the existence of Journalen. The fact that health care professionals generally do not inform patients about Journalen or encourage them to use the service has also been reported in earlier research but then concerning patients with cancer [24]. A reasonable interpretation of this neglect is that the earlier-mentioned concerns raised by health care professionals function as an obstacle.

Finally, with regard to effects on involvement in care, the results were also similar between the 2 groups. However, patients with mental health issues gave significantly lower ratings to actual effects on the relationship with health care and regarding shared decision-making. These results could possibly be related to the concerns on the effects of the therapeutic alliance that mental health professionals have raised [17]. In contrast to these results, a previous study showed that patients with cancer gave significantly higher ratings than all other patient groups and on all items regarding the effects of involvement in care [24]. Earlier studies, not focusing on specific patient groups [25], have shown that patients' web-based access to medical records has improved the possibility for patients to engage in shared decision-making, something that Rexhepi et al [24] also showed for patients with cancer in Sweden. Similar studies of patients with mental health issues are very few to date. Petersson and Erlingsdóttir [18] showed that most of their answering professionals did *not* experience a higher patient involvement. This survey indicates that patients with mental health issues, regarding participation in decisions related to their care, do not experience the same positive effect as other patient groups. This conclusion is thus in line with the health professionals' view [18]. This issue is clearly worthy of additional exploration.

Limitations

As already mentioned, at the time of the survey, only patients from the regions Skåne and Kronoberg could access mental health notes, since the introduction in 2013. Consequently, some of the patients with mental health issues who answered the survey could not yet access the mental health notes, while others in fact could. All patients with mental health issues could, however, access information on somatic care. An additional Mann-Whitney *U* test was used to compare answers between patients with mental health issues who could and could not access mental health notes. For most of the areas covered in this paper, no significant differences could be found between these 2 groups. In light of this study limitation, when it comes to actual experiences, the study is focusing more on experiences of the PAEHR among patients with mental health issues in general than mental health notes in particular. However, regarding attitudes, the study captures the ideas of patients with mental health issues regarding access to his or her mental health notes. It is also important to recognize that answers were gathered through self-report. A unique patient might have had contact with health care for numerous reasons. We cannot be sure whether there are patients with mental health issues in the "others" group who did not disclose their mental health status in the survey. Moreover, one could discuss how homogenous the mental health group is, given the diversity in both type of diagnosis and severity in symptoms that might be visible within the group. The survey did not capture this diversity.

Conclusions

The study was based on a national patient survey where 19.5% (n=504) of respondents indicated that they experienced a mental health disease. The objectives of the paper were to examine the use, attitudes, and experiences of patients with mental health issues by reading their notes in the PAEHR and, moreover, whether their experiences differ from other patient groups, and if so, how.

A first conclusion, on an overall level, is that patients with mental health issues are as positive in their attitudes toward the access of personal health information in Journalen as other patient groups. This conclusion agrees with previous research. A second conclusion is that patients with mental health issues use Journalen differently in 2 manners, as compared to other patient groups: they check the record for inaccuracies regarding care and information content and they tend to block access to mental health notes for professionals from other parts of the health care system. These differences in usage were not known from previous research. A third conclusion is that patients with mental health issues have somewhat other experiences from Journalen than other patient groups, in that they are less likely to find it supportive of shared decision-making between themselves and their doctor. This was not known from previous research. A final conclusion is that the clinicians' worry about possible harm to this patient group does not find support by the current empirical evidence. Patients with mental health issues with access to their mental health notes reported the same positive attitudes toward Journalen as did patients with mental health issues with only access to their somatic health notes.

There are many previous studies on how patients access PAEHR and their attitudes to the introduction of such eHealth services. This study contributes with knowledge, through its comparative research design, on how PAEHRs are perceived by patients with mental health issues as compared to other patients. Further research is however needed in this area. For example, the study contributed with insights regarding different usage patterns. It would be valuable with empirical insights and explanations of why patients with mental health issues are somewhat more critical regarding the accuracy of the information content. From the perspective of professionals, previous research has predicted that patient access to mental health notes will have consequences on what and how the professionals write the notes. This study indicates interesting paths for further investigation of that issue. A practical implication is, however, that professionals do not need to be overly concerned about potential harm to the patients. Patients with mental health issues use Journalen and its information by the same reasons as other patient groups. Patients with mental health issues are as positive to the effects on communication with health care as other patient groups, which is in line with previous research.

Patients with mental health issues are a vulnerable group, where professionals anticipate that patients may get confused, judged, worried, or angry when reading their notes. This study did not find support for that. Other studies have reported benefits from accessing the mental health notes, such as feelings of increased engagement, control over their health, and trust toward the professionals. Finally, this study contributes with the insight that the group of patients with mental health issues finds it less possible to engage in shared decision-making as compared to other patient groups. Further research could help us better understand why we need to know more about the obstacles to patient participation and how Journalen can be used to better address practical issues related to feelings of engagement, control, and trust.

Acknowledgments

The authors would like to thank the colleagues in the national Development of Online Medical Records and E-Health Services consortium who participated in the data collection. The authors would also like to thank INERA and Evry for distributing the survey through the national patient portal.

Authors' Contributions

JM led the work and analyzed all the data. JM and GM did most of the study design, and all authors contributed equally in writing and editing the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Detailed results from all the survey questions covered in this study.

[DOCX File , 53 KB - mental_v11i1e48008_app1.docx]

References

1. Wass S, Vimarlund V. The role of PAEHRs in patient involvement. J Med Syst 2018;42(11):210 [FREE Full text] [doi: 10.1007/s10916-018-1070-y] [Medline: 30255404]
2. Bell SK, Folcarelli P, Fossa A, Gerard M, Harper M, Leveille S, et al. Tackling ambulatory safety risks through patient engagement: what 10,000 patients and families say about safety-related knowledge, behaviors, and attitudes after reading visit notes. J Patient Saf 2021;17(8):e791-e799. [doi: 10.1097/PTS.0000000000000494] [Medline: 29781979]
3. Rexhepi H, Åhlfeldt RM, Cajander Å, Huvila I. Cancer patients' attitudes and experiences of online medical records. 2015 Presented at: Proceedings of the 17th International Symposium on Health Information Management Research (ISHIMR); June 25-26; York, United Kingdom.
4. Grünloh C, Myreteg G, Cajander Å, Rexhepi H. "Why do they need to check me?" Patient participation through eHealth and the doctor-patient relationship: qualitative study. J Med Internet Res 2018;20(1):e11 [FREE Full text] [doi: 10.2196/jmir.8444] [Medline: 29335237]
5. Peck P, Torous J, Shanahan M, Fossa A, Greenberg W. Patient access to electronic psychiatric records: a pilot study. Health Policy and Technol 2017;6(3):309-315. [doi: 10.1016/j.hlpt.2017.06.003]

6. Dobscha SK, Denneson LM, Jacobson LE, Williams HB, Cromer R, Woods S. VA mental health clinician experiences and attitudes toward OpenNotes. Gen Hosp Psychiatry 2016;38:89-93. [doi: 10.1016/j.genhosppsych.2015.08.001] [Medline: 26380876]
7. Petersson L, Erlingsdóttir G. Open notes in Swedish psychiatric care (part 1): survey among psychiatric care professionals. JMIR Ment Health 2018;5(1):e11 [FREE Full text] [doi: 10.2196/mental.9140] [Medline: 29396386]
8. Salmi L, Brudnicki S, Isono M, Riggare S, Rodriquez C, Schaper LK, et al. Six countries, six individuals: resourceful patients navigating medical records in Australia, Canada, Chile, Japan, Sweden and the USA. BMJ Open 2020;10(9):e037016 [FREE Full text] [doi: 10.1136/bmjopen-2020-037016] [Medline: 32933961]
9. Cromer R, Denneson LM, Pisciotta M, Williams H, Woods S, Dobscha SK. Trust in mental health clinicians among patients who access clinical notes online. Psychiatr Serv 2017;68(5):520-523 [FREE Full text] [doi: 10.1176/appi.ps.201600168] [Medline: 28142383]
10. Pisciotta M, Denneson LM, Williams HB, Woods S, Tuepker A, Dobscha SK. Providing mental health care in the context of online mental health notes: advice from patients and mental health clinicians. J Ment Health 2019;28(1):64-70. [doi: 10.1080/09638237.2018.1521924] [Medline: 30468100]
11. DesRoches CM, Bell SK, Dong Z, Elmore J, Fernandez L, Fitzgerald P, et al. Patients managing medications and reading their visit notes: a survey of OpenNotes participants. Ann Intern Med 2019 Jul 02;171(1):69-71. [doi: 10.7326/M18-3197] [Medline: 31132794]
12. O'Neill S, Chimowitz H, Leveille S, Walker J. Embracing the new age of transparency: mental health patients reading their psychotherapy notes online. J Ment Health 2019;28(5):527-535. [doi: 10.1080/09638237.2019.1644490] [Medline: 31364902]
13. Klein JW, Peacock S, Tsui JI, O'Neill SF, DesRoches CM, Elmore JG. Perceptions of primary care notes by patients with mental health diagnoses. Ann Fam Med 2018;16(4):343-345 [FREE Full text] [doi: 10.1370/afm.2287] [Medline: 29987083]
14. Denneson LM, Chen JI, Pisciotta M, Tuepker A, Dobscha SK. Patients' positive and negative responses to reading mental health clinical notes online. Psychiatr Serv 2018;69(5):593-596 [FREE Full text] [doi: 10.1176/appi.ps.201700353] [Medline: 29493408]
15. Denneson LM, Pisciotta M, Hooker ER, Trevino A, Dobscha SK. Impacts of a web-based educational program for veterans who read their mental health notes online. J Am Med Inform Assoc 2019;26(1):3-8 [FREE Full text] [doi: 10.1093/jamia/ocy134] [Medline: 30445648]
16. Bärkås A, Scandurra I, Rexhepi H, Blease C, Cajander Å, Hägglund M. Patients' access to their psychiatric notes: current policies and practices in Sweden. Int J Environ Res Public Health 2021;18(17):9140 [FREE Full text] [doi: 10.3390/ijerph18179140] [Medline: 34501730]
17. Denneson LM, Cromer R, Williams HB, Pisciotta M, Dobscha SK. A qualitative analysis of how online access to mental health notes is changing clinician perceptions of power and the therapeutic relationship. J Med Internet Res 2017;19(6):e208 [FREE Full text] [doi: 10.2196/jmir.6915] [Medline: 28615152]
18. Petersson L, Erlingsdóttir G. Open notes in Swedish psychiatric care (part 2): survey among psychiatric care professionals. JMIR Ment Health 2018;5(2):e10521 [FREE Full text] [doi: 10.2196/10521] [Medline: 29929946]
19. Kahn MW, Bell SK, Walker J, Delbanco T. A piece of my mind. Let's show patients their mental health records. JAMA 2014;311(13):1291-1292. [doi: 10.1001/jama.2014.1824] [Medline: 24691603]
20. Mental disorders. World Health Organization. 2022. URL: https://www.who.int/news-room/fact-sheets/detail/mental-disorders [accessed 2023-04-04]
21. Statistik om psykisk ohälsa i Sverige [Statistics on mental illness in Sweden]. Folkhälsomyndigheten [Public Health Agency of Sweden]. 2022. URL: https://www.folkhalsomyndigheten.se/livsvillkor-levnadsvanor/psykisk-halsa-och-suicidprevention/statistik-psykisk-halsa/ [accessed 2023-04-04]
22. Moll J, Rexhepi H, Cajander Å, Grünloh C, Huvila I, Hägglund M, et al. Patients' experiences of accessing their electronic health records: national patient survey in Sweden. J Med Internet Res 2018;20(11):e278 [FREE Full text] [doi: 10.2196/jmir.9492] [Medline: 30389647]
23. Psykiatrin i siffror. Vuxenpsykiatri [Psychiatric care in numbers. Adult psychiatry]. Uppdrag psykisk hälsa [Mission mental health]. 2022. URL: https://www.uppdragpsykiskhalsa.se/wp-content/uploads/2022/05/Psykiatrin_i_siffror_VUP_2021_tillganglighetsanpassad.pdf [accessed 2023-10-17]
24. Rexhepi H, Moll J, Huvila I. Online electronic healthcare records: comparing the views of cancer patients and others. Health Informatics J 2020;26(4):2915-2929 [FREE Full text] [doi: 10.1177/1460458220944727] [Medline: 32914701]
25. Fossa AJ, Bell SK, DesRoches C. OpenNotes and shared decision making: a growing practice in clinical transparency and how it can support patient-centered care. J Am Med Inform Assoc 2018;25(9):1153-1159 [FREE Full text] [doi: 10.1093/jamia/ocy083] [Medline: 29982659]

Abbreviations

EHR: electronic health record
PAEHR: patient accessible electronic health record

Edited by J Torous; submitted 08.04.23; peer-reviewed by T Scott Duncan, E Matthews; comments to author 31.05.23; revised version received 20.10.23; accepted 29.10.23; published 02.02.24.

Please cite as:
Moll J, Myreteg G, Rexhepi H
Experiences of Patients With Mental Health Issues Having Web-Based Access to Their Records: National Patient Survey
JMIR Ment Health 2024;11:e48008
URL: https://mental.jmir.org/2024/1/e48008
doi:10.2196/48008
PMID:38306158

Original Paper

Impact of Digital Inclusion Initiative to Facilitate Access to Mental Health Services: Service User Interview Study

Amy Oliver[1,2], BA, MSc; Ella Chandler[1], BSc; Julia A Gillard[1,3], BSc, MSc, PhD, DClinPsy

[1]Camden and Islington NHS Foundation Trust, London, United Kingdom

[2]Division of Psychology and Language Sciences, University College London, London, United Kingdom

[3]School of Psychology and Sport Science, Anglia Ruskin University, Cambridge, United Kingdom

Corresponding Author:
Julia A Gillard, BSc, MSc, PhD, DClinPsy
School of Psychology and Sport Science
Anglia Ruskin University
East Road
Cambridge, CB1 1PT
United Kingdom
Phone: 44 (0)1245 493131
Email: julia.gillard@aru.ac.uk

Abstract

Background: Digital exclusion, characterized by a lack of access to digital technology, connectivity, or digital skills, disproportionally affects marginalized groups. An important domain impacted by digital exclusion is access to health care. During COVID-19, health care services had to restrict face-to-face contact to limit the spread of the virus. The subsequent shift toward remote delivery of mental health care exacerbated the digital divide, with limited access to remote mental health care delivery. In response, Camden and Islington National Health Service Foundation Trust launched the innovative Digital Inclusion Scheme (DIS).

Objective: This study aimed to examine the impact of facilitating digital inclusion in mental health access. Camden and Islington National Health Service Foundation Trust implemented the trust-wide DIS for service users who were digitally excluded, that is, were without devices or connectivity or reported poor digital skills. The scheme provided access to a loan digital device (a tablet), internet connectivity devices, and mobile data, as well as personalized digital skills support.

Methods: The DIS went live in October 2021 and received 106 referrals by June 2022. Semistructured interviews were conducted with 12 service users to ask about their experience of accessing the DIS. A thematic analysis identified themes and subthemes relating to the extent of their digital exclusion before engaging with the scheme and the impact of accessing a scheme on their ability to engage with digital technology and well-being.

Results: There were 10 major themes. A total of 6 themes were related to factors impacting the engagement with the scheme, including digital exclusion, relationship to the trust, the importance of personalized digital support, partnership working, device usability and accessibility, and personal circumstances. The remaining 4 themes spoke to the impact of accessing the scheme, including improved access to services, impact on well-being, financial implications, and a greater sense of empowerment.

Conclusions: Participants reported an increased reliance on technology driving the need for digital inclusion; however, differences in motivation for engaging with the scheme were noted, as well as potential barriers, including lack of awareness, disability, and age. Overall, the experience of accessing the DIS was reported as positive, with participants feeling supported to access the digital world. The consequences of engaging with the scheme included greater perceived access to and control of physical and mental health care, improved well-being, and a greater sense of empowerment. An overview of the lessons learned are provided along with suggestions for other health care settings that are looking to implement similar schemes.

(JMIR Ment Health 2024;11:e51315) doi:10.2196/51315

KEYWORDS

digital exclusion; digital inclusion; video consultation; COVID-19; tablet loan scheme; mental health; telemedicine; digital divide; digital inequality; technology

Introduction

Background

The emergence of COVID-19 in the United Kingdom led to a rapid transformation in the provision of health services, as it became necessary to reduce face-to-face contact to limit the spread of the virus [1]. As the general population experienced a marked and a prolonged deterioration in mental health, remote care and telehealth practices were rapidly adopted by health care systems to ensure continuity of care [2-4]. Disruption to community and third-sector services further proved detrimental to already at-risk populations, including those with preexisting mental health conditions and severe mental illness or those affected by financial destitution and homelessness [5,6]. As such, the integration of digital solutions into health care became inevitable, with advocates noting greater flexibility to meet individual needs and digital transformation as key to implementing the National Health Service (NHS) long-term transformation strategy [7-9].

However, concerns were raised by practitioners and service users as to the feasibility of engaging in telehealth interventions or other forms of digital health care delivery. These included the ability of staff and service users to navigate devices and platforms, the lack of physical or financial resources to obtain technology, poor connectivity, concerns around cybersecurity, and physical privacy when providing or receiving telehealth care in a home environment [1,8,10,11]. Digital exclusion, characterized by lack of access, connectivity, or digital skills, is another domain that presents a challenge to the remote delivery of health care with marginalized groups particularly affected [12]. This includes older adults [3], homeless people [13], immigrants [14], people with disabilities [12], and individuals with severe mental illness (eg, psychosis) [15].

The term *digital divide* conceptualizes this inequality, differentiating between those who have access and those who do not [16]. It consists of 3 levels [17]: *lack of access* to information and communications technology (ICT), for example, computer, tablet or mobile phone devices, or data; *lack of digital skills or confidence*; and *lack of awareness, or interest*, in the opportunities digital technology can provide [18]. The 3 strands of the digital divide are not linear, but rather they are interactional and reciprocal; one may have a high level of digital ability; however, without an ICT device, one cannot use these skills, while the likelihood of being digitally excluded is exacerbated by societal inequities, and the lack of access can itself perpetuate marginalization [19-21]. The relationship between digital exclusion and preexisting inequalities thus limits individuals' ability to participate in the digital world; can lead to greater social isolation [20]; and prohibit access to education, training, and employment opportunities [22] or, critically, access to mental and physical health services [13].

With the shift toward remote delivery of mental health services exacerbating the digital divide in the wake of COVID-19, the Camden and Islington NHS Foundation Trust (C&I NHS FT) launched the innovative Digital Inclusion Scheme (DIS) to support digitally excluded service users accessing community mental health services. The scheme provided access to a loan ICT device (a tablet), internet connectivity devices, and mobile data, as well as personalized digital skills support to enhance literacy and confidence. This scheme aimed to build on lessons learned from previous initiatives, which had predominantly focused on increasing digital skills and competency to bridge the digital divide [3,11,21,23,24]. However, the lack of physical access to resources remained an immediate concern for those affected by digital exclusion [25]. Similarly, interventions restricted in time (eg, NAViGO) [12] or limited in reach, for example, only available for a small number of inpatients [4]. Interventions also tended to support organizations rather than individuals (eg, Digital Communities Wales) [26] or link together various community organizations rather than providing holistic support, for example, 100% Digital Leeds [27]. Pertinent was also the marked absence of robust evaluation regarding the impact of such schemes.

Objective

Following its launch in October 2021, this evaluation therefore aimed to understand the impact of the DIS on service users, as well as to identify the barriers and facilitators to implementing such schemes in future. Specifically, it aimed to (1) summarize the project and communicate the lessons learned to date, (2) understand the barriers and facilitators to accessing digital technology within a DIS within a population of mental health service users, and (3) explore the impact on service users who used the scheme and any impact it had on their well-being.

It is anticipated this evaluation will inform clinicians, service planners, and commissioners as to whether this type of scheme is feasible and acceptable in the context of an NHS trust and provide an example design to those both within the NHS and wider community groups.

Methods

Design

This evaluation used an intervention mixed methods framework, combining quantitative and demographic data from the referral form and baseline questionnaire (part 1) and semistructured qualitative interviews to understand the impact of the scheme (part 2). This evaluation encompasses referrals received between October 2021 and June 2022.

Setting

C&I NHS FT is a large inner-city NHS trust, which provides mental health and substance misuse services. Due to the COVID-19 pandemic and social distancing requirements, most NHS services transitioned to delivering treatment remotely, leaving many service users without continued access to care. In response, C&I NHS FT established a DIS to support service users at risk of digital exclusion. The scheme went live in October 2021 and received 106 referrals by June 2022.

Initiative

The DIS provided a response to the 3 key areas of digital exclusion: provision of devices; connectivity; and support to build skills, confidence, and motivation to engage with digital technology. To this end, the trust purchased 185 Samsung Galaxy Tablets, of which 25 supported 4G connectivity in addition to being Wi-Fi enabled. Additionally, 100 SIM cards, each with 20 GBs of free data for 6 months, were obtained from Vodafone as part of the "charities.connected" initiative [28].

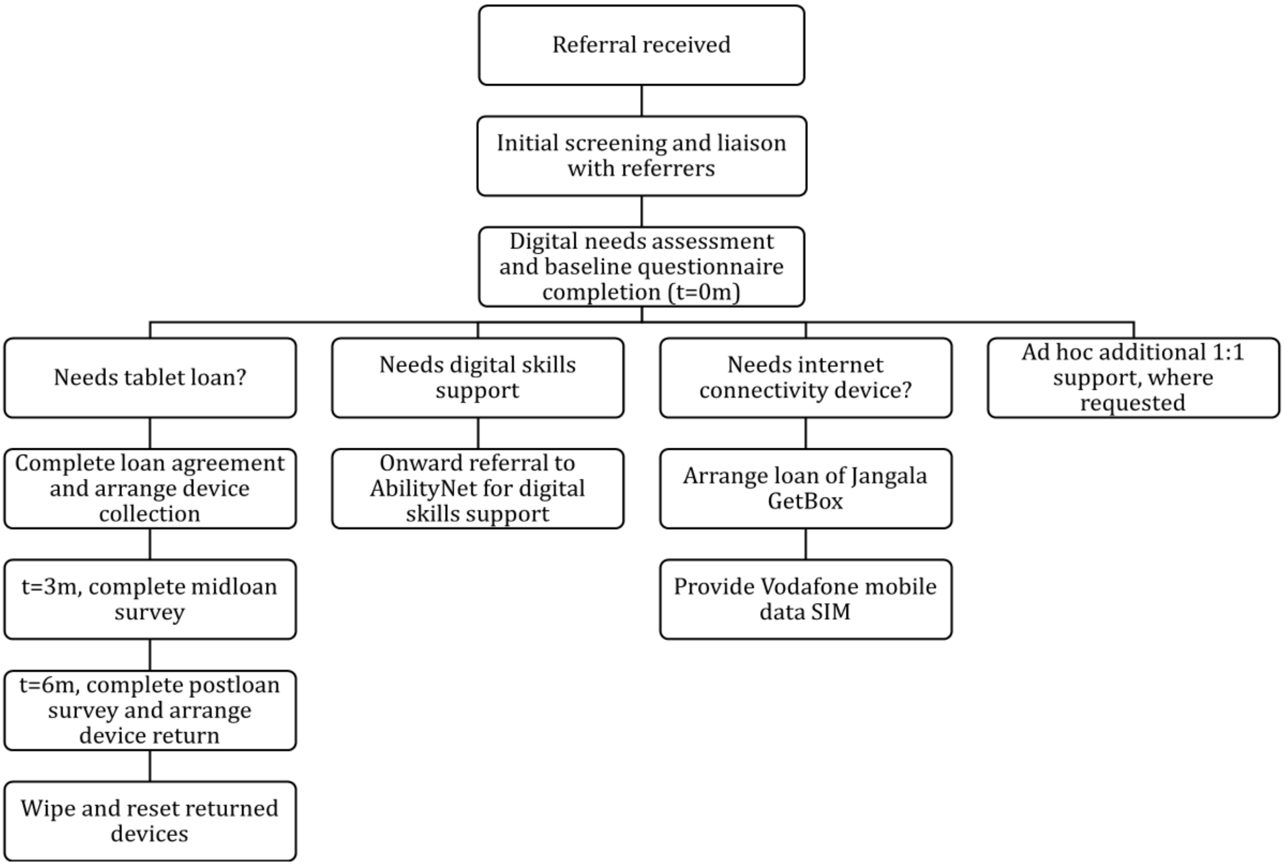
The project also extended preexisting partnerships with the UK-based not-for-profit Jangala [29] and AbilityNet, a UK charity [30]. Jangala loaned C&I NHS FT 30 Get Boxes, which is a mobile router that operated using a SIM card and provided simultaneous connectivity for up to 10 devices. AbilityNet used their web-based resources and network of community-based volunteers to provide 1:1 tailored digital support. The trust also invested in 185 protective cases and 20 Bluetooth keyboards to support accessibility needs.

A digital inclusion officer (DIO) was employed by the trust on a 0.4 full time equivalent basis to facilitate the scheme. The DIO was embedded within a clinical team and line managed by a clinical psychologist who could provide risk and safeguarding advice where needed. A digital workspace was developed on ServiceNow (ServiceNow, Inc) [31], which was integrated into the wider IT infrastructure. This was used by the DIO to administer the scheme, operating as both a stock inventory and log of all loans, as well as providing a digital workflow to track the loan process.

Pathway of Support

Service users were referred via a web-based form by the C&I staff members and screened by the DIO for eligibility and suitability (see Figure 1 for referral pathway). To be eligible for the scheme, service users had to be (1) currently accessing mental health services within C&I NHS Foundation Trust and (2) referred by a mental health professional or key worker who worked within the trust. In addition, they had to be at risk of digital exclusion, experiencing (1) lack of access to digital devices, (2) lack of internet connectivity, and (3) lack of skills or confidence in accessing digital technology.

Figure 1. Digital inclusion referral pathway.



Following the receipt of the referral, the DIO would screen the referral for any risk information and liaise with the referrer before contacting the service user to complete an initial telephone or in-person digital needs assessment. The digital needs assessment involved completing a baseline questionnaire (Multimedia Appendix 1) to understand individuals' digital skills and confidence in accessing digital technology. The DIO would also clarify service users' current digital needs and confirm eligibility for the scheme. Once service users opted into engaging with the scheme, including the receipt of a tablet loan, the DIO would discuss the practicalities of the loan, as well as the terms of the loan agreement, which detailed that service users were able and encouraged to use the devices for personal use under the conditions of the loan agreement.

The DIO then provided the service user with a tablet loan, a connectivity device, or free mobile data SIMs and facilitated an onward referral to the partner organization AbilityNet for digital skills support. The DIO would also provide further ad hoc support around digital skills training or other digital resources. The DIO would meet service users within a familiar environment (ie, home visits) or within the trust premises, considering risk and safeguarding concerns, where identified.

Over the course of the loan, service users were contacted by telephone or email on 2 occasions to complete a midloan survey after 3 months and a postloan survey after 6 months, unless referrers or service users requested a loan extension. At the end of the loan period, the DIO would liaise with the service user and referrer to arrange the return of the device and assess further needs. Throughout their engagement with the scheme, service users were also encouraged to bring any concerns or questions about their devices to the DIO for ad hoc support. The DIO could also facilitate a referral for additional digital skills training through the partnership with AbilityNet at any point throughout the loan period. Once the device had been returned, the DIO would liaise with trust IT services to wipe and reset the device for a new loan cycle.

Participants

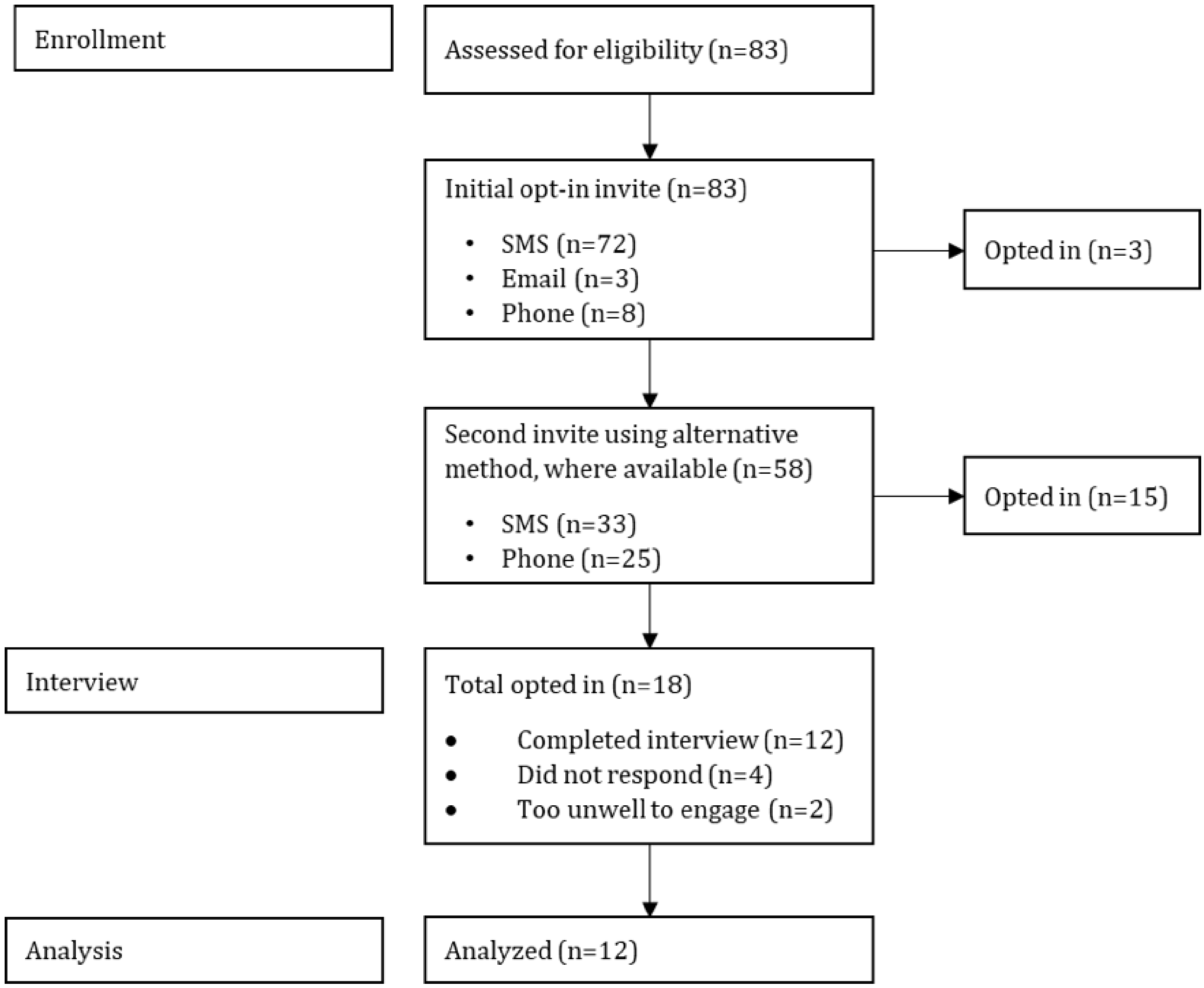
Part 1

In total, 106 service users were referred to the scheme between October 2021 and June 2022. Service users were considered to have engaged with the scheme if they had completed a baseline questionnaire (Multimedia Appendix 1) following the referral. Of the 106 service users referred, 83 (78.3%) completed the baseline questionnaire, and the remainder 23 (21.7%) either did not wish to take part in the scheme or did not engage, for example, not responding to phone calls or SMS text messages. The inclusion criteria were being a current user of mental health services within C&I NHS Foundation Trust, over the age of 18 years, and at risk of digital exclusion.

Part 2

The study used convenience sampling of individuals who had prior engagement with the DIS by completing, at minimum, the baseline questionnaire during the initial digital needs assessment. As such, 83 service users were eligible to take part in the formative service evaluation and were invited. A total of 12 participants were recruited via SMS text message, email, or telephone and were subsequently interviewed (Figure 2). Recruitment ceased once data saturation had been reached. Efforts were made to invite service users who chose not to engage with the scheme, and all the participants were informed that their contribution to the service evaluation would have no bearing on their current or future engagement with the scheme, which aimed to reduce the risk of bias.

Figure 2. Recruitment pathway.



Measures

Part 1: Baseline Questionnaire

The baseline questionnaire was derived from *The Digital Inclusion Evaluation Toolkit: Bank of Outcomes, Indicators and Survey Questions* [32] and consisted of 10 items, including questions such as "Do you consider your digital skills to be good, average, or poor?" "How confident are you in using digital technology?" and "Is there anything that limits you from using the internet at the present moment?" (Multimedia Appendix 1).

The questionnaire aimed to ascertain service users' self-reported level of digital skills and confidence in accessing and using digital technology, as well as their frequency and type of use. It also provided insight into what factors led to service users being digitally excluded by reporting what limited their use of the internet and whether COVID-19 had made accessing the internet and using digital technology more important. It also established what resources they required from the scheme to meet their digital needs. service users were able to indicate whether they needed assistance with acquiring equipment, accessing data connectivity, and support with using digital technology and digital resources on the internet.

Part 2: Interview Schedule

A semistructured interview schedule (Multimedia Appendix 2) was created to obtain qualitative data to understand the impact of the DIS. Each interview lasted between 30 and 40 minutes and was conducted by the same researcher (AO) to minimize interrater variability [33]. The questions were based on the initial aims and perceived outcomes of the DIS and preexisting literature [34] while also following guidance on good practice for service evaluations [35] and ensuring any potential ethical issues were considered [36]. Interview topics covered factors contributing to digital exclusion, questions relating to primary outcomes, including eliciting the perceived barriers and facilitators to engagement with the scheme (eg, "What helped or hindered you in engaging with the support offered by the scheme?" and "What changes would you make to improve the scheme?"), as well as questions relating to their experience of engaging with partner organizations, such as AbilityNet or Jangala. The interviews also attempted to elicit potential secondary outcomes, including but not limited to the impact of engagement on well-being, social interactions, ability to access C&I NHS FT or other health services, leisure, finances, education, and employment opportunities. Where applicable, participants were also asked about their experience of accessing the partnership support available, such as that offered by AbilityNet.

Procedure

Part 1

Baseline demographic, digital needs, and digital literacy information was obtained through the referral form and baseline questionnaire, which was carried out by the DIO with the service users' consent, upon referral to the scheme.

The Initiative section provides further details on the baseline digital needs assessment.

Part 2

A link to an evaluation opt in form was sent to all eligible service users if a mobile number had been provided in the initial referral. An email was sent to those who did not provide a telephone contact and for those with only a landline number, and the invitation to the those who chose to participate were contacted by their preferred method, whereby the project was explained, and interviews were scheduled.

Before the interview, participants were provided an information sheet (Multimedia Appendix 3). Further demographic information was obtained after the main interview, and participants were given opportunities to ask any questions. The information sheet also included guidance on further support available should any questions or concerns arise after the session.

The interviews were recorded using the Microsoft Word voice recorder. Interviews were automatically transcribed verbatim by Microsoft Word and anonymized by removing personally identifiable content. The final transcription was checked against the original audio recording for errors and corrected manually by the first author (AO).

Statistical Data Analysis

Part 1

The baseline questionnaire contained self-reported data of participants' digital skills, confidence, and literacy, as well as current engagement with digital technology. Quantitative data from the referral form and baseline questionnaire was analyzed using SPSS Statistics (version 28.0.0.0; IBM Corp) and JASP (version 0.16.3; JASP Team, 2022) using descriptive statistics.

Part 2

Thematic analysis [37] was chosen to explore the qualitative experience of service users accessing the DIS from a realist approach [38]. Each interview was transcribed by the main researcher and checked by another (JAG). Each transcript was reviewed before coding the data using a combination of inductive and deductive approaches [37], deriving codes from the data within the primary areas of interest. Transcripts were repeatedly checked against the emergent themes before a list of themes and subthemes were collated through discussion between 2 authors (AO and JAG). The analysis was conducted using NVivo 12.7.0 software (QSR International) before the results were synthesized into a thematic analysis.

Ethical Considerations

Ethical approval was obtained from the Anglia Ruskin University School Research Ethics Panel (ETH2122-0635; July 29, 2022) and the C&I NHS FT Clinical and Information Governance Department (July 12, 2022). Informed consent was obtained from the participants before the interview (Multimedia Appendix 4)

Results

Part 1: Baseline Digital Needs and Demographics

Of the 83 service users who completed the baseline questionnaire, 52% (n=43) were female, and 48% (n=40) were male, with a mean age of 52.42 (SD 13.47) years.

Almost all reported requiring a device (n=78, 94%), with approximately half reporting a need for internet connectivity (n=45, 54%) or digital skills support (n=57, 69%). One-third of referrals (n=31, 37%) indicated requiring support across all domains of need.

Of those reporting a lack of devices, 77 (93%) were provided with a loan tablet; 40 (48%) were provided with a Vodafone SIM card to access mobile data; and a further 15 (18%) were given SIM-enabled Jangala Get Boxes, which acted as a stand-alone internet connectivity device. A total of 29 (35%) service users requiring extra digital skills support were referred to AbilityNet by the DIO. While the scheme could provide additional 1:1 support with the DIO, some service users were also signposted to a Recovery College [39] course focused on improving digital skills, cofacilitated by the DIO or other digital support delivered by services across the C&I NHS FT or in the wider community.

At baseline, service users reported various factors that limited their ability to access the internet (Table 1), with the lack of devices and lack of knowledge and skills in using devices representing the most common barriers to accessing the internet. Those who did access the internet did so for a variety of purposes (Table 1), with more than half of respondents (n=43, 52%) indicating entertainment as a primary driver. Overall, 70% (n=57) of service users reported that access to digital technology and the internet had become more important as a result of the COVID-19 pandemic (Table 1).

Table 1. Participant characteristics and baseline digital skills, confidence, and use before participation in the Digital Inclusion Scheme (n=83).

Characteristic	Values
Gender, n (%)	
Female	43 (52)
Male	40 (48)
Age[a] (y)	
Overall, mean (SD)	52.42 (13.47)
Range	21-88
Female, mean (SD)	53.83 (14.36)
Male, mean (SD)	50.86 (12.41)
Ethnicity, n (%)	
Asian or Asian British	2 (2)
Black or Black British	12 (14)
Kurdish	1 (1)
White	50 (60)
Mixed	4 (5)
Not known	14 (17)
Level of digital skills, n (%)	
Good	13 (16)
Average	26 (31)
Poor	44 (53)
Confidence in use of digital technology, n (%)	
Extremely confident	7 (8)
Somewhat confident	20 (24)
Neutral	13 (16)
Somewhat not confident	19 (23)
Extremely not confident	24 (29)
Frequency of internet use, n (%)	
Every day	36 (43)
A few times a week	13 (16)
A few times a month	6 (7)
Never	28 (34)
Self-reported limits on internet use, n (%)	
I did not have a device	40 (48)
I struggle to know how to use my device	39 (47)
Other	34 (41)
I did not have an internet connection	28 (34)
My device was not good quality	14 (17)
I had an internet connection but it was too slow	6 (7)
There was no reason for me to use my device	4 (5)
I do not have time	4 (5)
None of my friends and family are online	0 (0)
Self-reported use of digital technology, n (%)	
For entertainment (ie, to watch videos)	43 (52)

Characteristic	Values
Other	36 (43)
To speak to friends and family	24 (29)
To stay up to date (ie, read the news or Twitter)	22 (27)
To access other physical and/or mental health services	23 (28)
To do online shopping	20 (24)
Nothing, I didn't use it	20 (24)
To speak to my GP	20 (24)
To learn	18 (22)
To speak to my clinician, care coordinator, or service	11 (13)
To access other online mental wellbeing support	7 (8)
To earn money	3 (4)
To access social services	3 (4)

[a]Age unknown: n=5 (male: n=3; female: n=2).

Part 2: Semistructured Interview Results

Overview

A total of 12 semistructured interviews were conducted between July 6, 2022, and July 14, 2022, either over the phone (n=11) or via Microsoft Teams (n=1). This obtained the qualitative data focusing on the service users' experience of the DIS as well as basic demographic information (Table 2).

The results from the semistructured interviews are presented in subsequent sections. There were 10 major themes in total. These 10 themes could be clustered into two superordinate themes: (1) factors impacting participant's engagement and (2) the impact of accessing the scheme. For the former (1), the 6 major themes that were shared among participants included digital exclusion, relationship to trust, the importance of personalized digital support, partnership working, device usability and accessibility, and personal circumstances. In the latter superordinate theme (2), we identified 4 major themes: improved access to services, impact on well-being, financial implications, and a greater sense of empowerment.

Table 2. Demographics of participants (n=12).

Characteristic	Values
Age (y), n (%)	
25-34	2 (17)
35-44	2 (17)
45-54	1 (8)
55-64	4 (33)
65-74	1 (8)
75-84	2 (17)
Gender, n (%)	
Male	5 (42)
Female	6 (50)
Gender fluid or gender queer	1 (8)
Ethnicity, n (%)	
Asian or Asian British (Bangladeshi)	1 (8)
Black or Black British (African)	1 (8)
Black or Black British (Caribbean)	1 (8)
White British	5 (42)
White Irish	2 (17)
Prefer not to say	1 (8)
Other (White British Traveller)	1 (8)
Employment status, n (%)	
Unemployed	4 (33)
Unable to work	4 (33)
Student	1 (8)
Retired	2 (17)
Unknown	1 (8)
Highest level of education, n (%)	
Secondary school	4 (33)
Sixth form college or further education (A-levels, BTEC[a], etc)	4 (33)
Higher education or university (diploma, bachelor’s degree, etc)	2 (17)
Postgraduate education	2 (17)
Disability or long-term physical or mental health condition, n (%)	
Yes	11 (92)
Prefer not to say	1 (8)
Relationship status, n (%)	
Married or civil partnership	2 (17)
Living with someone, but not married or in civil partnership	1 (8)
Single	8 (67)
Widowed	1 (8)
Household income over last year (£[b]), n (%)	
<10,000	7 (58)
10,001-20,000	3 (25)

Characteristic	Values
Prefer not to say	2 (17)
Support accessed, n (%)	
Tablet loan (C&I NHS FT[c])	12 (100)
Internet connectivity router (Jangala)	4 (33)
Mobile data (Vodafone)	7 (58)
Digital skills support referral (AbilityNet)	2 (17)
Other digital support or skills training (including DIO[d])	6 (50)
Keyboard	3 (25)

[a]BTEC: Business and Technology Education Council.

[b]Average exchange rate in July 2022: £1 GBP=US $1.2.

[c]C&I NHS FT: Camden and Islington National Health Service Foundation Trust.

[c]DIO: digital inclusion officer.

Engagement With Scheme

Digital Exclusion

Of the 12 participants, 3 (25%) participants reflected on how COVID-19 had made digital technology more important, with 7 (58%) participants reporting that they felt "forced" to learn how to use technology. They noted that before accessing the scheme, the increasing reliance on digital technology had resulted in them feeling isolated and excluded:

> *It's really important, especially since COVID.... Everything has moved online.* [P1]

> *I got people bothering me...she said you gotta learn how to do emails and stuff like that.* [P2]

> *It's just something we, we never could do before. [We felt] isolated, left behind.* [P10]

> *I remember using the clunky old computers with the floppy disks and all of that. Very dim memories. I didn't use [new technology] particularly well...I kind...of felt excluded in the sense that [these were] things I couldn't possibly buy over the years. I didn't have the capacity to keep up to speed with it because I let it slip for so many years.* [P11]

A total of 5 (42%) participants reported that the sudden digital transformation in the wake of COVID-19 left them unable to access crucial mental or physical health support, information (eg, pertaining to health care or local services), and government and banking services, as well as being unable to pay bills:

> *I have to transfer onto the Universal Credit scheme. That was another reason for really having to get to grips with the IT because you can only access that Universal Credit thing via IT, you can't do it postally anymore. So that was kind of a worry in the back of my mind, I knew it was coming up at some point, and some point this year I'll have to fill out one of these multi page online forms which I'm not looking forward to.* [P11]

> *Anything to do with art I am kind of interested in, and yoga. And I said, well I can't...because my camera doesn't work...an external camera or...external mic, none of that worked.* [P6]

Moreover, participants reported that their digital exclusion before accessing the scheme was driven by a lack of suitable devices or internet access, as well as defective or unsuitable equipment. One-third (4/12, 33%) of the participants reported that their device did not work or was not fit for purpose when needing to engage in education, employment, or therapy. It was noted that, as a result, they were unable to engage with web-based mental health appointments offered, with some participants opting for telephone consultations instead:

> *Before I saw my therapist, I was being assessed, but because of COVID, I couldn't go out. My computer didn't work. So...in the end it was all done on the phone with another therapist.* [P6]

> *A major impact I was finding it difficult to study. I'm a student, so it makes a big impact. I had a mobile phone, but I wasn't able to use anything like for typing and submitting assessments and assignments and things like that.* [P1]

In addition, the level of self-reported competency was low before accessing the scheme. One-third (4/12, 33%) of participants reported never having used digital technology, whereas 7 (58%) struggled despite having some experience with technology though limited to laptops or computers. Only 1 (8%) participant reported being proficient, but that they were unable to afford their own device. Digital confidence was also low, with 10 (83%) participants reflecting on the challenges they faced using devices, even those who professed to have had some experience:

> *I would say familiar. But I wouldn't say expert, you know what I mean? I still get lost sometimes.* [P6]

> *I never used anything like that [loan tablet] before in my life. I used to have a landline. So that's the most...we used as communication.* [P12]

Relationship to the Trust

The relationship with trust staff was key in facilitating or encouraging the service users' engagement with the scheme.

All 12 participants were unaware of the existence of the scheme before being referred by staff members; however, there were differences in the motivation to engage with the scheme. One participant found that accessing the scheme helped to foster their own agency and independence:

> *I knew [about the DIS], from my key worker, 'cause I didn't know about you guys. [I engaged because] I couldn't keep on going to my key worker with every problem I had.* [P4]

A total of 5 (42%) participants acknowledged they felt more inclined to engage with the scheme due to it being recommended by their health professional:

> *It was through my care worker.... I remember her phoning me up and saying I've got you a tablet...so I said, well all that sounds great. I'll, I'll proceed with, with this thing, you know, and it's turned out pretty decent for me to have one, you know.* [P9]

> *I think without being pushed into it [by the referrer], we might never have actually...ever thought we would possibly even attempt to use [a device], you know.* [P7]

However, for one participant, being referred by their health professional led to them feeling obliged to accept the help, despite the benefits:

> *I mean, all these things have a kind of psychological element, so when you're...offered something by your...long-term nurse or therapist, you're more psychological inclined to please them and say yes, I'll try that, it's a good idea.* [P11]

Nonetheless, 2 (17%) participants reported feeling more positive toward the scheme, as it was a service provided by the NHS, rather than a "corporation" (P9), with another participant reporting that they "would feel more vulnerable if Barclays had given it [the device] to me." Similarly, another participant felt that accessing NHS support and resources through the technology by the scheme was safer in general:

> *I also like the fact that it [is an NHS scheme], 'cause I doubt very much that Joe Public can get access to what I'm getting access to through [the NHS]. So, it's helping [the NHS] to help me, you know?* [P6]

Personalized Digital Support

All participants were positive about the role of the DIO, explaining that they felt "safe" (P2) and that the process was "easy" and the DIO "supportive" (P3) and, overall, worked "extremely well" (P7). Participants appreciated that the DIO was always "available to be contacted" (P11) if there were any problems:

> *I liked [the DIO]. [They were] very friendly and approachable, and made it very easy because, as I said I've got agoraphobia and mental health and stuff, so when I'm accessing services or things, I find it quite difficult, although I'm quite chatty on the, outwardly but, I do like self-judge and stuff quite a lot so, yeah, [they are] just, like, [they are] friendly, [they] made me feel at ease, so it wasn't a traumatic experience...* [P8]

Similarly, P8 felt that the support offered was relevant and tailored to their needs:

> *I think it was quite thorough really, because [they] offered an extension of help and stuff if I needed it...and there were extra apps and stuff on the tablet that were, I felt like they were tailored towards what I had said, like the PTSD [posttraumatic stress disorder] thing, so and I think you mentioned that, that possibly your IT department would put apps on there so...so actually, I think you've done that quite well actually.* [P8]

However, 7 (58%) participants reported needing further digital support to use the device, with 1 participant finding the amount of web-based content "daunting" (P11). In total, 5 (42%) participants identified needing additional support from the DIO, with 6 (50%) receiving digital skills support from external partners, such as AbilityNet or community groups. Despite this, 10 (83%) reported still feeling unsure or uncertain; for example, P7 worried that they "might press the wrong buttons" and resorted to trial and error instead of asking for help. One participant reported feeling uncomfortable asking for further support:

> *I guess I didn't want to be too pushy about it. I didn't know what the boundaries were.* [P11]

P11 also described feeling that the scheme assumed a higher level of digital competency, which they perceived as impacting on the level of support provided:

> *That's a criticism that should be levelled at lots of IT schemes, sometimes they assume too much at the beginning. It was even true of the IT skills course at [the C&I NHS FT service], just a lot was assumed at the earliest stage.* [P11]

The service received by the participants was also at times delayed by the staffing arrangements of the scheme, with one participant's ability to join a web-based course affected by the sole DIO being on leave:

> *I'm still waiting for the next art class to start because I did one session, and they were all very happy and then I got an email saying they're not happy. The other participants. And I did start quite late, but I mean that wasn't quite my fault as I couldn't get hold of the...tablet for ages.* [P6]

Partnership Working

A further essential component of the scheme organization was the partnerships established with external organizations, including AbilityNet for personalized digital support, Jangala for connectivity devices, and Vodafone for SIM cards with free data.

While 1 (8%) participant described the partnership support as "invaluable" and the connectivity device provided by Jangala as "simple" to use, participants were nonetheless more inclined to seek additional support from the DIO (5/12, 42%), as opposed to the designated partner organization (2/12, 17%). For example, while P7 reported that the telephone support provided by the AbilityNet volunteer was "helpful," they also noted limitations:

> *I don't know what I could say about [external support] really, because...I just think that it was very short and quite brief...it's really not enough time to get things to sink in..., we just need that bit more help like that...just if it could have continued for even for four or five weeks, once a week or once a fortnight, or just something like, that would have been would have been very helpful.* [P7]

> *It was just...so basic, I couldn't say if it was brilliant or not brilliant. It worked. What they said was correct...[But] well you know me with this phone, it's often turned off...So, it just never matched up.* [P11]

Device Usability and Accessibility

The tablets were described as "easy to set up" (P3) and "helpful" (P1). The physical flexibility of a tablet was also cited as a positive (eg, as it allowed the tablet to be used outdoors or facilitate engagement with therapeutic services from the safety of their home environment). For example, P3 explained how accessing a portable device, such as the loan device, enabled them to participate privately in group therapy:

> *I didn't feel as comfortable, I would say, to participate in group therapy without it, without having that privacy element [physically being able to move the tablet to a private space] in it. And so, when I did join via video link it did...I felt more inclusive, and you know able to participate properly in the group.* [P3]

The additional items available to loan (eg, a keyboard and case) were also reported to be "a massive help" (P8). However, 2 (17%) participants raised concerns about the restrictions placed on the loan device by C&I NHS FT for security purposes. As one participant explained:

> *I just didn't like the fact that I couldn't access some of the services... due to it being sort of owned by the trust. With Google and Google services and things like that.* [P1]

Other negatives of using a tablet included it being "quite heavy in weight" (P9) and not as practical, or familiar, as a computer:

> *I think it can be difficult to edit PDF's, especially on tablets, it's not quite as straightforward as a computer.* [P4]

Concerns around the safety or implications of use were also cited as barriers to fully using the loan device. In total, 2 participants did not engage with the loan device once it had been loaned to them, with one instead opting to use an older device provided by their daughter in case they would damage the equipment in any way:

> *We were quite nervous about getting [a tablet] because we were worried that we might press the wrong buttons, we might do something wrong.* [P7]

> *I was afraid to, you know, in case I'd mess it all up. So, my daughter said just use mine and if you if you mess it up, you know what I mean, 'cause it's old.* [P5]

One participant also reported misunderstanding the terms of the tablet loan, limiting their use, believing that the tablet was "not something you use for personal use" (P3).

A total of 5 (42%) participants expressed being worried that in accessing digital technology, they may become more vulnerable to fraud, cyberattacks, or hackers accessing their personal or private information. Consequently, some participants limited their use of the loan device due to concerns around web-based security and privacy, for example:

> *I'd be very nervous of putting in a debit card number and buying something online, I'm thinking somebody is going to scam me here, something's going to go wrong. I'm still nervous of that.* [P11]

> *I just don't want anything personal or private on that because it's not actually mine.* [P6]

Personal Circumstances

In general, disability and poor mental health affected the participants' ability to engage with the scheme and the digital world more generally. A total of 10 (83%) participants reported having a disability, and nearly half (5/12, 42%) reported mental health difficulties that affected their daily lives and impacted their ability to engage with the scheme:

> *It's me, personally, not the service available.... I just lost confidence in everything.* [P5]

In total, 2 (17%) participants reported difficulties with reading and writing, which hindered their ability to access the scheme or loan device, despite acknowledging its potential benefits:

> *It would help an awful lot of people; you know what I mean. But not in the situation I am in.... It's my spelling ability [that hindered me in engaging with the tablet]...You've got to kinda spell things.* [P12]

Beliefs that older age negatively affect technology use were noted by a third of participants (4/12, 33%), both as a causal and maintaining factor for digital exclusion. In total, 2 participants stated a preference for not engaging with technology due to their age:

> *I grew up without, before that technology came out. So, I'm in my 50s now and I'd prefer to stay in my old school world.* [P9]

> *At my age, what do I want a computer for?* [P12]

A total of 3 (25%) participants expressed feeling motivated to access the digital world despite this. However, the need for additional support to overcome the perceived age-related difficulties in learning to use technology was also highlighted:

> *You know as you get older, obviously it's a bit more difficult to pick things up, and you know you need a little bit more than a couple of, sort of, half an hour or even an hour session. I know you can learn by just using it, but you really do need a bit more help in actually using it, you know.* [P7]

> *I do find tablets a little bit difficult because I think it might be my age.... I find I'm a little bit disconnected with like iPad and things like that. I find it easier to use computers.* [P8]

Impact of Accessing Scheme

Access to Services

The scheme provided "access to the digital world" (P8), with 4 (33%) participants reporting that they had accessed C&I NHS FT services as a direct result of the scheme. For example, P6, had previously struggled to manage their posttraumatic stress disorder symptoms during lockdown and described the benefits of being able to attend web-based yoga classes using their loaned tablet, rather than attending in person:

> *Communicating with people is becoming increasingly difficult for me. But I feel very comfortable at home, and this has nothing to do with COVID. This is just me shutting down, basically. So being able to stay within my own home environment, safe environment, nobody knows where I live. But I have access to yoga...because of who I am...the very idea of going to a place, in a room with a lot of strangers and try and do yoga or tai chi in a mask is just "no, I'm sorry."* [P6]

Improved access to education and employment was noted, with 5 (42%) participants attending web-based courses or using their device to facilitate their studies following engagement with the scheme:

> *It has given me a sense of, you know, of other possibilities and things like that. It's kind of, you know, got me out of the house and all that [to attend a course].* [P11]

A total of 10 (83%) participants also acknowledged the necessity of digital technology and the access it provided. For example, P1 reported the tablet to be an "essential item," and P7 found they frequently relied upon it:

> *The first thing I, I would look for if I wanted help, or wanted to look for an organisation, I would look on [the tablet] straight away, you know, and ask them what I wanted and up comes an awful lot of options, you know.* [P7]

Impact on Well-Being

A total of 10 (83%) participants reported that the scheme had a positive effect on their well-being. Accessing the scheme had improved participants' mood, built confidence and self-esteem, relieved stress, aided sleep, provided "comfort" (P9), and proved to be a distraction from anxiety. A total of 9 (75%) participants also reported feeling included, more connected, and less isolated; no longer the "odd ones out" (P7) and being able to "re-access parts of my old life via things like social media" (P8):

> *In my general well-being, it's actually good for me. I'll give you 100%, I'll tell you that much. Like literally I go home now when I finish my course at 4, I just go down there and then put music on and start cleaning the house.* [P2]

It also provided opportunities for entertainment (eg, browsing the internet, playing chess, or listening to music); staying up-to-date (eg, watching the news); and social interactions (eg, video calling friends or using Facebook to keep in touch with others):

> *Well, maybe I listen to music. Maybe I listen to something educational, you know depends on what mood I'm in, you know.* [P4]

One participant felt that accessing the tablet had "open[ed] up the world to me" (P6), as they did not feel comfortable interacting with people outside of their home:

> *Communicating with people is becoming increasingly difficult for me. But I feel very comfortable at home, and this has nothing to do with COVID. This is just me shutting down, basically. So being able to stay within my own home environment, safe environment, nobody knows where I live.* [P6]

However, 3 (25%) participants also expressed potential negative impacts on well-being. For example, P11 expressed concern around using the tablet "too much" and described how having access to a tablet and the digital world also led to feelings of disconnection and isolation, noting that it can be "antisocial...sitting alone with a screen" where "you become wrapped up in [an]...internal and isolated world":

> *I get a call from someone to come on up a coffee or something and say no I'm do[ing] this online...because I wasn't, you know, really au fait with all the IT skills, I thought no no no, I don't wanna lose my thread and lose what I'm on here...So, there is an addictive aspect. It's really kind of... ensnare you in a kind of isolating mental practice.* [P11]

One participant was conscious of limiting their reliance on technology when accessing mental health support to challenge avoidance behaviors linked to their mental health difficulty:

> *So, for me, I prefer physical appointments even though I hate the agoraphobia..., but it's important for me to still do that otherwise I won't go out.* [P8]

Financial Implications

Before engaging with the scheme, 7 (58%) participants cited the unaffordability of technology as a reason for being digitally excluded. While participants did not have to pay to access the scheme, P7 noted that they had decided to invest in Wi-Fi when they knew they were going to be loaning a tablet. P6 noted that if there had been any extra cost involved, it would have led them to reconsider their involvement. A total of 2 (17%) participants reported that the scheme had potentially saved them money:

> *Yes, I'm, I have listened to certain pieces of music, in the past I would have you know, saved up for a few weeks and then bought a couple of CDs, for example, yeah.* [P11]

> *I guess that aspect of it has saved money. So, and like sending letters and replying and tracking stuff like- using my phone or my travel money.* [P8]

However, accessing the scheme provided other financial advantages; for example, it enabled P8 to manage their money on the web:

> *I use the internet banking, so it's been paramount and stuff so. ...As well, my bank is Co-Op, so they're very few and far between. That's not even one close to me, so when it comes down to banking, internet is vital.* [P8]

Meanwhile, P1 and P2 reported that the scheme was a stopgap, enabling them to access a device, while they saved to purchase their own:

> *I couldn't afford to purchase a new, a new device.... It's given me that time to sort of bridge that gap and so I can afford to buy myself something when the loan finishes.* [P1]

P2 noted that after their involvement in the scheme was over, they hoped to purchase a device themselves or obtain one via a charity or other means:

> *I put away like £5 away each week. Just to think, maybe I can get a tablet as well, you know...I would like to have one.* [P2]

Empowerment

For most participants, the scheme provided a sense of empowerment. More than half of the participants (7/12, 58%) reported that using a tablet and having access to the digital world was either a new experience or one that broadened their world. One participant described the amount of content on the web described as a "revelation" (P11). This led participants to feel like the scheme had enabled them to learn new skills, whether it was how to access music on the web, how to use a tablet (ie, rather than a computer or laptop), or how to access the internet more generally. As a result, a quarter of participants (3/12, 25%) described that access to a tablet, connectivity, and skills now allowed them to have more control over their health, for example, by looking up information about medications, tests, and health issues or by organizing physical health appointments at the hospital, general practitioner, or dentist:

> *[I was] writing my appointments down on pieces of paper and things like that, so yeah. It's a lot to manage. Like I've had at least 4 appointments each week for the past three months. And that's like other-dentist, doctors, psychiatry and hospital tests...I've been able to manage myself a little bit better [using the tablet].* [P8]

Participants also felt empowered by being able to independently manage their finances (eg, web-based banking [P8]), access information (eg, health care [P7] and government services and benefits [P11]), or practical support (eg, web-based shopping [P1] or transport apps [P11]):

> *[What] we have done though is look up things like pensions, changes to universal credits, whether we qualify for all these new things that are coming up. So, information might change on the government websites.* [P10]

Participants were unanimous in reporting that the scheme was of high value and that other trusts should look to implement similar schemes:

> *Well, I think, I think it's, well it's the best thing...I don't think that we've ever really had anything like this happen to us before and it has made our lives very different and a lot more interesting for both of us. And I do think that if other trusts did it, I think it's a wonderful idea.* [P7]

P6 also reflected on their concern that the scheme was only available as a result of COVID-19, with digital exclusion predating COVID-19:

> *I don't know if I'd have been given [a tablet], if COVID had never happened. So, I guess that's a factor. And it might be a factor in the future for you lot. I mean the Government love to make excuses as to why they can't find money.... Yeah, it's hypothetical for me to kind of, put in a negative, I mean it's a hypothetical negative, right? I wonder what my life would have been like without it. And was COVID instrumental in me having it? That's the question I would ask.* [P6]

One participant highlighted how it gave service users the opportunity to engage with novel technology and build their confidence:

> *To me [the loaned device] was just a better option than us going to a shop and getting one and not having the first idea even how to switch it on [...] Without just having a little bit of help which we got, we wouldn't really have much idea...perhaps you're a bit shy about getting into new technology...and I think without being pushed into it, we might never have actually, you know, ever thought we, we would possibly even attempt to use one.* [P7]

Discussion

Principal Findings

This study highlighted the experience of digitally excluded service users engaging in an innovative tablet loan scheme. The scheme aimed to mitigate the detrimental impact of digital exclusion on service users who lacked the necessary equipment, skills, or confidence to access mental health services or effectively engage with digital ways of working at a societal level.

Findings align with prior research on the 3-level digital divide, emphasizing the negative consequences of COVID-19–related restrictions on health service access and well-being of individuals with preexisting mental health difficulties [18]. Semistructured interviews revealed 10 major themes describing the facilitators and barriers to engagement with the scheme and subsequent impact on well-being. Factors limiting engagement included disability; poor mental health; financial constraints; and concerns around web-based safety and conditions of the loan device use, linked to lower skills and confidence. This aligns with prior research on the limited uptake of digital inclusion initiatives [12,40]. While some participants additionally reported a belief that older age influenced support needs, it was not within the scope and nature of this study to specifically assess whether age determined technological ability. However, the expressed views highlight the importance of tackling stereotypes and providing enhanced digital support for older adults, where required [41,42].

Factors facilitating engagement with the scheme included the relationship to referrers and the wider organization, as well as the personalized approach to providing digital support. Staff played a crucial role facilitating access to the scheme, but motivation varied among participants. Most expressed a desire to learn and understand technology, while others wanted to use the device to access basic functions. Some felt obligated to accept the help due to their relationship with their health care professionals, others explicitly chose not to engage with digital technology or the scheme, challenging the assumption of digital inclusion initiatives providing a desired solution. Moreover, most participants described lacking awareness of the scheme before being referred, which may be attributed to communication being limited to trust staff initially to manage limited capacity and resources before advertising the scheme more widely through posters and trust-wide communications. Regardless of their own level of engagement or participation in the scheme, participants unanimously emphasized the importance of the DIO. Participants also praised the overall value of the scheme and suggested that other NHS trusts would be well-serviced in implementing similar schemes. All participants described the wide-ranging impact the DIS had on their lives; facilitating the provision of mental health and other services; enabling participants to take control of their health care and finances; and providing access to social, educational, and entertainment opportunities. Participants also reported an improvement in well-being and gaining a sense of empowerment, without incurring additional costs. However, limited resources and the time-limited nature of the loan constrained the support offered by the DIS, a concern also seen in other initiatives limited in time or reach [4,12,42].

Lessons Learned

Understanding service user motivation and needs is crucial for tailored support, ensuring it is relevant and not presumptive. Future initiatives should prioritize individual consultation before implementation of schemes, even if time-consuming, to ensure full engagement and use of the support offered [43].

A dedicated DIO is a vital resource, both in administering the scheme and safeguarding service user needs. It was necessary to ensure that there was a thorough risk assessment as well as to facilitate a person-centered approach. Given the complexity of mental health needs, the DIO in turn benefited from ad hoc clinical supervision to manage risk and ensure that service user needs were met sensitively and appropriately.

Several participants noted reluctance or concern around using the loan device fully as it was not their own. Gifting devices may allow for greater engagement or use; however, finite resources were a major limitation in the scheme. Evaluations such as this highlighting the benefits of digital inclusion projects across health care services are vital to demonstrate the value of investment [44].

Considering the specific use and accessibility of loan devices is critical. It may increase engagement if the service users' specific needs were considered to determine which equipment to provide (eg, laptop rather than tablet) and the duration and type of support required to prevent the subsequent reexclusion.

Schemes should be tailored to those who need and want it most. service users who are digitally excluded will not have access to web-based promotion or lack in understanding the value of digital technology, thus not feeling compelled to ask for support. It may be helpful to be creative in increasing awareness through physical means (eg, posters or leaflets) and workshops and by advertising support offers across community services [42]. It is also helpful to connect with staff, for example, by attending team meetings, to highlight the value of the scheme and thus facilitate the access for service users under their care.

Limitations and Future Considerations

Challenges in evaluating patient-centered programs include biased sampling or unreliable self-report data [18,43]. In addition, the method of contacting participants exposed digital divides, as those unable to opt in on the web had fewer opportunities to share their views, while the influence of health factors may have impacted others' ability to participate [11]. Moreover, by using a convenience sample drawn only from the service users who had previously engaged in the scheme, there was a risk of social desirability bias, with participants providing more positive evaluations, while insights from service users who were unable to or did not want to engage with the scheme were not captured. To address these limitations, the interviewer ensured that all the participants were aware that both positive and negative evaluations were welcome and would not negatively impact their current or future access to the scheme. In addition, the qualitative nature of the study did not allow for an investigation into the causal relationship between specific factors and digital exclusion. Nevertheless, understanding the demographics, motivations, and challenges experienced by those who participated in the scheme provided valuable insights into the populations that require support and those who may need more targeted interventions [42].

Finally, one might query whether the need for digital inclusion initiatives is still required, given the lessened immediate impact of COVID-19–related disruption on services. However, while this scheme was introduced in the height of COVID-19, the NHS long-term plan had long recommended a continued expansion of digital transformation initiatives across the United Kingdom [9]. Mounting evidence compiled before, during, and after COVID-19 suggests that web-based therapeutic interventions are at least equivalent to in-person alternatives [7] and are increasingly being trialed and evaluated in specialist health care services, such as prisons [45]. Outside of the NHS, many essential public services have also moved on the web, with the UK government committing to increased spending and investment to further improve efficiency and ease of access [46].

But despite the wider financial benefits of implementing web-based services [47], many still express a desire for in-person availability of support [47]. Furthermore, nearly a quarter of working-age adults report living with a disability that impacts their day-to-day functioning [48], alongside greater prevalence of mental health conditions and lower employment rates [49]. Given the disproportionate impact on these groups [50], it will be necessary to consider the impact of digital transformation initiatives on access to health care, public services, and wider society for those who are not digitally enabled.

The debate around who should bear the responsibility of digital inclusion initiatives continues, with community organizations and libraries often serving as popular suggestions [44,51]. Others have argued for interventions on a macrolevel through policy change [24,25]. However, in this paper, we present the notion that when presented with disrupted access to health care, providing support at source, that is, by the health care provider itself, can have a profound impact on uptake and response.

Conclusions

This evaluation aimed to demonstrate the potential of a digital inclusion initiative to enable NHS mental health service users to access the digital world. By examining engaged service users and their experiences within the scheme, valuable insights have been gained, addressing a previously existing knowledge gap. The study offers recommendations for future digital inclusion efforts while also highlighting the importance of providing equitable access to digital technology and support within health care contexts, especially given the rise in remotely delivered health care services.

Acknowledgments

This paper represents a service evaluation supported by Camden and Islington National Health Service (NHS) Foundation Trust and Anglia Ruskin University. The views expressed are those of the authors and not necessarily those of the NHS, Camden and Islington NHS Foundation Trust, or the Department of Health and Social Care.

Authors' Contributions

This project was designed by JAG and AO. AO recruited participants with the support of EC. AO conducted and transcribed the interviews, as well as extracted key themes, alongside JAG. The manuscript was prepared by AO, EC, and JAG.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Camden and Islington National Health Service Foundation Trust Digital Inclusion Project (baseline survey).
[DOCX File , 21 KB - mental_v11i1e51315_app1.docx]

Multimedia Appendix 2
Semistructured interview schedule.
[DOCX File , 27 KB - mental_v11i1e51315_app2.docx]

Multimedia Appendix 3
Participant information sheet.
[DOCX File , 21 KB - mental_v11i1e51315_app3.docx]

Multimedia Appendix 4
Camden and Islington National Health Service Foundation Trust Digital Inclusion Scheme evaluation consent form.
[DOCX File , 20 KB - mental_v11i1e51315_app4.docx]

References

1. Chen JA, Chung WJ, Young SK, Tuttle MC, Collins MB, Darghouth SL, et al. COVID-19 and telepsychiatry: early outpatient experiences and implications for the future. Gen Hosp Psychiatry 2020;66:89-95 [FREE Full text] [doi: 10.1016/j.genhosppsych.2020.07.002] [Medline: 32750604]
2. Daly M, Sutin AR, Robinson E. Longitudinal changes in mental health and the COVID-19 pandemic: evidence from the UK Household Longitudinal Study. Psychol Med 2020 Nov 13;52(13):2549-2558. [doi: 10.1017/s0033291720004432]
3. Fisk M, Livingstone A, Pit SW. Telehealth in the context of COVID-19: changing perspectives in Australia, the United Kingdom, and the United States. J Med Internet Res 2020 Jun 09;22(6):e19264 [FREE Full text] [doi: 10.2196/19264] [Medline: 32463377]
4. Kanellopoulos D, Bueno Castellano C, McGlynn L, Gerber S, Francois D, Rosenblum L, et al. Implementation of telehealth services for inpatient psychiatric COVID-19 positive patients: a blueprint for adapting the milieu. Gen Hosp Psychiatry 2021 Jan;68:113-114 [FREE Full text] [doi: 10.1016/j.genhosppsych.2020.08.011] [Medline: 32972728]

5. Burton A, McKinlay A, Aughterson H, Fancourt D. Impact of the COVID-19 pandemic on the mental health and well-being of adults with mental health conditions in the UK: a qualitative interview study. J Ment Health 2023 Dec 29;32(6):1040-1047. [doi: 10.1080/09638237.2021.1952953] [Medline: 34323635]
6. Adams EA, Parker J, Jablonski T, Kennedy J, Tasker F, Hunter D, et al. A qualitative study exploring access to mental health and substance use support among individuals experiencing homelessness during COVID-19. Int J Environ Res Public Health 2022 Mar 15;19(6):3459 [FREE Full text] [doi: 10.3390/ijerph19063459] [Medline: 35329147]
7. Stepanova E, Thompson A, Yu G, Fu Y. Changes in mental health services in response to the COVID-19 pandemic in high-income countries: a rapid review. BMC Psychiatry 2024 Feb 06;24(1):103 [FREE Full text] [doi: 10.1186/s12888-024-05497-6] [Medline: 38321403]
8. Vera San Juan N, Shah P, Schlief M, Appleton R, Nyikavaranda P, Birken M, et al. Service user experiences and views regarding telemental health during the COVID-19 pandemic: a co-produced framework analysis. PLoS One 2021 Sep 16;16(9):e0257270 [FREE Full text] [doi: 10.1371/journal.pone.0257270] [Medline: 34529705]
9. The NHS long term plan. National Health Service. 2019 Aug. URL: https://www.longtermplan.nhs.uk/wp-content/uploads/2019/08/nhs-long-term-plan-version-1.2.pdf [accessed 2024-04-03]
10. Lakeman R, Crighton J. The impact of social distancing on people with borderline personality disorder: the views of dialectical behavioural therapists. Issues Ment Health Nurs 2021 May 15;42(5):410-416. [doi: 10.1080/01612840.2020.1817208] [Medline: 32931341]
11. Liberati E, Richards N, Parker J, Willars J, Scott D, Boydell N, et al. Remote care for mental health: qualitative study with service users, carers and staff during the COVID-19 pandemic. BMJ Open 2021 Apr 22;11(4):e049210 [FREE Full text] [doi: 10.1136/bmjopen-2021-049210] [Medline: 33888531]
12. Digital inclusion in mental health: a guide to help increase choice and improve access to digital mental health services. Mental Health Network. 2020 Dec. URL: https://tinyurl.com/y98d7c3t [accessed 2022-06-07]
13. Dorney-Smith S, Gill N. Improving digital health access for excluded groups. Br J Healthc Manag 2017 Aug 02;23(8):354-356. [doi: 10.12968/bjhc.2017.23.8.354]
14. Bastick Z, Mallet-Garcia M. Double lockdown: the effects of digital exclusion on undocumented immigrants during the COVID-19 pandemic. New Media Soc 2022 Feb 11;24(2):365-383. [doi: 10.1177/14614448211063185]
15. Tobitt S, Percival R. Switched on or switched off? A survey of mobile, computer and internet use in a community mental health rehabilitation sample. J Ment Health 2019 Feb;28(1):4-10. [doi: 10.1080/09638237.2017.1340623] [Medline: 28675329]
16. Hoffman DL, Novak TP, Schlosser A. The evolution of the digital divide: how gaps in internet access may impact electronic commerce. J Comput Mediat Commun 2000;5(3). [doi: 10.1111/j.1083-6101.2000.tb00341.x]
17. van Dijk JA. The Deepening Divide: Inequality in the Information Society. Thousand Oaks, CA: SAGE Publications; 2005.
18. Sharp M. Revisiting digital inclusion: a survey of theory, measurement and recent research. Blavatnik School of Government. 2022 Mar 31. URL: https://tinyurl.com/54yazb5p [accessed 2022-06-14]
19. Schlief M, Saunders KR, Appleton R, Barnett P, Vera San Juan N, Foye U, et al. Synthesis of the evidence on what works for whom in telemental health: rapid realist review. Interact J Med Res 2022 Sep 29;11(2):e38239 [FREE Full text] [doi: 10.2196/38239] [Medline: 35767691]
20. Seifert A. The digital exclusion of older adults during the COVID-19 pandemic. J Gerontol Soc Work 2020 May 13;63(6-7):674-676. [doi: 10.1080/01634372.2020.1764687] [Medline: 32401181]
21. Watson A, Mellotte H, Hardy A, Peters E, Keen N, Kane F. The digital divide: factors impacting on uptake of remote therapy in a South London psychological therapy service for people with psychosis. J Ment Health 2022 Dec 28;31(6):825-832. [doi: 10.1080/09638237.2021.1952955] [Medline: 34319202]
22. Alkureishi MA, Choo ZY, Rahman A, Ho K, Benning-Shorb J, Lenti G, et al. Digitally disconnected: qualitative study of patient perspectives on the digital divide and potential solutions. JMIR Hum Factors 2021 Dec 15;8(4):e33364 [FREE Full text] [doi: 10.2196/33364] [Medline: 34705664]
23. Al-Saffar M, Noble M, Patel T, Tahir S, Schneider F. A digital first primary care model in the National Health Service (NHS): a perspective from a digital health provider. Int J Digit Health 2021;1(1):11 [FREE Full text] [doi: 10.29337/ijdh.35]
24. Weil J, Kamber T, Glazebrook A, Giorgi M, Ziegler K. Digital inclusion of older adults during COVID-19: lessons from a case study of older adults technology services (OATS). J Gerontol Soc Work 2021 Sep 22;64(6):643-655. [doi: 10.1080/01634372.2021.1919274] [Medline: 33882782]
25. Aissaoui N. The digital divide: a literature review and some directions for future research in light of COVID-19. Glob Knowl Mem Commun 2021 Feb 04;71(8/9):686-708. [doi: 10.1108/gkmc-06-2020-0075]
26. Digital communities Wales. Cwmpas. 2022. URL: https://cwmpas.coop/what-we-do/services/digital-communities-wales/ [accessed 2022-12-07]
27. Tutin J, Atherton B. 100% Digital Leeds update. Leeds City Council. 2022 Apr 6. URL: https://tinyurl.com/3ehne764 [accessed 2022-12-07]
28. charities.connected. Vodafone. URL: https://www.vodafone.co.uk/mobile/everyone-connected/charities-connected [accessed 2022-12-07]
29. Enabling internet access for people in need. Jangala. URL: https://www.janga.la [accessed 2022-12-07]

30. Home page. AbilityNet. URL: https://abilitynet.org.uk [accessed 2022-12-07]
31. Version 3.3.070. ServiceNow. URL: https://www.servicenow.com [accessed 2022-12-07]
32. The digital inclusion evaluation toolkit: bank of outcomes, indicators and survey questions. Just Economics. 2017 Mar. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/605093/DigitalInclusion_BankOfOutcomes.pdf [accessed 2022-12-07]
33. O'Connor C, Joffe H. Intercoder reliability in qualitative research: debates and practical guidelines. Int J Qual Methods 2020 Jan 22;19:160940691989922. [doi: 10.1177/1609406919899220]
34. Herlitz L, Harden A. Can the use of technology help to reduce social isolation and loneliness? An in-depth study of digital inclusion projects for older people living with, or at risk, of social isolation and loneliness before and during the COVID-19 pandemic. Connect Hackney. 2021. URL: https://www.connecthackney.org.uk/wp-content/uploads/2021/07/Digital-Inclusion-Summary-Report-07_2021.pdf [accessed 2022-06-14]
35. Best practice in the ethics and governance of service evaluation guidelines for evaluators and commissioners of evaluation in health and social care. West of England Evaluation Strategy Group. URL: https://arc-w.nihr.ac.uk/Wordpress/wp-content/uploads/2021/02/Full-ethics-guidelines-revised-Nov-2020.pdf [accessed 2022-12-07]
36. Chen LY, Fawcett TN. Service evaluation: a grey area of research? Nurs Ethics 2019 Jun 20;26(4):1172-1185. [doi: 10.1177/0969733017742961] [Medline: 29262740]
37. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
38. Hewitt G, Sims S, Harris R. The realist approach to evaluation research: an introduction. Int J Ther Rehabil 2012 May;19(5):250-259. [doi: 10.12968/ijtr.2012.19.5.250]
39. Recovery College. North London Mental Health Partnership. URL: https://www.northlondonmentalhealth.nhs.uk/search/service/recovery-college-224 [accessed 2022-12-07]
40. Digital exclusion: a review of Ofcom's research on digital exclusion among adults in the UK. Ofcom. 2022 Mar 30. URL: https://www.ofcom.org.uk/__data/assets/pdf_file/0022/234364/digital-exclusion-review-2022.pdf [accessed 2022-11-01]
41. Helsper EJ, Eynon R. Digital natives: where is the evidence? British Educational Res J 2013 Jan 02;36(3):503-520 [FREE Full text] [doi: 10.1080/01411920902989227]
42. Mouland J. The digital age: new approaches to supporting people in later life get online. Centre for Ageing Better. 2018 May. URL: https://ageing-better.org.uk/sites/default/files/2018-06/The-digital-age.pdf [accessed 2022-09-04]
43. Díaz Andrade A, Techatassanasoontorn AA. Digital enforcement: rethinking the pursuit of a digitally ‐ enabled society. Inf Syst J 2020 Sep 16;31(1):184-197. [doi: 10.1111/isj.12306]
44. Petrou C, Jameel L, Nahabedian N, Kane F. A call for digital inclusion initiatives in mental health services: an integrative review. J Psychiatr Ment Health Nurs 2023 Oct 06;30(5):911-941. [doi: 10.1111/jpm.12931] [Medline: 37022715]
45. Elison S, Weston S, Davies G, Dugdale S, Ward J. Findings from mixed-methods feasibility and effectiveness evaluations of the "Breaking Free Online" treatment and recovery programme for substance misuse in prisons. Drugs Educ Prev Policy 2015 Oct 01;23(2):176-185 [FREE Full text] [doi: 10.3109/09687637.2015.1090397]
46. Press release: 50 top online government services to be easier and faster to use than ever before. Cabinet Office. 2022 Jun. URL: https://tinyurl.com/mu8t8e9k [accessed 2023-04-04]
47. Build back stronger: the final report of renew normal: the people's commission on life after COVID-19. DEMOS. 2021 Mar. URL: https://demos.co.uk/wp-content/uploads/2021/03/Build-Back-Stronger.pdf [accessed 2024-05-06]
48. Kirk-Wade E. UK disability statistics: prevalence and life experiences. House of Commons Library. 2023 Aug 23. URL: https://researchbriefings.files.parliament.uk/documents/CBP-9602/CBP-9602.pdf [accessed 2024-05-16]
49. Adam M, Brown A, Diong J. Official statistics: employment of disabled people 2022. Government of UK. 2023 Jan 26. URL: https://tinyurl.com/y8hdvc5k [accessed 2024-05-16]
50. 2023 consumer digital index. Lloyds Bank. 2023 Nov. URL: https://www.lloydsbank.com/assets/media/pdfs/banking_with_us/whats-happening/231122-lloyds-consumer-digital-index-2023-report.pdf [accessed 2024-05-16]
51. Real B. Bridging digital divides during COVID-19: findings from the 2020-2021 Connecticut State Library Digital Inclusion Survey. Public Libr Q 2021 Jun 23;40(4):283-309. [doi: 10.1080/01616846.2021.1938918]

Abbreviations

C&I NHS FT: Camden and Islington National Health Service Foundation Trust
DIO: digital inclusion officer
DIS: Digital Inclusion Scheme
ICT: information and communications technology
NHS: National Health Service

Edited by J Torous; submitted 17.08.23; peer-reviewed by B Greer, F Kane; comments to author 22.03.24; revised version received 24.05.24; accepted 25.05.24; published 26.07.24.

Please cite as:
Oliver A, Chandler E, Gillard JA
Impact of Digital Inclusion Initiative to Facilitate Access to Mental Health Services: Service User Interview Study
JMIR Ment Health 2024;11:e51315
URL: https://mental.jmir.org/2024/1/e51315
doi:10.2196/51315
PMID:

Original Paper

Skill Enactment Among University Students Using a Brief Video-Based Mental Health Intervention: Mixed Methods Study Within a Randomized Controlled Trial

Hayley M Jackson[1], BA; Philip J Batterham[1], PhD; Alison L Calear[1], PhD; Jeneva L Ohan[2,3], PhD; Louise M Farrer[1], PhD

[1]Centre for Mental Health Research, National Centre for Epidemiology and Population Health, The Australian National University, Acton ACT, Australia
[2]School of Psychological Science, University of Western Australia, Perth, Australia
[3]Telethon Kids Institute, Nedlands, Australia

Corresponding Author:
Hayley M Jackson, BA
Centre for Mental Health Research
National Centre for Epidemiology and Population Health
The Australian National University
63 Eggleston Road
Acton ACT, 2601
Australia
Phone: 61 416913323
Email: hayley.jackson@anu.edu.au

Abstract

Background: Mental health problems are common among university students, yet many students do not seek professional help. Digital mental health interventions can increase students' access to support and have been shown to be effective in preventing and treating mental health problems. However, little is known about the extent to which students implement therapeutic skills from these programs in everyday life (ie, skill enactment) or about the impact of skill enactment on outcomes.

Objective: This study aims to assess the effects of a low-intensity video-based intervention, Uni Virtual Clinic Lite (UVC-Lite), in improving skill enactment relative to an attention-control program (primary aim) and examine whether skill enactment influences symptoms of depression and anxiety (secondary aim). The study also qualitatively explored participants' experiences of, and motivations for, engaging with the therapeutic techniques.

Methods: We analyzed data from a randomized controlled trial testing the effectiveness of UVC-Lite for symptoms of depression and anxiety among university students with mild to moderate levels of psychological distress. Participants were recruited from universities across Australia and randomly assigned to 6 weeks of self-guided use of UVC-Lite (243/487, 49.9%) or an attention-control program (244/487, 50.1%). Quantitative data on skill enactment, depression, and anxiety were collected through baseline, postintervention, and 3- and 6-month follow-up surveys. Qualitative data were obtained from 29 intervention-group participants through open-ended questions during postintervention surveys (n=17, 59%) and semistructured interviews (n=12, 41%) after the intervention period concluded.

Results: Mixed model repeated measures ANOVA demonstrated that the intervention did not significantly improve skill enactment ($F_{3,215.36}$=0.50; P=.68). Skill enactment was also not found to influence change in symptoms of depression ($F_{3,241.10}$=1.69; P=.17) or anxiety ($F_{3,233.71}$=1.11; P=.35). However, higher levels of skill enactment were associated with lower symptom levels among both intervention and control group participants across time points (depression: $F_{1,541.87}$=134.61; P<.001; anxiety: $F_{1,535.11}$=73.08; P<.001). Inductive content analysis confirmed low levels of skill enactment among intervention group participants. Participants were motivated to use techniques and skills that were perceived to be personally relevant, easily integrated into daily life, and that were novel or had worked for them in the past.

Conclusions: The intervention did not improve skill enactment or mental health among students with mild to moderate psychological distress. Low adherence impacted our ability to draw robust conclusions regarding the intervention's impact on outcomes. Factors influencing skill enactment differed across individuals, suggesting that it may be necessary to tailor therapeutic skills and engagement strategies to the individual user. Theoretically informed research involving collaboration with end users is needed to understand the processes underlying skill enactment in digital mental health interventions.

Trial Registration: Australian New Zealand Clinical Trials Registry ACTRN12621000375853; https://tinyurl.com/7b9ar54r

(JMIR Ment Health 2024;11:e53794) doi:10.2196/53794

KEYWORDS

university students; young people; internet; computer-assisted therapy; engagement; skill enactment; depression; anxiety; randomized controlled trial; mobile phone

Introduction

Background

University students are a population in need of mental health support. Before the COVID-19 pandemic, findings from the World Health Organization World Mental Health International College Student initiative indicated that approximately one-third of students met the criteria for a mental disorder each year [1,2]. Mood and anxiety disorders are the most common diagnoses [1]. Research conducted since the beginning of the pandemic has generally indicated an increase in the prevalence of mental health problems among university students [3,4]. It is estimated that 30% to 40% of students experienced elevated levels of depression and anxiety symptoms during the first 2 years of the pandemic alone [5,6], with the highest rates reported for students in low- and middle-income countries [4]. Mental health problems among students have been associated with a range of negative outcomes, including reduced academic performance, lower engagement in campus life, poorer interpersonal relationships, and higher dropout rates [2,7-9]. However, evidence from upper-middle and high-income countries suggests that only 25% of students with a mental disorder receive treatment in a given year [2]. Common barriers to accessing mental health services include lack of time, stigma, high treatment costs, and insufficient capacity of university counseling centers to meet student demand [10,11]. Some of these barriers may be addressed through the provision of evidence-based digital mental health interventions (DMHIs) [2,12].

University students from various countries and regions have indicated a willingness to use digital tools for mental health–related information and support, highlighting benefits in improved access, privacy, confidentiality, and reduced costs [13]. Moreover, several systematic reviews have indicated that DMHIs targeting tertiary students can be effective for symptoms of depression and anxiety when delivered under trial conditions [12,14,15], with the most robust support for programs based on cognitive behavioral therapy (CBT) [12]. However, there is also evidence to suggest that the benefits of DMHIs are diminished in real-world settings [16,17], with high rates of dropout and a lack of sustained engagement frequently reported as key challenges to program effectiveness and successful implementation in both the general and student populations [15,18]. Although strategies such as the use of reminders, coaching, and engagement facilitation interventions offer the potential to redress suboptimal levels of engagement [19-21], their effectiveness has so far been variable, and a recent systematic review concluded that more research is needed to ensure that engagement strategies are effective and acceptable for end users [22].

Efforts to improve engagement with DMHIs are further complicated by a lack of consensus regarding how to define and measure engagement [22]. Most previous studies have focused on program use indicators of engagement (eg, number of logins or modules accessed) [23,24], reflecting the extent to which users are exposed to or interact with intervention materials. These indicators provide relatively objective information on use patterns and have been associated with small to moderate improvements in mental health outcomes in some instances [25,26]. However, they may not account for other factors that might be more closely related to positive outcomes, such as the degree to which users integrate and apply the therapeutic techniques from an intervention into their daily lives (ie, skill enactment) [27]. Recent recommendations, such as those provided by Li et al [24], Beintner et al [23], and Baumel [27], underscore the importance of investigating skill enactment as a distinct aspect of engagement that goes beyond the technology used to capture the cognitive and behavioral changes a user makes in their everyday life because of intervention [27-30]. Evaluating the strategies and techniques individuals implement from digital interventions is likely to be important both as a stand-alone outcome following intervention and because of its potential role as a mediator of treatment outcomes [31,32]. Skill enactment is likely to be an important outcome to consider in itself given suggestions that adopting healthy actions and routines may help maintain mental health or prevent future problems by improving the individual's ability to cope with stressors [31,32]. This may be especially relevant in interventions targeting relatively mild distress or symptoms where improvements in mental health outcomes may be small. Moreover, skill enactment has also been argued to be critical to understanding the impact of an intervention on symptom reduction [27,33]. This is because most therapeutic programs are predicated on the theoretical understanding that positive change occurs by facilitating the uptake of adaptive cognitive and behavioral skills [27].

The proposed importance of skill enactment has some support from recent studies that have demonstrated that DMHIs can effectively increase levels of skill enactment among adults with at least clinically mild symptoms of depression or anxiety at mid or postintervention assessments [34-36]. These studies also tend to indicate that improvements in skill enactment (ie, increase in the use of cognitive or behavioral skills, such as cognitive reframing and behavioral activation) are associated with small to moderate improvements in depression and anxiety [35-38]. However, a recent systematic review of digital CBT programs targeting depression and anxiety identified a gap in the literature on skill enactment among young people [30], and none of the included studies explored changes in skill enactment among a university student sample. Furthermore, qualitative or mixed methods investigations of skill enactment in digital interventions are rare [39], and we are not aware of any studies specifically exploring experiences of skill enactment among university students. As little is known about skill enactment in this population, there is a need for research examining skill enactment, its potential implications for mental health outcomes, and experiences among university students participating in DMHIs.

This Study

This mixed methods study investigated skill enactment among university students with mild to moderate distress who participated in a randomized controlled trial (RCT) of the Uni Virtual Clinic Lite (UVC-Lite), a low-intensity internet-based program designed to address depression, anxiety, and other mental health problems that commonly affect university students. The results from the trial have been submitted for publication [40] and indicated no differences between the UVC-Lite intervention and an attention-control program on any of the primary (symptoms of depression and generalized anxiety disorder) or secondary outcomes at postintervention measurement, 3-month, or 6-month follow-ups. Despite these nonsignificant results, understanding patterns and experiences related to skill enactment remains pertinent due to its potential relevance in maintaining mental health and improving future interventions. The primary aim of this study was to investigate whether the UVC-Lite intervention led to improvements in skill enactment. As per the study protocol, we also intended to investigate whether improvements in skill enactment mediated change in mental health outcomes (secondary aim). Although the intervention was not shown to be effective, we retained statistical analyses to explore the association between skill enactment and mental health outcomes. As there was a significant reduction in symptoms of depression and anxiety at posttest assessment across both conditions, there remained the potential to observe whether change in mental health outcomes was associated with skill enactment. We restricted our analysis to investigate the primary outcomes of depression and anxiety symptoms. We supplemented the quantitative component with partial data set analysis of semistructured interviews and free-text response data from participants who received access to UVC-Lite to examine patterns of skill enactment and explore participants' motivations for using the techniques delivered in the program. By collecting feedback from participants, it may be possible to refine the content and features of the intervention to enhance engagement and skill enactment.

We hypothesized that skill enactment would be higher in the UVC-Lite intervention condition than the attention-control condition at postintervention, 3-month, and 6-month follow-up assessments.

Methods

Overview

The CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) is presented in Multimedia Appendix 1 [41]. Data for this study were collected as part of an RCT that examined the effectiveness of UVC-Lite in reducing depression and anxiety symptoms among university students with mild to moderate levels of psychological distress. Participants were randomly assigned to either the UVC-Lite intervention or an attention-control program for a period of 6 weeks. A full description of the trial methods and primary outcomes, including information on allocation concealment, randomization procedures, and blinding, is provided in the main trial paper [40].

Participants

RCT Data

Trial participants were 487 undergraduate and postgraduate students, recruited and randomized into the trial between August 2021, and May 2022. Follow-up data were collected by January 2023. Students from all universities in Australia were invited to participate in the trial using a range of recruitment methods. These included targeted social media advertising, posts in university-related social media groups, and advertisements in university newsletters. In addition, assistance was sought from student associations and advocacy departments, student housing services, university marketing teams, survey management departments, university counseling and well-being clinics, and academic course coordinators to disseminate information about the trial to students via email and flyers. A total of 2865 students clicked on the study invitation and were screened for eligibility. Of these, 523 (18.25%) participants were initially screened into the trial, completed baseline measures, and were randomized to the UVC-Lite intervention (n=261, 49.9%) or attention-control group (n=262, 50.1%). A further 6.9% (36/523) of the participants were excluded after randomization due to being older than the eligible age range. Respondents were eligible for inclusion if they were (1) enrolled at an Australian university and resided in Australia, (2) aged between 18 and 25 years, and (3) scored between 8 and 17 on the Distress Questionnaire-5, indicating mild to moderate levels of psychological distress [42]. Students diagnosed with bipolar disorder, schizophrenia, posttraumatic stress disorder, or a personality disorder could participate if they met eligibility criteria and were receiving support or treatment for their disorder. Students who were ineligible due to high levels of psychological distress were provided with help-seeking resources.

Follow-Up Interviews

Purposive sampling was used to recruit participants for interview. All participants (131/243, 53.9%) randomized to the intervention condition who completed the postintervention survey were invited to provide additional feedback by telephone or videoconference about their experience with the modules. Of these, 38.9% (51/131) of the respondents indicated a willingness to provide their feedback. From this group, 33% (17/51) of the participants responded to the subsequent email invitation and 24% (12/51) participated in an interview. Reasons for not participating included loss of contact (3/51, 6%) and being overseas (2/51, 4%). In addition, 1 follow-up email was sent to participants if there was no response to the initial email. No incentives were offered for participation in the interview.

Intervention and Attention-Control Conditions

Intervention Condition: UVC-Lite

The UVC-Lite is a low-intensity internet-based intervention comprising 12 modules that target the common mechanisms underlying mental health problems in university students. Content for the intervention was drawn from the Uni Virtual Clinic, an web-based mental health intervention for university students that was shown to be effective in reducing symptoms of social anxiety and improving academic self-efficacy compared to a waitlist control group in a previous trial [43,44]. Each UVC-Lite module comprises a 3- to 6-minute video presenting psychoeducation about a mental health problem and a therapeutic technique designed to address the problem. The videos provide practical examples and guidance on how to apply the techniques in daily life. The modules also include optional exercises designed to encourage the practice of therapeutic techniques, self-monitoring quizzes (included in modules 1, 3, 4, 5, 7, and 11), and links to help-seeking resources provided at their university or in the community if necessary. The twelve modules include (1) dealing with depression and low mood (behavioral activation), (2) tackling negative and anxious thoughts (cognitive reframing), (3) dealing with anxiety (cognitive reframing, breathing, and grounding techniques), (4) managing study issues: procrastination and time management (practical strategies for time management), (5) perfectionism (challenging perfectionistic thoughts and behaviors), (6) coping with stress (mindfulness practice), (7) managing sleep issues (sleep hygiene), (8) social anxiety and shyness (behavioral experiments and exposure techniques), (9) relationships and loneliness (communication skills and social support), (10) social media use (practical strategies for reducing social media use), (11) body image (body functionality appreciation), and (12) thoughts of suicide (safety planning). Modules were delivered twice weekly over the 6-week intervention period.

Attention-Control Condition: General Health Information

Participants assigned to the attention-control condition received 2 emails per week containing a web link to a PDF document providing information on topics related to general health rather than mental health, approximately matched for completion time with the active intervention. A total of 12 topics were covered, including bone health, sun exposure, food hygiene, dietary supplements, kidney health, microbes, household burns, respiratory viruses, heart health, allergens, posture, and pancreas health. A similar program has previously been shown to be associated with no therapeutic reductions in depression [45].

Data Collection

Overview

At baseline, the following demographic data were collected: age, gender, ethnicity, living situation, employment status, study discipline and year of degree, study load, and international or domestic student status. Quantitative data on skill enactment, depression symptoms, and generalized anxiety symptoms were obtained using web-based self-report questionnaires at baseline, postintervention assessment (emailed to participants 6 weeks after completion of the baseline survey), and 3- and 6-month follow-ups (emailed to participants 3 and 6 months after receiving the postintervention survey). Participants received 2 email reminders to finish the postintervention and follow-up surveys. Qualitative data were collected at postintervention assessments (free-text responses) and through interviews conducted with participants following the completion of postintervention assessments between April and August 2022.

Skill Enactment

We assessed skill enactment using a list of 14 items developed to assess the frequency of enacting therapeutic techniques from the UVC-Lite program (eg, “I made time for activities that make me feel better” and “I challenged my thinking to be more realistic and helpful”). At each time point, participants rated how often they used each of the skills over the past 2 weeks on a 5-point scale using the following response options: 0 (never), 1 (rarely), 2 (sometimes), 3 (often), and 4 (very often). Item development was informed by existing CBT skills use questionnaires [46,47]. Exploratory factor analysis was conducted on the skill enactment items before computing total scores. The results of this analysis indicated that the scale could be treated as a single factor. One item was dropped from the measure due to a low factor loading, resulting in a 13-item measure (Multimedia Appendix 2 [48-50]). The internal consistency of the scale was acceptable (Cronbach α=0.84). The 13 items yielded a summed total score ranging from 0 to 52. Higher scores indicate more frequent skill enactment.

Mental Health Outcomes

Primary mental health outcomes were symptoms of depression, measured by the Patient Health Questionnaire-9 (PHQ-9) [51], and generalized anxiety disorder, measured by the Generalized Anxiety Disorder-7 (GAD-7) scale [52]. The PHQ-9 comprises 9 items rated on a 4-point scale, ranging from 0 (not at all) to 3 (nearly every day). Item scores are summed to produce an overall severity score ranging from 0 to 27, with higher scores indicating greater symptom severity (0-4: no symptoms, 5-9: mild symptoms, 10-14: moderate symptoms, and 15-27: severe symptoms). The GAD-7 comprises 7 items rated on the same 4-point scale as the PHQ-9. Summed scores produce an overall severity score ranging from 0 to 21 (0-4: no symptoms, 5-9: mild symptoms, 10-14: moderate symptoms, and 15-21: severe symptoms). Higher scores indicate greater symptom severity. Both scales have demonstrated robust psychometric properties in general population and university student samples [53-56]. In the current sample, internal consistency was acceptable at baseline (PHQ-9: Cronbach α=0.80; GAD-7: Cronbach α=0.84).

Program Use

JavaScript code was used to track intervention group participants' uptake and use behaviors. We report specific indicators in this study to provide an indication of exposure to the therapeutic content, including the number of modules accessed (ie, whether participants clicked to open each module), number of videos started (ie, >0% of the video watched), and number of therapeutic activities accessed (based on whether participants clicked through to access the exercise). For the control group, we tracked the number of modules accessed.

Qualitative Data

Semistructured interviews were conducted by the first author (HMJ), with ongoing supervision provided by the senior author (LMF). Overall, 7 interviews were conducted via Zoom (Zoom Video Communications), a web-based teleconferencing platform, and the remaining 5 interviews were conducted via telephone. The interview guide was developed by the researchers to investigate participants' thoughts about and use of the video modules and exercises, their suggestions for implementation in universities, and any barriers to the modules being used by university students. The full list of interview questions is provided in Multimedia Appendix 3, although only a subset of questions was the focus of investigation in this paper. Interviews were audio-recorded and transcribed verbatim using a speech-to-text transcription application (Otter.ai; Otter.ai, Inc). Transcripts were corrected by the first author while listening to the audio-recorded interviews.

Participants' satisfaction with and views of the intervention were also assessed during postintervention surveys by asking participants open-ended questions regarding whether there was anything they liked about the program, whether there was anything they disliked about the program, and whether they would change anything about the program.

Data Analysis

Quantitative Data

Quantitative data were analyzed using SPSS for Windows (version 28.0; IBM Corp), and significance was set at $P<.05$. Primary analyses were undertaken on an intention-to-treat basis. Mixed model repeated measures (MMRM) ANOVA was conducted to compare mean scores on skill enactment between the intervention and control groups, with measurement occasion as a within-groups factor and condition as a between-groups factor. MMRM techniques incorporate all available data under the assumption that data are missing at random [57]. Within-person variation was modeled using unstructured covariance matrices, and *df* was estimated using Satterthwaite correction. Subgroup analyses of intervention group participants who started ≥1 videos versus all of the control group were conducted to estimate the effectiveness of UVC-Lite among participants who accessed at least some of the therapeutic content. We opted to use this indicator of program use in the analyses as use of the videos was considered sufficient to promote acquisition and enactment of the therapeutic skills in this intervention and because the module exercises were optional. The influence of skill enactment on primary outcomes (PHQ-9 and GAD-7) was explored using MMRM ANOVA, including 3-way interaction terms for time×condition×skill enactment, as well as all main effects and 2-way interactions. Time-varying skill enactment, measured simultaneously with mental health outcomes, was used in the models. The 3-way interaction tested whether changes in outcome over time differed more for those in the intervention group with greater skill enactment, relative to the control group. The time × skill enactment interaction tested how a change in skill enactment was related to symptoms across different time points.

Qualitative Data

Qualitative data were analyzed using inductive content analysis [58,59], as per the analytic steps outlined by Vears and Gillam [60]. This approach is well suited to answering applied research questions [60]. After reading each of the interviews and free-text responses, the first author extracted all data potentially relevant to the research question to develop a working data set (ie, extracts referring to any aspects of in-program exercises or skill enactment) [61]. Data extracts were drawn from all (12/12, 100%) interview participants and 16.5% (17/103) of the participants who provided free-text responses during the postintervention survey, resulting in the analysis of data from 11.9% (29/243) of intervention group participants. The data then went through several iterations of inductive coding and similar codes were clustered together into content categories. Reporting was structured around the categories, drawing on illustrative quotes from interview transcripts and open-ended responses. Data extracts are presented verbatim, though some words (eg, "like" and "um") were removed to improve readability. Participants are identified by their gender, a coded number, and the type of data provided (eg, M1, interview=male, interview participant 1). Quirkos (offline version; Quirkos Software) was used to analyze qualitative data.

Ethical Considerations

The trial was registered at the Australian New Zealand Clinical Trials Registry (ACTRN12621000375853) and the ethical aspects of the research were approved by the Australian National University Human Research Ethics Committee (protocol #2020/412). All participants provided informed consent by ticking a checkbox on the web after viewing an information sheet containing key details about the study. To protect participant privacy, all data were deidentified. Participants who completed all 4 surveys were entered into a prize draw to win 1 of 10 Aus $100 (US $66.7) gift vouchers.

Results

Sample Characteristics and Participation Rates

The flow of participants through the study is presented in Figure 1 [40,62]. Of the 487 students who were eligible, randomized, and completed baseline measures, 265 (54.4%) completed postintervention surveys, 105 (21.6%) completed 3-month follow-up surveys, and 172 (35.3%) completed 6-month follow-up surveys. Table 1 presents the demographic data for the full sample by trial condition. Students from 40 academic institutions across Australia participated. Most participants were female (355/487, 72.9%), with an average age of 20.62 (SD 2.12) years. Participants were predominantly domestic students

(411/487, 84.4%) living either with family (209/487, 42.9%) or in on-campus accommodation (152/487, 31.2%). Most (328/487, 67.4%) participants were in some form of paid employment. Most students were studying full time (449/487, 92.2%) for an undergraduate degree (416/487, 85.4%); *t* tests (2-tailed) and chi square analyses demonstrated that there were no significant differences between the intervention and control conditions on any of the baseline demographic variables. Similarly, we found no significant differences between conditions on each of the outcomes of interest at baseline: skill enactment (t_{483}=−1.06; *P*=.29), depression ($t_{471.98}$=1.59; *P*=.11), and anxiety (t_{485}=1.41; *P*=.16). At preintervention, skill enactment was negatively correlated with both depression (*r*=−0.35; *P*<.001) and generalized anxiety (*r*=−0.21; *P*<.001) symptoms.

Figure 1. Flow of participants through the study. This figure is an abbreviated version of the flow diagram presented in the main trial paper by Farrer et al [40], which is published under Creative Commons Attribution 4.0 International License [62]. The figure has been adapted to include numbers on interview participation. DQ5: Distress Questionnaire-5; UVC-Lite: Uni Virtual Clinic Lite.

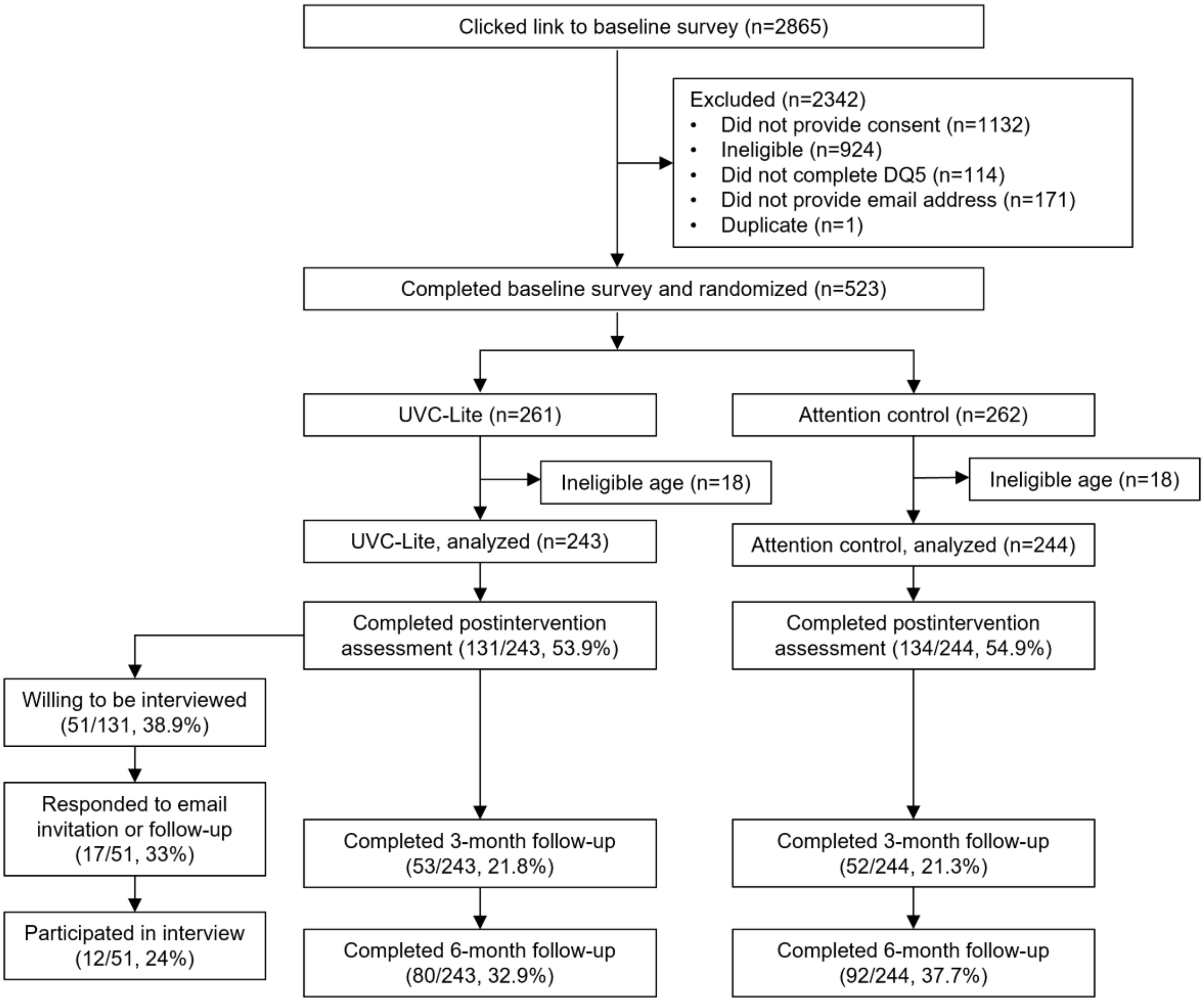


Table 1. Sample characteristics by condition.

Characteristic	UVC-Lite[a] (n=243)	Attention control (n=244)	Chi-square (*df*)	*t* test (*df*)	*P* value
Gender, n (%)			0.4 (2)	—[b]	.82
Male	52 (21.4)	58 (23.8)			
Female	180 (74.1)	175 (71.7)			
Other	11 (4.5)	11 (4.5)			
Ethnicity, n (%)			3.2 (4)	—	.53
Aboriginal and/or Torres Strait Islander, or Pacific Islander	6 (2.5)	2 (0.8)			
African or Middle Eastern	6 (2.5)	6 (2.5)			
Asian or Indian	65 (26.7)	75 (30.7)			
Caucasian or European	154 (63.4)	152 (62.3)			
Other	11 (4.5)	8 (3.3)			
Current living situation, n (%)			4.6 (6)	—	.59
With parents or family	100 (41.2)	109 (44.7)			
On-campus housing	82 (33.7)	70 (28.7)			
Friends off-campus	24 (9.9)	21 (8.6)			
Others off-campus	19 (7.8)	18 (7.4)			
With partner, children, or both	10 (4.1)	17 (7)			
Alone	7 (2.9)	9 (3.7)			
Other	1 (0.4)	0 (0)			
Hours per week in paid employment, n (%)			6.1 (5)	—	.30
None	89 (36.6)	70 (28.7)			
1-9	53 (21.8)	58 (23.8)			
10-19	68 (28)	67 (27.5)			
20-29	24 (9.9)	33 (13.5)			
30-39	8 (3.3)	15 (6.1)			
≥40	1 (0.4)	1 (0.4)			
Discipline of degree studied, n (%)[c]			5.5 (7)	—	.60
Health/medicine	83 (38.4)	71 (32)			
Arts/social sciences	37 (17.1)	42 (18.9)			
Science	39 (18.1)	45 (20.3)			
Engineering/computing	21 (9.7)	28 (12.6)			
Business/economics	11 (5.1)	14 (6.3)			
Law/criminology	14 (6.5)	15 (6.8)			
Education	11 (5.1)	6 (2.7)			
Tertiary preparation course	0 (0)	1 (0.5)			
Year of degree, n (%)[d]			4.3 (3)	—	.23
First-year undergraduate	93 (38.4)	86 (35.4)			
Later-year undergraduate	107 (44.2)	112 (46.1)			
Honors	5 (2.1)	13 (5.3)			
Postgraduate	37 (15.3)	32 (13.2)			
Study load, n (%)			0.4	—	.51

Characteristic	UVC-Lite[a] (n=243)	Attention control (n=244)	Chi-square (*df*)	*t* test (*df*)	*P* value
Full time	226 (93)	223 (91.4)			
Part time	17 (7)	21 (8.6)			
Student status, n (%)			0.1	—	.79
Domestic	204 (84)	207 (84.8)			
International	39 (16)	37 (15.2)			
Age (y), mean (SD)	20.67 (2.15)	20.57 (2.09)	—	−0.55 (485)	.58

[a]UVC-Lite: Uni Virtual Clinic Lite.

[b]Not applicable.

[c]Discipline of degree data provided by 216 participants in the UVC-Lite condition and 222 participations in the attention-control condition.

[d]Year of degree data provided by 242 participants in the UVC-Lite condition and 243 participants in the attention-control condition.

Program Use

The use of the UVC-Lite intervention among intervention group participants was low. Although 69.1% (168/243) of the participants accessed ≥1 (mean 5.10, SD 5.08) modules, more than half (129/243, 53.1%; mean 2.48, SD 3.89) of the participants did not start any of the videos, 28.8% (70/243) started 1 to 5 videos, 12.3% (30/243) started 6 to 11 videos, and only 5.8% (14/243) started all 12 videos. Engagement with the therapeutic activities was lower than engagement with the videos. Most participants (157/243, 64.6%; mean 1.31, SD 2.66) did not access any of the therapeutic activities; 27.2% (66/243) accessed 1 to 5 activities, and only 8.2% (20/243) of participants accessed ≥6 activities. Participants who received access to the attention-control program accessed a mean of 4.78 (SD 5.05) modules. More information on patterns of program use is provided in the study by Farrer et al [40].

Change in Skill Enactment

Table 2 presents the observed means and SDs for the outcomes of interest at each measurement occasion across the 2 conditions, and Table 3 provides the estimates of fixed effects from the MMRM ANOVA models. There was no significant overall interaction effect between measurement occasion and condition for skill enactment ($F_{3,215.36}$=0.50; *P*=.68) based on MMRM analyses. Planned contrasts demonstrated no significant interactions between time and condition on skill enactment scores at postintervention, 3-month, or 6-month follow-up assessments, indicating no differences in skill enactment between conditions. When subgroup analyses among those who started ≥1 videos (vs. the control group) were conducted, the overall effect of the intervention on skill enactment scores remained nonsignificant ($F_{3,179.78}$=2.08; *P*=.11). However, exploratory post hoc contrasts demonstrated that intervention group participants who accessed ≥1 videos had higher levels of skill enactment than participants in the control condition at postintervention ($t_{231.96}$=1.99; *P*=.048; 95% CI 0.02-4.01) and 6-month follow-up assessment ($t_{186.84}$=2.09; *P*=.04; 95% CI 0.13-4.64), though not at 3-month follow-up ($t_{132.85}$=0.58; *P*=.57; 95% CI –2.07 to 3.77). We also explored whether a change in skill enactment over time was modified by reliable improvement (reliable improvement vs no reliable improvement. Reliable improvement was defined as a decrease of ≥4 points on either the PHQ-9 or GAD-7 scales from pre to postintervention). The interaction between time, condition, and improvement was not significant ($F_{3,207.14}$=1.13; *P*=.34), but there was a significant 2-way interaction between time and improvement ($F_{3,207.14}$=8.78; *P*<.001). Post hoc contrasts indicated that, regardless of condition, participants demonstrating reliable improvement reported greater skill enactment at postintervention assessment ($t_{282.72}$=3.97; *P*<.001; 95% CI 2.53-7.52) and 3 months ($t_{133.09}$=2.02; *P*=.045; 95% CI 0.09-8.23) than those who did not. This was not significant at the 6-month follow-up ($t_{216.90}$=1.81; *P*=.07; 95% CI –0.23 to 5.39).

Table 2. Observed means for skill enactment, depression, and anxiety scores at baseline, postintervention, and at 3- and 6-month follow-up.

Outcome	UVC-Lite[a]		Attention control	
	Participants, n (%)	Scores, mean (SD)	Participants, n (%)	Scores, mean (SD)
Skill enactment				
Baseline	243 (100)	27.25 (8.24)	242 (99.2)	26.46 (8.09)
Postintervention	124 (51.0)	27.76 (7.31)	129 (52.9)	26.57 (8.14)
3-month follow-up	52 (21.4)	26.52 (8.97)	50 (20.5)	27.00 (9.14)
6-month follow-up	78 (32.1)	27.47 (7.12)	91 (37.3)	26.27 (6.54)
Depression (PHQ-9[b])				
Baseline	243 (100)	9.70 (5.11)	244 (100)	9.02 (4.34)
Postintervention	131 (53.9)	9.04 (5.31)	134 (54.9)	8.24 (4.80)
3-month follow-up	53 (21.8)	9.06 (5.92)	52 (21.3)	8.13 (4.57)
6-month follow-up	80 (32.9)	9.58 (6.10)	92 (37.7)	9.02 (5.60)
Generalized anxiety (GAD-7[c])				
Baseline	243 (100)	7.84 (4.53)	244 (100)	7.28 (4.13)
Postintervention	130 (53.5)	6.75 (4.19)	134 (54.9)	6.03 (3.90)
3-month follow-up	53 (21.8)	7.21 (4.84)	52 (21.3)	6.60 (4.36)
6-month follow-up	80 (32.9)	8.15 (5.52)	92 (37.7)	7.20 (4.77)

[a]UVC-Lite: Uni Virtual Clinic Lite.

[b]PHQ-9: Patient Health Questionnaire-9.

[c]GAD-7: Generalized Anxiety Disorder Scale-7.

Table 3. Estimates of fixed effects from mixed model repeated measures models.

Outcome and source	*F* test (*df*)	*P* value
Skill enactment		
Intercept	5607.77 (1, 357.91)	<.001
Time	0.72 (3, 215.36)	.54
Condition	1.10 (1, 357.91)	.29
Time×condition	0.50 (3, 215.36)	.68
Depression score (PHQ-9[a])		
Intercept	644.17 (1, 567.48)	<.001
Time	0.36 (3, 229.22)	.78
Condition	3.30 (1, 567.48)	.07
Skill enactment	134.61 (1, 541.87)	<.001
Time×condition	1.68 (3, 229.22)	.17
Time×skill enactment	1.19 (3, 241.10)	.31
Condition×skill enactment	1.36 (1, 541.87)	.24
Time×condition×skill enactment	1.69 (3, 241.10)	.17
Generalized anxiety score (GAD-7[b])		
Intercept	419.26 (1, 544.66)	<.001
Time	2.05 (3, 221.68)	.11
Condition	0.23 (1, 544.66)	.63
Skill enactment	73.08 (1, 535.11)	<.001
Time×condition	1.13 (3, 221.68)	.34
Time×skill enactment	1.16 (3, 233.71)	.32
Condition×skill enactment	0.15 (1, 535.11)	.70
Time×condition×skill enactment	1.11 (3, 233.71)	.35

[a]PHQ-9: Patient Health Questionnaire-9.

[b]GAD-7: Generalized Anxiety Disorder Scale-7.

Influence of Skill Enactment on Primary Outcomes

In MMRM models testing the effect of skill enactment on mental health outcomes, the 3-way interaction between time, condition, and skill enactment was not significant (PHQ-9: $F_{3,241.10}$=1.69; P=.17; GAD-7: $F_{3,233.71}$=1.11; P=.35), indicating no differential effects of the intervention for those with greater skill enactment. The 2-way interactions between time and skill enactment were also nonsignificant (PHQ-9: $F_{3,241.10}$=1.19; P=.32; GAD-7: $F_{3,233.71}$=1.16; P=.32). Planned contrasts confirmed these null results. However, a significant main effect of skill enactment was demonstrated on both depression ($F_{1,541.87}$=134.61; P<.001) and anxiety ($F_{1,535.11}$=73.08; P<.001) scores, indicating that higher levels of skill enactment were associated with lower symptom levels among both intervention and control group participants across time points. We also examined the influence of skill enactment among intervention group participants who started ≥1 videos (compared against all of the control group). While the main effect of skill enactment remained significant (PHQ-9: $F_{1,492.97}$=130.47; P<.001; GAD-7: $F_{1,501.66}$=071.95; P<.001), the 3-way interaction (PHQ-9: $F_{3,211.95}$=1.47; P=.23; GAD-7: $F_{3,204.92}$=0.9; P=.44) and the 2-way interaction effects between time and skill enactment (PHQ-9: $F_{3,211.95}$=1.28; P=.28; GAD-7: $F_{3,204.92}$=0.9; P=.44) was not significant.

Qualitative Analysis

Description of Participants

Students from 18 academic institutions were represented in the qualitative analysis, 79% (23/29) of whom were female. The mean age of participants who provided qualitative data was 20.97 (SD 2.23) years. Most participants were domestic students (26/29, 90%) studying at the undergraduate level (21/29, 72%) across a range of disciplines (eg, Health and Medicine and Arts and Social Sciences). Unlike most participants randomized to the intervention, engagement among this group was very high; participants accessed a mean of 10.86 (SD 2.37) modules, started a mean of 8.28 (SD 4.33) videos, and accessed a mean of 4.14 (SD 3.62) exercises.

Qualitative Categories

One category was identified from the interviews and open-ended survey responses regarding patterns of skill enactment among participants: (1) limited initial and ongoing skill enactment. Three categories were identified regarding the factors that influenced participants' use of the program techniques: (1) relevance and perceived fit, (2) practicality and ease of implementation, and (3) exploring novel approaches and reinforcing familiar ones.

Limited Initial and Ongoing Skill Enactment

Overall, participants' initial engagement with the exercises designed to facilitate the practice of the therapeutic techniques was limited. Where participants did engage, most stated that they tried only some of the exercises, and several participants described a pattern of engagement that lacked depth in terms of the time, effort, or attention they gave to the exercises. For some, this involved reading through the exercises without any or limited further engagement:

> *Like the exercises and things like that, like I read through them. I didn't really, I didn't do any of the exercises myself, but I did...I did read through them and look at them and they seemed pretty, pretty good.* [M1, interview]

Others described completing the exercises, but not in a "fully focused" way [M3, interview], and some commented that their engagement did not go beyond their initial interaction with the exercise in the web-based platform, especially if the exercise included a writing component:

> *For ones that required actually writing things down, if I had something in the moment that I wanted to...that I thought of as I was, like, reading through the exercises then I put them down. But I didn't, like you know, take it as homework and adopt it within my lifestyle.* [F2, interview, writing exercises]

In general, participants did not describe continuing to enact skills from the program after initially interacting with the web-based exercises, and one student noted that they were not "consistent" in their use of skills (F5, interview). At the same time, participants described finding the techniques they did try, which included a mixture of thought records, pleasant events schedules, behavioral experiments for social anxiety, procrastination strategies, body functionality appreciation, or mindfulness exercises, to be practical and useful, and several stated that they would use them in the future if needed:

> *But I can probably look back at that. So, I've downloaded all of them though, for like later reference if I need them.* [F6, interview, behavioral experiments for social anxiety]

Relevance and Perceived Fit

Several participants described the relevance of certain strategies or techniques to their specific symptoms or challenges as a key influence on their decisions to initially engage with the exercises. Techniques that were not seen to be relevant to students' symptoms were less likely to be used, whereas those that were more relevant were more likely to be used:

> *I think at the time, it probably applied to me a bit more. I think that I was probably feeling a bit low in the social skills area of life. Yeah.* [F8, interview, behavioral experiment for social anxiety]

Perceived compatibility between an individual's personal characteristics, their goals, and the techniques taught in the intervention was also mentioned by some participants as positively or negatively influencing their use of specific techniques. For example, one participant commented on the fit between a cognitive exercise (a thought record) and their personal characteristics:

> *And it...it kind of...I am a very logical and, like, a person who likes to process things. So, like, I liked the layout of it, that you could fill in a table and see...work it out.* [F1, interview, thought record exercise]

Although mentioned less often, another participant stated that their use of the techniques was motivated by their preexisting mental health goals:

> *Okay, so pretty much like I'd already been wanting to meditate. And I feel like the videos just like reinforced it.* [M4, interview, mindfulness exercise]

To address issues of relevance and fit, participants recommended tailoring or personalizing program content, for example, by modifying content according to the needs of the student or by highlighting that users can work through the exercises based on their needs and preferences. One participant suggested that "the exercises might be improved by having a few variations by which one might manage their issues" within each module [M5, free-text response].

Practicality and Ease of Implementation

The degree to which participants perceived the strategies to be practical and easy to implement in everyday life also affected their initial willingness to engage with and enact techniques. For example, techniques that were seen as enjoyable, rewarding, and having a tangible impact on mood (such as the pleasant events schedule) were easier to implement for some participants:

> *So then incorporating, I think it was like one achievable task that will give you a sense of achievement and one task for fun. And I'm not used to like scheduling in things just for fun. So that felt really rewarding.* [F4, interview, pleasant events schedule]

Conversely, some participants stated that they were less likely to use exercises that had a significant web-based writing component or that involved multiple steps (such as the thought records):

> *Yeah, there were...there was a lot to like write down and it was maybe if it was like a thought exercise like with just like a couple of steps to do I maybe would have done it more but just because I'm not at my laptop all day.* [F3, interview, thought record exercise]

Other factors that made it challenging to implement the exercises, in general, and in terms of particular topics, included forgetting and lack of time due to other demands:

> *Ah, time, I suppose...But I've had a very kind of busy schedule, which means I probably should have prioritized it, but I didn't as much as I should have.* [F6, interview, body functionality appreciation exercise]

> *...I did download them, but I just forgot about them.* [F13, free-text response]

Participants indicated that changing the presentation of the thought records to be less cognitively demanding (such as by breaking down the exercise into smaller steps) and providing calls to action (eg, using weekly reminders) could facilitate engagement. Several participants also suggested that the exercises should be more interactive:

> *Possibly make them a bit more interactive—maybe add a challenge for the week or something (to help with motivation for these exercises/activities).* [F19, free-text response]

Exploring Novel Approaches and Reinforcing Familiar Ones

Participants also stated that their use of certain strategies or techniques was shaped by previous exposure. Several participants commented that they used strategies that were new to them because the concepts sparked curiosity and motivated them to try out new ways of coping or relating to themselves:

> *And I don't know why I practiced that one in particular, because I'm happy with my body image. But it was just a really interesting notion of 'don't think about, like how your body looks, think about how it performs,' and things like that.* [F5, interview, body functionality appreciation exercise]

In contrast, some participants indicated that they used strategies with which they were already familiar because this reinforced previous knowledge and behaviors:

> *There was like, there was one of them...that is like, you know, I think I've heard it before. I think it's like a common thing. But it was just good to reinforce it. I guess.* [M2, interview, behavioral experiments for social anxiety]

> *Improving sleep...yeah, yeah so, I already kind of try and do some of that but it's helpful to kind of have a refresher.* [F6, interview, sleep hygiene exercise]

Discussion

Principal Findings

This mixed methods study aimed to examine skill enactment among university students participating in an RCT of a low-intensity video-based intervention (UVC-Lite). The intervention was not found to be effective in improving self-reported skill enactment. There was also no evidence to indicate that participants who reported greater skill enactment experienced greater improvement in symptoms. This finding is not surprising given that skill enactment remained relatively stable over time and because we did not observe an overall effect of the intervention on primary mental health outcomes in the main trial [40]. These results raise the possibility that it may be difficult to facilitate changes in skill enactment via the unguided use of brief transdiagnostic modules in a sample of university students with mild to moderate levels of distress. Several previous studies have demonstrated that DMHIs can lead to improvements in skill enactment, with some evidence indicating that these improvements mediate improvements in symptoms of depression and anxiety [35,36,38,63]. These factors may provide users with more opportunities to consolidate, practice, and benefit from program skills through increased repetition and reinforcement.

Use of the materials was also low in this trial (ie, approximately half of the participants either did not start any of the videos or accessed any of the activities). These results are comparable to previous studies of DMHIs for students and young people [64-66], including a recent study by Bolinski et al [67], which found that although most students started at least 1 session of an automated transdiagnostic intervention, they completed an average of only 2 (25%) out of the 8 mandatory modules. The low rates of program use in this study suggest that a key explanation for our null results may be insufficient exposure to the therapeutic content. Our subgroup analyses lent some support to this explanation, as contrasts indicated that the program may be effective in improving skill enactment at postintervention and 6-month follow-up for participants who accessed at least some of the intervention materials. However, this result should be interpreted with caution given that the analyses were post hoc and represented small effects, the overall model was not significant, and an effect was not observed at 3-month follow-up. Furthermore, improvements in skill enactment among this group also did not translate to improvements in depression or anxiety, perhaps because the analysis was not sufficiently powered to detect small effects that have been observed in other studies [36].

It is also possible that relatively high baseline levels of self-reported skill enactment (49.1%-83.6% of participants indicated that they used each of the skills, at least sometimes; median 70.8%, IQR 62.4%-78.5%) and low baseline severity of depression and anxiety symptoms affected our ability to detect change. We did find partial support for this idea in our post hoc analysis comparing skill enactment among participants who reported reliable improvement against those who did not. Higher skill enactment was reported at postintervention and 3-month follow-up among those who demonstrated reliable improvement in symptoms. However, there was no evidence to suggest that the intervention led to greater improvements in skill enactment in this group. Further investigation is needed to examine the conditions under which this intervention may be effective and to explore the relationship between skill enactment and outcomes in a larger sample of students with a higher need or more severe symptoms (ie, those with greater potential to experience reliable or clinically significant change).

The results of our qualitative analysis help to further elucidate the processes underlying skill enactment. Participants who provided qualitative data constituted a unique subset of users, as most had high levels of program use. However, even among this group, skill enactment was limited and often did not extend beyond students' initial interactions with the web-based exercises. Instead, participants tended to highlight that they planned to use the skills if they experienced a worsening of symptoms. Similar experiences were expressed by participants in 2 previous studies of adolescent and adult users of DMHIs, wherein the authors observed that some users adopt a passive and reactive approach to applying treatment principles [39,68]. Both in these previous studies and our study, this approach appeared to be underpinned by participants' judgments concerning their current and potential future need for strategies to address various mental health concerns, suggesting that tailoring intervention content and activities based on an initial assessment of symptoms may improve skill enactment. However, other studies have reported that students are not more likely to choose content matched to their symptoms based on clinical assessment [69,70], and some research has linked choice to improved outcome [71], which suggests that tailoring approaches may need to strike a balance between student-identified concerns and clinical recommendations.

In addition, participants also highlighted that they were motivated to use skills that were easily integrated into daily routines rather than those that involved a significant writing component (eg, thought record exercises). These findings are consistent with past research on the co-design and evaluation of DMHIs among young people, which found that this group has a preference for practical and easy to use information and an aversion to text-heavy content [72-74]. This might suggest that behavioral activities (eg, pleasant events scheduling) may be more useful among students with mild to moderate distress than cognition-focused exercises, consistent with the suggestions offered by Titov et al [32] that different actions may be important at different levels of mental health. Yet individual differences also appeared to have a substantial influence on participants' decisions to enact certain skills, for example, some participants highlighted that their engagement decisions were underpinned by judgments regarding whether the approach was consistent with their personal characteristics (eg, approach to problem-solving and preexisting goals), and others pointed to the newness or, conversely, the familiarity of the strategy as positively influencing skill enactment.

Suggestions offered by participants to improve engagement with the therapeutic activities included changing the presentation of writing-based activities to be less cognitively demanding (eg, using tunneling features), providing reminders, and enhancing interactivity (eg, using weekly challenges). These results align broadly with the recommendations for persuasive system design provided by Baumel et al [27], who suggested that calls to action (eg, reminding and inspiring users to use a technique) and load reduction (eg, gradual presentation of information to reduce cognitive load) may help encourage users of digital interventions to integrate therapeutic activities in everyday life. Gamification elements, such as the use of points and reward systems or leaderboards, might also be used to enhance interactivity and support motivation among students who face barriers to skill enactment, with a previous RCT of an internet-based intervention finding that adult participants who received concurrent access to a smartphone app encouraging the completion of small challenges in daily life had greater skill enactment than those who did not have concurrent access to the app [75].

Implications

This study was motivated by calls for research to measure and report indicators of meaningful and active engagement such as skill enactment in digital interventions [23,24,27]. However, low rates of program use made it challenging to draw meaningful conclusions regarding whether low-intensity interventions such as the UVC-Lite have the capacity to improve skill enactment among students, or about the potential effects of skill enactment on outcomes. Our results related to low program use and high dropout are common in several trials of digital CBT interventions for young people with relatively mild symptoms [18,76]. They suggest that a low perceived need for intervention among students with mild to moderate symptoms may undermine their motivation to fully engage with CBT-based programs. Lack of perceived need can reduce program uptake and participation [77], but it may also result in a lack of active engagement with therapeutic skills and limited ongoing behavior changes among those who do participate. This suggests that an important priority for future research would be to establish whether low-intensity programs can effectively improve levels of skill enactment and mental health outcomes among students with more severe symptoms. In addition, the potential for the large-scale implementation of unguided interventions in university settings suggests that another direction for future research would be to ascertain the most effective approaches for prevention and early intervention among students. Alternative approaches such as those that focus on preparation for difficult times or establishing healthy routines may be needed with this population. This might include not only mental health education to promote appropriate support seeking, but also well-being interventions (eg, mindfulness-based and positive psychological interventions), stress management programs, strength-based planning, lifestyle interventions, and CBT-based programs targeting less stigmatized challenges such as sleep disturbance.

The study also points to potential pathways for improving intervention design to promote skill enactment. Our qualitative findings related to individual differences in skill enactment suggest that 1 promising approach may be to tailor therapeutic content to the needs and preferences of individual users. For instance, real-time analysis of user behaviors could be used to dynamically tailor therapeutic recommendations and feedback [27], continuous engagement with certain activities might trigger recommendations for more complex activities, while a lack of engagement could trigger recommendations for alternative activities or feedback to address barriers. Tailoring therapeutic recommendations based on other user characteristics is also likely to be beneficial (eg, through assessment of user-identified concerns, goals, or existing mental health knowledge). However, this will require a better understanding of the processes underlying skill enactment among users of DMHIs, including factors that influence initial and ongoing cognitive and behavioral change in everyday life, as well as factors that affect the enactment of different types of skills, especially among different student groups. Qualitative methods involving consultation with end users and the application of structured frameworks of behavior change during data collection and interpretation (eg, theoretical domains framework [78]) may help address these aims.

Strengths and Limitations

Strengths of the study include the large sample relative to other trials testing web-based interventions in tertiary student populations [79-81], recruitment of participants from universities across Australia rather than a single institution, and a focus on maximizing ecological validity. However, high ecological validity was also a limitation due to high rates of attrition and low adherence. This may have led to inaccurate estimates of the treatment effect, although we used rigorous MMRM models to account for all available data. Our use of qualitative data to supplement the quantitative analysis represents an additional strength. However, we relied on a self-selected sample of students to provide these data, nearly all of whom were highly engaged. This may have resulted in a sample skewed toward those who had predominantly positive experiences of the intervention, and our findings may not apply to students who did not use or only minimally used the intervention. Questions regarding skill enactment were also asked as part of a larger qualitative study designed to examine students' views on implementing the program within universities, and thus we may not have captured other experiences related to skill enactment. Furthermore, we did not ask participants about potential factors other than the program that may have influenced skill enactment during the measurement period.

In addition, participation among males, gender-diverse students, and international students was low, which may limit the generalizability of the results to other populations. Although this reflects differences in the help-seeking and participation in mental health research more generally [82-84], ongoing efforts to improve participation are important to ensure DMHIs are acceptable and effective for these groups. Given the high number of international students attending university in Australia [85], future studies might also consider the utility of offering interventions in multiple languages to maximize accessibility for international and culturally and linguistically diverse students. We also relied on brief self-report scales to measure outcomes. Although assessments included validated measures of depression and anxiety, a bespoke scale was used to assess skill enactment as none of the standardized measures available at the time of trial design adequately captured the range of techniques covered in the UVC-Lite intervention [46,47,86]. Titov et al [32] have since developed and evaluated a comprehensive self-report questionnaire of mentally healthy actions that is not tied to a specific therapeutic approach and which may assist other researchers to evaluate skill enactment using a standardized approach. However, it will also be important for future studies to assess the fidelity of skill enactment (ie, the extent to which individuals enact skills in line with how they are taught), for example, through evaluations of homework or therapist ratings [30]. Moreover, delivery of the 3-month follow-up assessments was delayed by approximately 1 month for a proportion of participants due to an administrative error.

Conclusions

This study indicated that a transdiagnostic low-intensity video-based program was not effective in improving skill enactment among university students with mild to moderate levels of psychological distress. Our null results may be due to several factors, including characteristics of the intervention, low program use, low baseline severity of participants' symptoms, and high levels of baseline skill enactment. The results of our qualitative analysis suggested that initial and ongoing skill enactment was limited even among participants with high program use, and individual differences appeared to have a substantial influence on decisions to enact skills. Ongoing research with university students is needed to determine who can benefit from low-intensity digital interventions, what types of digital interventions may be most effective, and how we can improve engagement, both in terms of program use and skill enactment, to ensure that the potential benefits of DMHIs are realized. Qualitative and theory-informed research may lead to a better understanding of the mechanisms underlying skill enactment.

Acknowledgments

The authors would like to thank Dr Amelia Gulliver, Dr Liana Lench, and Dr Penelope Hasking who were involved as coinvestigators in the main trial. The authors would also like to thank Natasha Katruss for the assistance with developing the intervention materials and data collection. Funding for this project was provided by Australian Rotary Health. HMJ is supported by an Australian Government Research Training Program Scholarship and the Grace Groom Memorial Scholarship (2021). LMF is supported by an Australian Research Council Discovery Early Career Researcher Award (ARC DECRA; DE190101382). ALC is supported by a National Health and Medical Research Council (NHMRC Fellowship 1173146). The funders were not involved in the writing, reviewing, or approval of the manuscript, nor were they involved in the decision to submit the manuscript for publication.

Authors' Contributions

LMF, PJB, and ALC obtained project funding and designed the trial. HMJ developed the research goals and aims and designed the skill enactment measure in consultation with all other authors. LMF and HMJ implemented the study. HMJ analyzed the data. HMJ drafted the manuscript. All authors provided critical revisions to and approved the final manuscript.

Conflicts of Interest

PJB, ALC, and LMF are codevelopers of the Uni Virtual Clinic Lite (UVC-Lite) intervention. LMF was involved in the development of the original UVC-Lite intervention. The other authors declare no conflicts of interest.

Multimedia Appendix 1
CONSORT-eHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 92 KB - mental_v11i1e53794_app1.pdf]

Multimedia Appendix 2
Results of the exploratory factor analysis on skill enactment items.
[DOCX File , 18 KB - mental_v11i1e53794_app2.docx]

Multimedia Appendix 3
Interview questions.
[DOCX File , 17 KB - mental_v11i1e53794_app3.docx]

References

1. Auerbach RP, Mortier P, Bruffaerts R, Alonso J, Benjet C, Cuijpers P, WHO WMH-ICS Collaborators. WHO World Mental Health surveys international college student project: prevalence and distribution of mental disorders. J Abnorm Psychol 2018 Oct;127(7):623-638 [FREE Full text] [doi: 10.1037/abn0000362] [Medline: 30211576]
2. Bruffaerts R, Mortier P, Auerbach RP, Alonso J, Hermosillo De la Torre AE, Cuijpers P, WHO WMH-ICS Collaborators. Lifetime and 12-month treatment for mental disorders and suicidal thoughts and behaviors among first year college students. Int J Methods Psychiatr Res 2019 Jun;28(2):e1764 [FREE Full text] [doi: 10.1002/mpr.1764] [Medline: 30663193]
3. Buizza C, Bazzoli L, Ghilardi A. Changes in college students mental health and lifestyle during the COVID-19 pandemic: a systematic review of longitudinal studies. Adolescent Res Rev 2022 Aug 03;7(4):537-550. [doi: 10.1007/s40894-022-00192-7]
4. Li W, Zhao Z, Chen D, Peng Y, Lu Z. Prevalence and associated factors of depression and anxiety symptoms among college students: a systematic review and meta-analysis. J Child Psychol Psychiatry 2022 Nov 16;63(11):1222-1230. [doi: 10.1111/jcpp.13606] [Medline: 35297041]
5. Wang C, Wen W, Zhang H, Ni J, Jiang J, Cheng Y, et al. Anxiety, depression, and stress prevalence among college students during the COVID-19 pandemic: a systematic review and meta-analysis. J Am Coll Health 2023 Oct 01;71(7):2123-2130. [doi: 10.1080/07448481.2021.1960849] [Medline: 34469261]
6. Deng J, Zhou F, Hou W, Silver Z, Wong CY, Chang O, et al. The prevalence of depressive symptoms, anxiety symptoms and sleep disturbance in higher education students during the COVID-19 pandemic: a systematic review and meta-analysis. Psychiatry Res 2021 Jul;301:113863 [FREE Full text] [doi: 10.1016/j.psychres.2021.113863] [Medline: 33984824]
7. Mojtabai R, Stuart EA, Hwang I, Eaton WW, Sampson N, Kessler RC. Long-term effects of mental disorders on educational attainment in the National Comorbidity Survey ten-year follow-up. Soc Psychiatry Psychiatr Epidemiol 2015 Oct 17;50(10):1577-1591 [FREE Full text] [doi: 10.1007/s00127-015-1083-5] [Medline: 26082040]
8. Sharp J, Theiler S. A review of psychological distress among university students: pervasiveness, implications and potential points of intervention. Int J Adv Counselling 2018 Feb 2;40(3):193-212. [doi: 10.1007/s10447-018-9321-7]
9. Salzer MS. A comparative study of campus experiences of college students with mental illnesses versus a general college sample. J Am Coll Health 2012;60(1):1-7. [doi: 10.1080/07448481.2011.552537] [Medline: 22171723]
10. Lui JC, Sagar-Ouriaghli I, Brown JS. Barriers and facilitators to help-seeking for common mental disorders among university students: a systematic review. J Am Coll Health 2022 Sep 09:1-9. [doi: 10.1080/07448481.2022.2119859] [Medline: 36084266]
11. Stallman HM. University counselling services in Australia and New Zealand: activities, changes, and challenges. Aust Psychol 2020 Nov 12;47(4):249-253. [doi: 10.1111/j.1742-9544.2011.00023.x]
12. Harrer M, Adam SH, Baumeister H, Cuijpers P, Karyotaki E, Auerbach RP, et al. Internet interventions for mental health in university students: a systematic review and meta-analysis. Int J Methods Psychiatr Res 2019 Jun 26;28(2):e1759 [FREE Full text] [doi: 10.1002/mpr.1759] [Medline: 30585363]
13. Montagni I, Tzourio C, Cousin T, Sagara JA, Bada-Alonzi J, Horgan A. Mental health-related digital use by university students: a systematic review. Telemed J E Health 2020 Feb 01;26(2):131-146. [doi: 10.1089/tmj.2018.0316] [Medline: 30888256]
14. Farrer L, Gulliver A, Chan JK, Batterham PJ, Reynolds J, Calear A, et al. Technology-based interventions for mental health in tertiary students: systematic review. J Med Internet Res 2013 May 27;15(5):e101 [FREE Full text] [doi: 10.2196/jmir.2639] [Medline: 23711740]

15. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital mental health interventions for depression, anxiety, and enhancement of psychological well-being among college students: systematic review. J Med Internet Res 2019 Jul 22;21(7):e12869 [FREE Full text] [doi: 10.2196/12869] [Medline: 31333198]
16. Mohr DC, Weingardt KR, Reddy M, Schueller SM. Three problems with current digital mental health research . . . and three things we can do about them. Psychiatr Serv 2017 May 01;68(5):427-429 [FREE Full text] [doi: 10.1176/appi.ps.201600541] [Medline: 28412890]
17. Gilbody S, Littlewood E, Hewitt C, Brierley G, Tharmanathan P, Araya R, REEACT Team. Computerised cognitive behaviour therapy (cCBT) as treatment for depression in primary care (REEACT trial): large scale pragmatic randomised controlled trial. BMJ 2015 Nov 11;351:h5627 [FREE Full text] [doi: 10.1136/bmj.h5627] [Medline: 26559241]
18. Becker TD, Torous JB. Recent developments in digital mental health interventions for college and university students. Curr Treat Options Psych 2019 Jun 14;6(3):210-220. [doi: 10.1007/s40501-019-00178-8]
19. Batterham PJ, Calear AL, Sunderland M, Kay-Lambkin F, Farrer LM, Christensen H, et al. A brief intervention to increase uptake and adherence of an internet-based program for depression and anxiety (enhancing engagement with psychosocial interventions): randomized controlled trial. J Med Internet Res 2021 Jul 27;23(7):e23029 [FREE Full text] [doi: 10.2196/23029] [Medline: 34313595]
20. Farrer L, Christensen H, Griffiths KM, Mackinnon A. Internet-based CBT for depression with and without telephone tracking in a national helpline: randomised controlled trial. PLoS One 2011;6(11):e28099 [FREE Full text] [doi: 10.1371/journal.pone.0028099] [Medline: 22140514]
21. Mohr DC, Schueller SM, Tomasino KN, Kaiser SM, Alam N, Karr C, et al. Comparison of the effects of coaching and receipt of app recommendations on depression, anxiety, and engagement in the IntelliCare platform: factorial randomized controlled trial. J Med Internet Res 2019 Aug 28;21(8):e13609 [FREE Full text] [doi: 10.2196/13609] [Medline: 31464192]
22. Gan DZ, McGillivray L, Larsen ME, Christensen H, Torok M. Technology-supported strategies for promoting user engagement with digital mental health interventions: a systematic review. Digit Health 2022 Jun 01;8:20552076221098268 [FREE Full text] [doi: 10.1177/20552076221098268] [Medline: 35677785]
23. Beintner I, Vollert B, Zarski AC, Bolinski F, Musiat P, Görlich D, et al. Adherence reporting in randomized controlled trials examining manualized multisession online interventions: systematic review of practices and proposal for reporting standards. J Med Internet Res 2019 Aug 15;21(8):e14181 [FREE Full text] [doi: 10.2196/14181] [Medline: 31414664]
24. Li SH, Achilles MR, Werner-Seidler A, Beames JR, Subotic-Kerry M, O'Dea B. Appropriate use and operationalization of adherence to digital cognitive behavioral therapy for depression and anxiety in youth: systematic review. JMIR Ment Health 2022 Aug 17;9(8):e37640 [FREE Full text] [doi: 10.2196/37640] [Medline: 35976180]
25. Gan DZ, McGillivray L, Han J, Christensen H, Torok M. Effect of engagement with digital interventions on mental health outcomes: a systematic review and meta-analysis. Front Digit Health 2021;3:764079 [FREE Full text] [doi: 10.3389/fdgth.2021.764079] [Medline: 34806079]
26. Donkin L, Christensen H, Naismith SL, Neal B, Hickie IB, Glozier N. A systematic review of the impact of adherence on the effectiveness of e-therapies. J Med Internet Res 2011 Aug 05;13(3):e52 [FREE Full text] [doi: 10.2196/jmir.1772] [Medline: 21821503]
27. Baumel A. Therapeutic activities as a link between program usage and clinical outcomes in digital mental health interventions: a proposed research framework. J Technol Behav Sci 2022 Feb 22;7(2):234-239. [doi: 10.1007/s41347-022-00245-7]
28. Piotrowska PJ, Tully LA, Lenroot R, Kimonis E, Hawes D, Moul C, et al. Mothers, fathers, and parental systems: a conceptual model of parental engagement in programmes for child mental health-connect, attend, participate, enact (CAPE). Clin Child Fam Psychol Rev 2017 Jun 2;20(2):146-161 [FREE Full text] [doi: 10.1007/s10567-016-0219-9] [Medline: 27914017]
29. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
30. Jackson HM, Calear AL, Batterham PJ, Ohan JL, Farmer GM, Farrer LM. Skill enactment and knowledge acquisition in digital cognitive behavioral therapy for depression and anxiety: systematic review of randomized controlled trials. J Med Internet Res 2023 May 31;25:e44673 [FREE Full text] [doi: 10.2196/44673] [Medline: 37256673]
31. Hundt NE, Mignogna J, Underhill C, Cully JA. The relationship between use of CBT skills and depression treatment outcome: a theoretical and methodological review of the literature. Behav Ther 2013 Mar;44(1):12-26. [doi: 10.1016/j.beth.2012.10.001] [Medline: 23312423]
32. Titov N, Dear BF, Bisby MA, Nielssen O, Staples LG, Kayrouz R, et al. Measures of daily activities associated with mental health (things you do questionnaire): development of a preliminary psychometric study and replication study. JMIR Form Res 2022 Jul 05;6(7):e38837 [FREE Full text] [doi: 10.2196/38837] [Medline: 35788101]
33. Donkin L, Hickie IB, Christensen H, Naismith SL, Neal B, Cockayne NL, et al. Rethinking the dose-response relationship between usage and outcome in an online intervention for depression: randomized controlled trial. J Med Internet Res 2013 Oct 17;15(10):e231 [FREE Full text] [doi: 10.2196/jmir.2771] [Medline: 24135213]
34. Birney AJ, Gunn R, Russell JK, Ary DV. MoodHacker mobile web app with email for adults to self-manage mild-to-moderate depression: randomized controlled trial. JMIR Mhealth Uhealth 2016 Jan 26;4(1):e8 [FREE Full text] [doi: 10.2196/mhealth.4231] [Medline: 26813737]

35. Forand NR, Barnett JG, Strunk DR, Hindiyeh MU, Feinberg JE, Keefe JR. Efficacy of guided iCBT for depression and mediation of change by cognitive skill acquisition. Behav Ther 2018 Mar;49(2):295-307 [FREE Full text] [doi: 10.1016/j.beth.2017.04.004] [Medline: 29530267]
36. Terides MD, Dear BF, Fogliati VJ, Gandy M, Karin E, Jones MP, et al. Increased skills usage statistically mediates symptom reduction in self-guided internet-delivered cognitive-behavioural therapy for depression and anxiety: a randomised controlled trial. Cogn Behav Ther 2018 Jan;47(1):43-61. [doi: 10.1080/16506073.2017.1347195] [Medline: 28724338]
37. Morgan AJ, Mackinnon AJ, Jorm AF. Behavior change through automated e-mails: mediation analysis of self-help strategy use for depressive symptoms. Behav Res Ther 2013 Feb;51(2):57-62 [FREE Full text] [doi: 10.1016/j.brat.2012.11.002] [Medline: 23262114]
38. Bisby MA, Titov N, Dear BF, Karin E, Wilhelms A, Nugent M, et al. Examining change in the frequency of adaptive actions as a mediator of treatment outcomes in internet-delivered therapy for depression and anxiety. J Clin Med 2022 Oct 11;11(20):6001 [FREE Full text] [doi: 10.3390/jcm11206001] [Medline: 36294322]
39. Berg M, Malmquist A, Rozental A, Topooco N, Andersson G. Knowledge gain and usage of knowledge learned during internet-based CBT treatment for adolescent depression - a qualitative study. BMC Psychiatry 2020 Sep 10;20(1):441 [FREE Full text] [doi: 10.1186/s12888-020-02833-4] [Medline: 32912154]
40. Farrer LM, Gulliver A, Katruss N, Fassnacht DB, Kyrios M, Batterham PJ. A novel multi-component online intervention to improve the mental health of university students: Randomised controlled trial of the Uni Virtual Clinic. Internet Interv 2019 Dec;18:100276 [FREE Full text] [doi: 10.1016/j.invent.2019.100276] [Medline: 31890625]
41. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
42. Batterham PJ, Sunderland M, Carragher N, Calear AL, Mackinnon AJ, Slade T. The Distress Questionnaire-5: population screener for psychological distress was more accurate than the K6/K10. J Clin Epidemiol 2016 Mar;71:35-42. [doi: 10.1016/j.jclinepi.2015.10.005] [Medline: 26464194]
43. Farrer LM, Gulliver A, Katruss N, Bennett K, Bennett A, Ali K, et al. Development of the Uni Virtual Clinic: an online programme for improving the mental health of university students. Br J Guid Counc 2020 Feb 20;48(3):333-346. [doi: 10.1080/03069885.2020.1729341]
44. Farrer LM, Gulliver A, Katruss N, Fassnacht DB, Kyrios M, Batterham PJ. A novel multi-component online intervention to improve the mental health of university students: randomised controlled trial of the Uni Virtual Clinic. Internet Interv 2019 Dec;18:100276 [FREE Full text] [doi: 10.1016/j.invent.2019.100276] [Medline: 31890625]
45. Glozier N, Christensen H, Naismith S, Cockayne N, Donkin L, Neal B, et al. Internet-delivered cognitive behavioural therapy for adults with mild to moderate depression and high cardiovascular disease risks: a randomised attention-controlled trial. PLoS One 2013 Mar 26;8(3):e59139 [FREE Full text] [doi: 10.1371/journal.pone.0059139] [Medline: 23555624]
46. Terides MD, Dear BF, Karin E, Jones MP, Gandy M, Fogliati VJ, et al. The frequency of actions and thoughts scale: development and psychometric validation of a measure of adaptive behaviours and cognitions. Cogn Behav Ther 2016 Apr;45(3):196-216. [doi: 10.1080/16506073.2016.1149876] [Medline: 26926484]
47. Jarrett RB, Vittengl JR, Clark LA, Thase ME. Skills of cognitive therapy (SoCT): a new measure of patients' comprehension and use. Psychol Assess 2011 Sep;23(3):578-586 [FREE Full text] [doi: 10.1037/a0022485] [Medline: 21319902]
48. O’connor BP. SPSS and SAS programs for determining the number of components using parallel analysis and Velicer’s MAP test. Behav Res Methods Instrum Comput 2000 Sep;32(3):396-402. [doi: 10.3758/BF03200807]
49. Cattell RB. The Scree test for the number of factors. Multivariate Behav Res 1966 Apr 01;1(2):245-276. [doi: 10.1207/s15327906mbr0102_10] [Medline: 26828106]
50. Costello AB, Osborne J. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. Pract Assess Res Eval 2005;10:7. [doi: 10.4135/9781412995627.d8]
51. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
52. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
53. Löwe B, Decker O, Müller S, Brähler E, Schellberg D, Herzog W, et al. Validation and standardization of the generalized anxiety disorder screener (GAD-7) in the general population. Med Care 2008 Mar;46(3):266-274. [doi: 10.1097/MLR.0b013e318160d093] [Medline: 18388841]
54. Kroenke KM, Spitzer RL, Williams JB, Löwe B. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. Gen Hosp Psychiatry 2010;32(4):345-359. [doi: 10.1016/j.genhosppsych.2010.03.006] [Medline: 20633738]
55. Byrd-Bredbenner C, Eck K, Quick V. GAD-7, GAD-2, and GAD-mini: psychometric properties and norms of university students in the United States. Gen Hosp Psychiatry 2021 Mar;69:61-66. [doi: 10.1016/j.genhosppsych.2021.01.002] [Medline: 33571925]
56. Kim YE, Lee B. The Psychometric Properties of the Patient Health Questionnaire-9 in a sample of Korean university students. Psychiatry Investig 2019 Dec;16(12):904-910 [FREE Full text] [doi: 10.30773/pi.2019.0226] [Medline: 31870089]

57. Verbeke G, Molenberghs G. Linear Mixed Models for Longitudinal Data. New York, NY: Springer; 2000.
58. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
59. Erlingsson C, Brysiewicz P. A hands-on guide to doing content analysis. Afr J Emerg Med 2017 Sep;7(3):93-99 [FREE Full text] [doi: 10.1016/j.afjem.2017.08.001] [Medline: 30456117]
60. Vears DF, Gillam L. Inductive content analysis: a guide for beginning qualitative researchers. Focus Health Prof Educ 2022 Mar 31;23(1):111-127. [doi: 10.11157/fohpe.v23i1.544]
61. Braun V, Clarke V. Thematic Analysis: A Practical Guide. Thousand Oaks, CA: Sage Publications; 2022.
62. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2024-08-20]
63. Seeley JR, Sheeber LB, Feil EG, Leve C, Davis B, Sorensen E, et al. Mediation analyses of internet-facilitated cognitive behavioral intervention for maternal depression. Cogn Behav Ther 2019 Jul 12;48(4):337-352 [FREE Full text] [doi: 10.1080/16506073.2018.1513554] [Medline: 30311850]
64. Musiat P, Conrod P, Treasure J, Tylee A, Williams C, Schmidt U. Targeted prevention of common mental health disorders in university students: randomised controlled trial of a transdiagnostic trait-focused web-based intervention. PLoS One 2014;9(4):e93621 [FREE Full text] [doi: 10.1371/journal.pone.0093621] [Medline: 24736388]
65. Radomski AD, Bagnell A, Curtis S, Hartling L, Newton AS. Examining the usage, user experience, and perceived impact of an internet-based cognitive behavioral therapy program for adolescents with anxiety: randomized controlled trial. JMIR Ment Health 2020 Feb 07;7(2):e15795 [FREE Full text] [doi: 10.2196/15795] [Medline: 32022692]
66. D'Adamo L, Paraboschi L, Grammer AC, Fennig M, Graham AK, Yaeger LH, et al. Reach and uptake of digital mental health interventions based on cognitive-behavioral therapy for college students: a systematic review. J Behav Cogn Ther 2023 Jun;33(2):97-117 [FREE Full text] [doi: 10.1016/j.jbct.2023.05.002] [Medline: 37724304]
67. Bolinski F, Kleiboer A, Neijenhuijs K, Karyotaki E, Wiers R, de Koning L, et al. Challenges in recruiting university students for web-based indicated prevention of depression and anxiety: results from a randomized controlled trial (ICare Prevent). J Med Internet Res 2022 Dec 14;24(12):e40892 [FREE Full text] [doi: 10.2196/40892] [Medline: 36515986]
68. Bendelin N, Hesser H, Dahl J, Carlbring P, Nelson KZ, Andersson G. Experiences of guided internet-based cognitive-behavioural treatment for depression: a qualitative study. BMC Psychiatry 2011 Jun 30;11(1):107. [doi: 10.1186/1471-244x-11-107]
69. Palacios JE, Richards D, Palmer R, Coudray C, Hofmann SG, Palmieri PA, et al. Supported internet-delivered cognitive behavioral therapy programs for depression, anxiety, and stress in university students: open, non-randomised trial of acceptability, effectiveness, and satisfaction. JMIR Ment Health 2018 Dec 14;5(4):e11467 [FREE Full text] [doi: 10.2196/11467] [Medline: 30552076]
70. Ciharova M, Cuijpers P, Amanvermez Y, Riper H, Klein AM, Bolinski F, et al. Use of tailoring features and reasons for dropout in a guided internet-based transdiagnostic individually-tailored cognitive behavioral therapy for symptoms of depression and/or anxiety in college students. Internet Interv 2023 Dec;34:100646 [FREE Full text] [doi: 10.1016/j.invent.2023.100646] [Medline: 38099094]
71. Hilvert-Bruce Z, Rossouw PJ, Wong N, Sunderland M, Andrews G. Adherence as a determinant of effectiveness of internet cognitive behavioural therapy for anxiety and depressive disorders. Behav Res Ther 2012 Aug;50(7-8):463-468. [doi: 10.1016/j.brat.2012.04.001] [Medline: 22659155]
72. Gericke F, Ebert DD, Breet E, Auerbach RP, Bantjes J. A qualitative study of university students' experience of Internet‐based CBT for depression. Couns and Psychother Res 2021 Aug 17;21(4):792-804. [doi: 10.1002/capr.12465]
73. Hanley T, Wyatt C. A systematic review of higher education students' experiences of engaging with online therapy. Couns and Psychother Res 2020 Dec 14;21(3):522-534. [doi: 10.1002/capr.12371]
74. Gan DZ, McGillivray L, Larsen ME, Bloomfield T, Torok M. Promoting engagement with smartphone apps for suicidal ideation in young people: development of an adjunctive strategy using a lived experience participatory design approach. JMIR Form Res 2023 Jun 06;7:e45234 [FREE Full text] [doi: 10.2196/45234] [Medline: 37279058]
75. Boettcher J, Magnusson K, Marklund A, Berglund E, Blomdahl R, Braun U, et al. Adding a smartphone app to internet-based self-help for social anxiety: a randomized controlled trial. Comput Human Behav 2018 Oct;87(6):98-108. [doi: 10.1016/j.chb.2018.04.052]
76. Clarke AM, Kuosmanen T, Barry MM. A systematic review of online youth mental health promotion and prevention interventions. J Youth Adolesc 2015 Jan;44(1):90-113. [doi: 10.1007/s10964-014-0165-0] [Medline: 25115460]
77. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
78. Atkins L, Francis J, Islam R, O'Connor D, Patey A, Ivers N, et al. A guide to using the theoretical domains framework of behaviour change to investigate implementation problems. Implement Sci 2017 Jun 21;12(1):77 [FREE Full text] [doi: 10.1186/s13012-017-0605-9] [Medline: 28637486]
79. Karyotaki E, Klein AM, Ciharova M, Bolinski F, Krijnen L, de Koning L, et al. Guided internet-based transdiagnostic individually tailored cognitive behavioral therapy for symptoms of depression and/or anxiety in college students: a randomized

controlled trial. Behav Res Ther 2022 Mar;150:104028 [FREE Full text] [doi: 10.1016/j.brat.2021.104028] [Medline: 35066365]
80. Day V, McGrath PJ, Wojtowicz M. Internet-based guided self-help for university students with anxiety, depression and stress: a randomized controlled clinical trial. Behav Res Ther 2013 Jul;51(7):344-351. [doi: 10.1016/j.brat.2013.03.003] [Medline: 23639300]
81. Rahmadiana M, Karyotaki E, Schulte M, Ebert DD, Passchier J, Cuijpers P, et al. Transdiagnostic internet intervention for Indonesian university students with depression and anxiety: evaluation of feasibility and acceptability. JMIR Ment Health 2021 Mar 05;8(3):e20036 [FREE Full text] [doi: 10.2196/20036] [Medline: 33666553]
82. Batterham PJ, Neil AL, Bennett K, Griffiths KM, Christensen H. Predictors of adherence among community users of a cognitive behavior therapy website. Patient Prefer Adherence 2008 Feb 02;2:97-105 [FREE Full text] [Medline: 19920949]
83. Woodall A, Morgan C, Sloan C, Howard L. Barriers to participation in mental health research: are there specific gender, ethnicity and age related barriers? BMC Psychiatry 2010 Dec 02;10:103 [FREE Full text] [doi: 10.1186/1471-244X-10-103] [Medline: 21126334]
84. LaMontagne AD, Shann C, Lolicato E, Newton D, Owen PJ, Tomyn AJ, et al. Mental health-related knowledge, attitudes and behaviours in a cross-sectional sample of australian university students: a comparison of domestic and international students. BMC Public Health 2023 Jan 25;23(1):170 [FREE Full text] [doi: 10.1186/s12889-023-15123-x] [Medline: 36698138]
85. Education at a glance. Organisation for Economic Cooperation and Development. URL: https://www.oecd-ilibrary.org/education/education-at-a-glance-2023_e13bef63-en [accessed 2024-04-29]
86. Kanter JW, Mulick PS, Busch AM, Berlin KS, Martell CR. The Behavioral Activation for Depression Scale (BADS): psychometric properties and factor structure. J Psychopathol Behav Assess 2006 Oct 25;29(3):191-202. [doi: 10.1007/s10862-006-9038-5]

Abbreviations

CBT: cognitive behavioral therapy
CONSORT-eHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
DMHI: digital mental health intervention
GAD-7: Generalized Anxiety Disorder-7
MMRM: mixed model repeated measures
PHQ-9: Patient Health Questionnaire-9
RCT: randomized controlled trial
UVC-Lite: Uni Virtual Clinic Lite

Edited by J Torous; submitted 18.10.23; peer-reviewed by J Passchier, DM Mirea; comments to author 12.02.24; revised version received 05.03.24; accepted 21.06.24; published 21.08.24.

Please cite as:
Jackson HM, Batterham PJ, Calear AL, Ohan JL, Farrer LM
Skill Enactment Among University Students Using a Brief Video-Based Mental Health Intervention: Mixed Methods Study Within a Randomized Controlled Trial
JMIR Ment Health 2024;11:e53794
URL: https://mental.jmir.org/2024/1/e53794
doi:10.2196/53794
PMID:

Original Paper

Patient Perspectives on AI for Mental Health Care: Cross-Sectional Survey Study

Natalie Benda[1], PhD; Pooja Desai[2], MA; Zayan Reza[3], MPH; Anna Zheng[4]; Shiveen Kumar[5], BS; Sarah Harkins[1], BSN, RN; Alison Hermann[6], MD; Yiye Zhang[7], PhD; Rochelle Joly[8], MD; Jessica Kim[7], MS; Jyotishman Pathak[7], PhD; Meghan Reading Turchioe[1], PhD, MPH, RN

[1]School of Nursing, Columbia University, New York, NY, United States
[2]Department of Biomedical Informatics, Columbia University, New York, NY, United States
[3]Mailman School of Public Health, Columbia University, New York, NY, United States
[4]Stuyvestant High School, New York, NY, United States
[5]College of Agriculture and Life Science, Cornell University, Ithaca, NY, United States
[6]Department of Psychiatry, Weill Cornell Medicine, New York, NY, United States
[7]Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, United States
[8]Department of Obstetrics and Gynecology, Weill Cornell Medicine, New York, NY, United States

Corresponding Author:
Natalie Benda, PhD
School of Nursing
Columbia University
560 168th Street
New York, NY, 10032
United States
Phone: 1 917 426 3069
Email: nb3115@cumc.columbia.edu

Abstract

Background: The application of artificial intelligence (AI) to health and health care is rapidly increasing. Several studies have assessed the attitudes of health professionals, but far fewer studies have explored the perspectives of patients or the general public. Studies investigating patient perspectives have focused on somatic issues, including those related to radiology, perinatal health, and general applications. Patient feedback has been elicited in the development of specific mental health care solutions, but broader perspectives toward AI for mental health care have been underexplored.

Objective: This study aims to understand public perceptions regarding potential benefits of AI, concerns about AI, comfort with AI accomplishing various tasks, and values related to AI, all pertaining to mental health care.

Methods: We conducted a 1-time cross-sectional survey with a nationally representative sample of 500 US-based adults. Participants provided structured responses on their perceived benefits, concerns, comfort, and values regarding AI for mental health care. They could also add free-text responses to elaborate on their concerns and values.

Results: A plurality of participants (245/497, 49.3%) believed AI may be beneficial for mental health care, but this perspective differed based on sociodemographic variables (all $P<.05$). Specifically, Black participants (odds ratio [OR] 1.76, 95% CI 1.03-3.05) and those with lower health literacy (OR 2.16, 95% CI 1.29-3.78) perceived AI to be more beneficial, and women (OR 0.68, 95% CI 0.46-0.99) perceived AI to be less beneficial. Participants endorsed concerns about accuracy, possible unintended consequences such as misdiagnosis, the confidentiality of their information, and the loss of connection with their health professional when AI is used for mental health care. A majority of participants (80.4%, 402/500) valued being able to understand individual factors driving their risk, confidentiality, and autonomy as it pertained to the use of AI for their mental health. When asked who was responsible for the misdiagnosis of mental health conditions using AI, 81.6% (408/500) of participants found the health professional to be responsible. Qualitative results revealed similar concerns related to the accuracy of AI and how its use may impact the confidentiality of patients' information.

Conclusions: Future work involving the use of AI for mental health care should investigate strategies for conveying the level of AI's accuracy, factors that drive patients' mental health risks, and how data are used confidentially so that patients can determine with their health professionals when AI may be beneficial. It will also be important in a mental health care context to ensure the patient–health professional relationship is preserved when AI is used.

(JMIR Ment Health 2024;11:e58462) doi:10.2196/58462

KEYWORDS

artificial intelligence; AI; mental health; patient perspectives; patients; public survey; application; applications; health care; health professionals; somatic issues; radiology; perinatal health; Black; professional relationship; patient-health; autonomy; risk; confidentiality; machine learning; digital mental health; computing; coding; mobile phone

Introduction

Background

The potential of artificial intelligence (AI) to transform health care has been touted since the early 2010s [1-4]. In health care applications, practitioners commonly operationalize AI by training machine learning algorithms using large retrospective data sets to perform human reasoning tasks, such as identifying issues (eg, anomalies in medical images), predicting events (eg, disease incidence), recommending treatments (eg, pharmacogenomics), detecting patterns (eg, finding symptom clusters), and generating text (eg, for clinical decision support rules).

AI has already made significant strides in the field of medical imaging, aiding health professionals at various stages, including improving image quality [5], guiding image acquisition [6], risk-stratifying images to be reviewed by a specialist (ie, a radiologist) [7,8], and interpreting images [9,10]. More recently, predictive AI has been leveraged to detect mental health–related issues, including major depressive disorders [11,12], stress, anxiety [13], bipolar disorder [14], and even suicide [15,16]. AI has also commonly been used for treatment selection in fields such as oncology and mental health [17,18]. Despite the demonstrated predictive accuracy of AI, relatively few of the predictive AI tools created are implemented in everyday clinical care [19], and even fewer tools have demonstrated a positive clinical impact compared to current standards of care [20-22]. Furthermore, of the 89 unique articles in 2 systematic reviews of clinical trials evaluating predictive AI, none focused on mental health–related conditions [20,21].

Due to the gap between the predictive AI's accuracy and its lack of observed impact on health outcomes, many researchers in many countries have studied health professionals' perceptions of AI-based tools and related implementation challenges [16,23-28]. However, patients' perspectives of AI have been understudied [29-31]. While some predictive AI developers may not intend for patients to view the AI's output on their own, it has become more likely that patients have access to predictive AI output due to recent advances in patient data ownership and access. The US 21st Century Cures Act, for example, prevents blocking information from patients, requiring health organizations and insurance providers to give patients access to their eHealth information without delay or expense [32]. This may result in a patient seeing a predictive AI risk score before a discussion with their health care team. In a 2020 predictive AI preimplementation study, health professionals stressed the importance of keeping the patient in the information loop when the AI predicts a risk or recommends a treatment to justify to the patient why they may require further support [24]. In addition to practical considerations, there is an ethical imperative to ensure patients understand how their data are being used, what predictive AI may reveal, and what the insight means, especially for sensitive issues, such as mental health care concerns [33]. Before we design solutions for communicating AI information to patients, it is important to understand the public's perceived benefits, comfort, concerns, and values related to AI use, particularly for mental health care [34].

To address the deficit in knowledge regarding patient perspectives on AI, Khullar et al [35] conducted a survey of a nationally representative panel of the US-based population. While most respondents reported a perceived benefit of using AI in health care, comfort with AI unsurprisingly varied based on the accuracy, transparency, and clinical application of AI (eg, reading a chest x-ray vs making a cancer diagnosis) [35]. The survey conducted by Khullar et al [35] focused on somatic applications of AI, leaving questions regarding public perceptions of AI for mental health care applications unanswered. However, others have explored narrower issues related to feedback on specific mental health care apps and specific prediction tasks (eg, the prediction of suicide) [36,37].

Objectives

AI applications for mental health are rapidly increasing as patients gain greater access and ownership of their data. Given the ethical concerns regarding the creation and use of AI and the stigmas surrounding mental health care, understanding patients' perceptions of whether and how AI may be appropriately used for mental health care is critical [38,39]. This study adapts and extends the survey conducted by Khullar et al [35] to evaluate patient perspectives on the use of AI for mental health care applications. We specifically surveyed members of the public to gain patient perspectives on AI applications for mental health. Khulllar et al [35] did not explore values regarding AI use, that is, what patients' priorities for effective, appropriate AI use for mental health care are. We also explored these values in this study using a bioethics-informed framework. The specific research questions (RQs) guiding our work were as follows:

- RQ 1: Do the public perceive AI to be beneficial for mental health care? RQ 1 Equity: Do perceived benefits differ by sociodemographic factors?
- RQ 2: How concerned is the public about common issues related to AI use in mental health care?
- RQ 3: What types of predictive tasks are the public comfortable with AI executing in mental health care applications?

- RQ 4: What are the public's values related to AI use for mental health care?

We also elicited open-ended responses from participants to add to their quantitative feedback.

Methods

Study Design

In our study, we conducted a 1-time, cross-sectional survey of US-based adults in September 2022. We sampled a general US adult population to elicit the public's perspectives on AI for mental health. We partnered with Prolific (Prolific Academic Ltd), a web-based survey sampling platform, to recruit participants. Prolific provides access to an international sample of verified users (>100,000 users residing in the United States) who are willing to be involved in survey research studies. Prolific matches eligible participants with research studies, streamlining the recruitment, data collection, and compensation processes. Prospective participants had to be verified Prolific users aged ≥18 years who were fluent in written and spoken English to be eligible.

Ethical Considerations

This study was approved by the BRANY Institutional Review Board (protocol 22-01024358). Before answering the questionnaire, participants read an information sheet and consented to participate in this specific study. Participants who completed the full questionnaire were compensated at an hourly rate of US $13.60 based on Prolific's policies.

Questionnaire Design

We designed our questionnaire such that it mimicked items asked in a previous study by Khullar et al [35] but applied to perspectives specifically related to AI for mental health, instead of AI for health care broadly. Question categories related to AI for mental health care were as follows: (1) perceived benefits of AI, (2) concerns about AI, and (3) comfort with using AI for specific predictive tasks. Adapting questions related to perceived benefits and concerns predominantly involved updating the terms "health" or "health care" to "mental health" or "mental health care," respectively. In the questionnaire developed by Khullar et al [35], questions regarding predictive tasks included those on reading a screening tool (ie, a chest x-ray); making a diagnosis for 2 different conditions, with 1 being more severe (pneumonia and cancer); and telling a patient they had either of the 2 aforementioned conditions and making a treatment recommendation. Our team worked with a trained psychiatrist to construct tasks following similar patterns but pertaining to mental health care, adding 2 more tasks (resulting in a total of 7 tasks) to explore more sensitive concepts relating to mental health. Multimedia Appendix 1 presents the questions in each of the aforementioned categories along with the question from which they were adapted as applicable.

We also extended the questionnaire to understand participant's values pertaining to AI design and implementation for mental health care to facilitate a more patient-centered design of future AI applications for mental health. This section asked patients to rate their level of importance regarding various statements pertaining to AI for mental health care, as informed by the constructs of MITRE's bioethical framework. Multimedia Appendix 1 displays the value statements presented to participants based on the relevant bioethics constructs.

In addition to the questions on perspectives and values, participants also answered questions on sociodemographic characteristics, including personal characteristics, health literacy, subjective numeracy, previous mental health care experience, and pregnancy history (results reported in a separate manuscript). The full battery of sociodemographic questions is presented in Multimedia Appendix 1.

Finally, following the sections regarding concerns and values, the survey contained open-ended questions to allow people to provide free-text responses with additional concerns or values.

We designed the battery of questions with input from experts in AI, human-centered design, and psychiatry and the author of the original survey from which the questions were adapted. The survey questions underwent 2 rounds of pilot testing to improve their comprehensibility and understand the amount of time needed to complete the questionnaire. The question-and-answer design was optimized and pilot tested for completion on both desktops and mobiles (ie, smartphones) to ensure those with different devices or preferences could participate in the study.

Participants

All participants were recruited from Prolific's survey sampling panel and were verified users who agreed to participate in research studies via the Prolific website. Our sample included those aged ≥18 years, residing in the United States, and with the ability to speak and read English. We recruited a sample representative of the adult US population in terms of age, race, and gender, according to the US Census. We initially recruited 530 survey respondents, of whom 30 (5.7%) did not begin the survey after reading the informed consent document, resulting in a total of 500 (94.3%) respondents. All 500 respondents finished the survey (0 incomplete responses) over a median time of 15 minutes and 24 seconds.

Data Collection

Our team designed and programmed the questionnaire using the Qualtrics XM (Qualtrics) platform. Participants received an invitation to complete the questionnaire through Prolific and then proceeded by clicking on a secure, anonymous link to Qualtrics. Participants could complete the survey using any smartphone, tablet, or computer, provided they had an internet connection. Participants then completed the questionnaire. There was no time limit for completing the questionnaire, and participants had the option to pause and resume completing the questionnaire at a later time. Participants also had the option to discontinue the survey at any time.

Data Analysis

Quantitative Analysis

The first level of analysis involved assessing descriptive statistics to understand trends in participant perceptions and values. We also selected an outcome of interest (perceived benefit of AI for mental health) and created a logistic regression model to better understand whether perceived benefits may differ by sociodemographic factors, specifically age, gender, race, education, financial resources, mental health history, and self-rated health literacy [40]. The α value for all analyses was set at .05, and the R software (version 4.2.1; The R Foundation) was used. An analysis of a subset of this data (only those of participants reporting female sex at birth) related to differences in perspectives based on pregnancy history has been reported in a separate manuscript [41].

Qualitative Analysis

We analyzed free-text responses through an inductive thematic analysis and a constant comparative process. One analyst initially reviewed the codes and created a draft codebook. Free-text responses to the 2 open-ended questions were analyzed using a singular coding scheme. A second analyst then used the coding scheme to independently dual code each free-text response. The analysts met with a third team member to resolve discrepancies, coding via consensus and updating the codebook throughout the discussion. Once detailed codes had been developed and 50% of the initial coding was completed, the team completed axial coding, coming up with higher-level summary themes to describe patterns in the detailed codes.

Results

Participant Characteristics

Table 1 describes the demographic makeup of the 500 adult, US-based survey respondents sampled using the Prolific platform [42]. Respondents were nationally representative based on race, age, and gender.

Table 1. Participant demographics (N=500).

Participant characteristics	Values
Age (y)	
Median (IQR)	46 (31-59)
Mean (SD; range)	46 (16; 18-93)
Gender, n (%)	
Women	249 (49.8)
Men	238 (47.6)
Transgender	1 (0.2)
Something else	9 (1.8)
Prefer not to answer	3 (0.6)
Race, n (%)	
Asian	25 (5)
Black or African American	66 (13.2)
White	388 (77.6)
Other or prefer not to answer[a]	21 (4.2)
Perceived financial resources, n (%)	
More than enough	65 (13)
Enough	271 (54.2)
Not enough	156 (31.2)
Prefer not to answer	8 (1.6)
Mental health history[b], n (%)	
Yes	215 (43)
No	271 (54.2)
Prefer not to answer	14 (2.8)
Health literacy[c], n (%)	
Adequate	369 (73.8)
Inadequate	131 (26.2)

[a]Answer options included American Indian or Alaskan Native, Native Hawaiian or Pacific Islander, and prefer not to answer, and multiple options could be selected.

[b]The question asked was "Have you ever been told that you have mental illness?"

[c]Measured using the Brief Health Literacy Screener developed by Chew et al [40].

RQ 1: Perceived Benefits of AI for Mental Health

Participants were first asked, "Overall, in the next 5 years, do you think AI will make mental health care in the United States..." Answer options included "much better," "somewhat better," "minimal change," "somewhat worse," "much worse," and "don't know." We computed a logistic regression model such that "much better" and "somewhat better" were classified as 1, and the other responses were classified as 0, excluding the 3 (0.6) participants, among the total 500 participants, answering, "don't know." Among 497 included respondents, 245 (49.3%) respondents believed that AI would make mental health care better or much better. Table 2 reveals that participants of Black or African American race (*P*=.04; odds ratio [OR] 1.76, 95% CI 1.03-3.05) and those with lower health literacy (*P*=.004; OR 2.19, 95% CI 1.29-3.78) were significantly more likely to endorse that AI would make mental health care somewhat or much better. Women, by contrast, were significantly less likely to endorse this statement (*P*=.046; OR 0.68, 95% CI 0.46-0.99).

Table 2. Logistic regression results: impact of sociodemographic variables on the perceived benefit of artificial intelligence (AI) for mental health.

Variable	β estimate	Odds ratio (95% CI)	*P* value
Intercept	0.378	—[a]	.37
Age (y)	–0.006	0.99 (0.98-1.01)	.36
Gender (woman)	–0.388	0.68 (0.46-0.99)	.046[b]
Race (Black or African American)	0.567	1.76 (1.03-3.05)	.04[b]
Perceived financial resources (not enough)	0.003	1.00 (0.66-1.51)	.99
Mental illness history (yes)	–0.044	0.96 (0.65-1.42)	.83
Health literacy (inadequate)	0.782	2.19 (1.29-3.78)	.004[b]

[a]Not available.

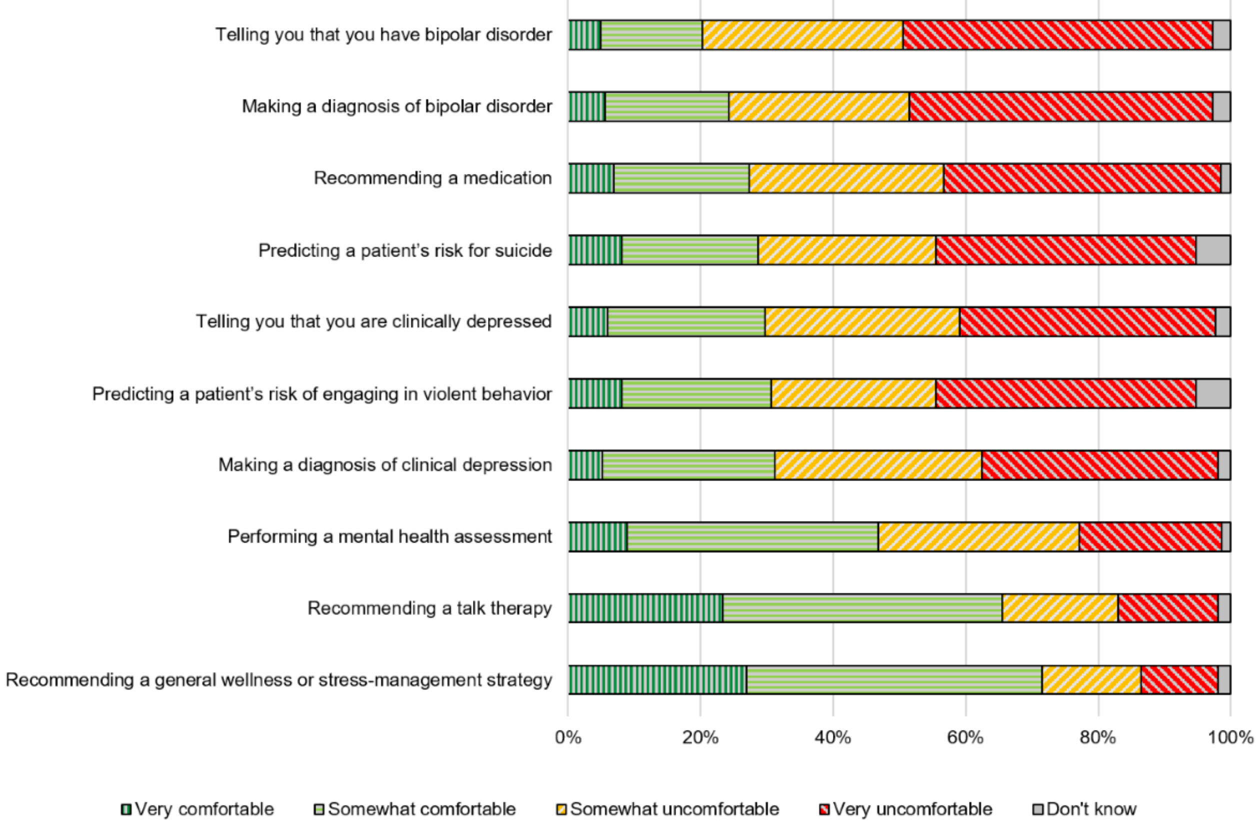
[b]Statistically significant based on α<.05.

Concerns, Comfort With Predictive Tasks, and Values (Quantitative)

RQ 2: Concerns Regarding AI for Mental Health Care

On the basis of the survey conducted by Khullar et al [35], we asked participants their level of concern (very concerned, somewhat concerned, not concerned, and don't know) related to 6 potential challenges of using AI for mental health care (Figure 1). Participants reported being somewhat or very concerned about AI making the wrong diagnosis (402/500, 80.4%), leading to inappropriate treatment (435/500, 87%), or leading to them not knowing their mental health care provider well (409/500, 81.8%). Participants reported being very or somewhat concerned regarding spending less time with their mental health care professional (346/500, 69.2%) and their confidentiality (302/500, 60.4%) but expressed relatively less concern regarding increased costs (217/500, 43.4%).

Figure 1. Reported levels of perceived concerns regarding artificial intelligence (AI) use for mental health.

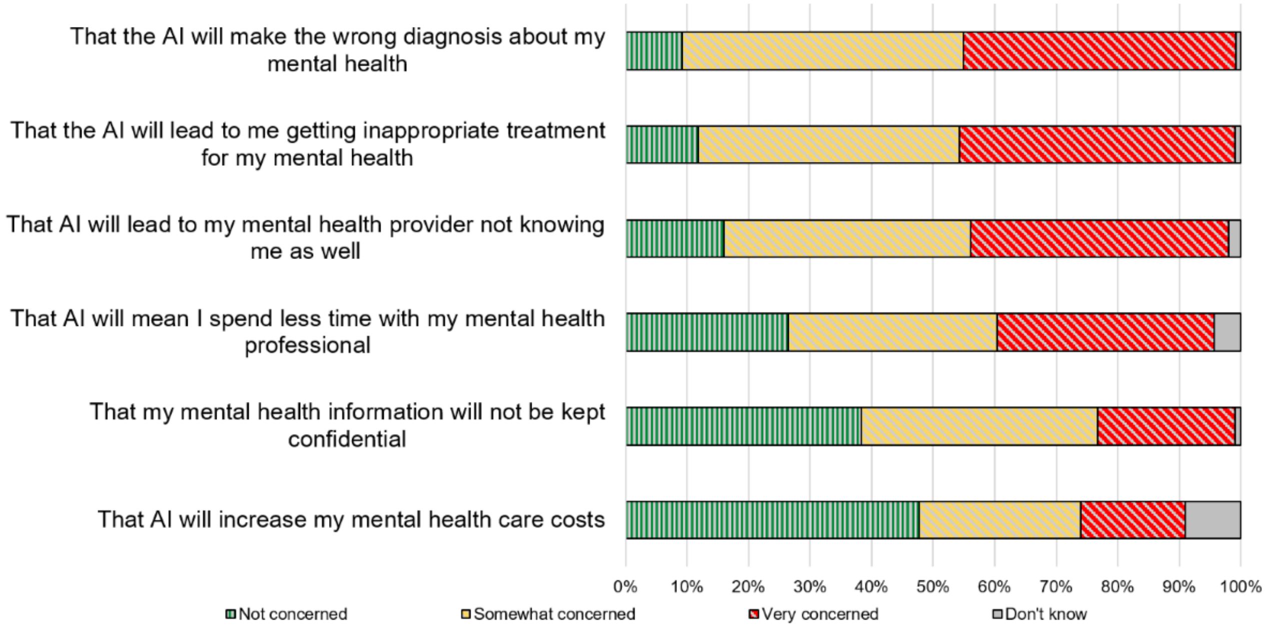


RQ 3: Comfort With AI Accomplishing Mental Health Care Tasks

Next, we asked patients their level of comfort with AI performing various tasks instead of their mental health care professional. We assessed a range of tasks (ie, assessment, diagnosis, diagnosis delivery, and treatment recommendation) and mental health issues of varied levels of perceived severity (ie, depression, bipolar disorder, and suicide), as shown in Figure 2. Participants were the most comfortable (reporting being very or somewhat comfortable) with recommendations of nonpharmacological interventions, including general wellness management strategies (357/500, 71.4%) and talk therapy (328/500, 65.6%). Participants expressed moderate comfort with AI performing a mental health care assessment but were less comfortable (with only 20% to 32% selecting very or somewhat comfortable) with various prediction, diagnosis, and diagnosis delivery tasks, as well as a medication recommendation task. Participants were the least comfortable with diagnosis delivery tasks (ie, telling someone directly they have a mental health condition), including for clinical depression (123/500, 24.6%) and bipolar disorder (103/500, 20.6%).

We also asked participants their level of comfort sharing mental health information with a (human) mental health care professional, an AI chatbot, or an AI program that treats disease to improve it. We chose these categories to understand perspectives of common uses of patient data for AI, such as the use of patient data to build models that make predictions or help treat diseases as compared to the use of patient information directly for patient support (eg, an AI chatbot). Patients were the most comfortable (very or somewhat) sharing information with a mental health care professional (389/500, 77.8%), followed by sharing with an AI program that treats disease (300/500, 60%) and then sharing with an AI chatbot (238/500, 47.6%).

Figure 2. Reported level of comfort with artificial intelligence (AI), instead of a mental health professional, conducting various tasks.



RQ 4: Values Related to AI for Mental Health

Overview

We replicated a series of questions from Khullar et al [35] assessing patient values of various aspects of AI, including transparency, explainability and performance, responsibility, and the effect of AI on trust in health professionals (Table 3).

Table 3. Summary of questions regarding values related to artificial intelligence (AI) for mental health (N=500).

Value, question, and answer choice	Respondents, n (%)
Transparency of AI use	
How important do you think it is that you are told when an AI program has played a big role in your mental health diagnosis or treatment?	
Not important	13 (2.6)
Somewhat important	111 (22.2)
Very important	364 (72.8)
Don't know	12 (2.4)
How important do you think it is that you are told when an AI program has played a small role in your mental health diagnosis or treatment?	
Not important	39 (7.8)
Somewhat important	128 (25.6)
Very important	253 (50.6)
Don't know	10 (2)
[Clinical scenario][a] How important is it that your doctor tells you that the computer program helped make this decision?	
Not important	25 (5)
Somewhat important	66 (13.2)
Very important	388 (77.6)
Don't know	21 (4.2)
Explainability and performance	
How comfortable would you be receiving a mental health diagnosis from a computer program that made the right diagnosis 90% of the time but could not explain why it made the diagnosis?	
Very comfortable	15 (3)
Somewhat comfortable	94 (18.8)
Somewhat uncomfortable	184 (36.8)
Very uncomfortable	199 (39.8)
Don't know	8 (1.6)
How comfortable would you be receiving a mental health diagnosis from a computer program that made the right diagnosis 98% of the time but could not explain why it made the diagnosis?	
Very comfortable	63 (12.6)
Somewhat comfortable	138 (27.6)
Somewhat uncomfortable	172 (34.4)
Very uncomfortable	117 (23.4)
Very comfortable	10 (2)
Don't know	63 (12.6)
Responsibility and AI	
Imagine that your mental health professional and a computer program work together to treat your mental illness and a medical error occurs. An example of a medical error is getting a diagnosis that was wrong, or a treatment that was not needed. Who is responsible? (Select all that apply.)[b]	
The mental health professional	412 (82.4)
The company that made the computer program	30 (6)
The hospital or clinic that bought the computer program	20 (4)
The government agency that approved the computer program	6 (1.2)
Someone else	12 (2.4)
No one	1 (0.2)

Value, question, and answer choice	Respondents, n (%)
Don’t know	19 (3.8)
Imagine that you have a sleeping disorder that might be due to a mental health issue. You have a test done. Your doctor uses a computer program that says the sleeping disorder might be mental health-related, so you start medication to treat it. The medication leads to bad side effects. After another doctor evaluates your sleeping disorder, it turns out it was NOT mental health related. Who, if anyone, is to blame? (Select all that apply.) [b]	
The mental health professional	408 (81.6)
The company that made the computer program	35 (7)
The hospital or clinic that bought the computer program	14 (2.8)
The government agency that approved the computer program	5 (1)
Someone else	6 (1.2)
No one	19 (3.8)
Don’t know	13 (2.6)
Imagine that your hospital recently started using a computer program to help diagnose mental health problems. Who do you think has checked to make sure the computer program is safe before it is rolled out? (Select all that apply.)[b]	
The mental health professional	135 (27)
The company that made the computer program	229 (45.8)
The hospital or clinic that bought the computer program	73 (14.6)
The government agency that approved the computer program	31 (6.2)
Someone else	4 (0.8)
No one	16 (3.2)
Don’t know	12 (2.4)
Effect of AI on trust in mental health professionals	
Imagine that you have some symptoms that have been bothering you for a while, such as difficulty sleeping, eating, and focusing on work. You visit a doctor who runs some tests and he says he does NOT think you have any mental health issue. He also puts your symptoms into a computer program that can make the right diagnosis about 80% of the time, but can’t say why it chose the diagnoses. It says you DO have mental health issue. How does the computer program affect your view?	
It would not affect my trust of the mental health professional’s assessment	85 (17)
It would make me question the mental health professional’s assessment	265 (53)
I do not know if it would change my view of the mental health professional’s assessment.	137 (27.4)
Don’t know	13 (2.6)

[a]Scenario wording: “Imagine that you have been told that you have been diagnosed with depression, a common mental illness that affects your mood, thoughts, and behavior. In the past, your doctor would have decided whether to prescribe a medication or refer you for psychotherapy depending on the type of symptoms you have and how severe they are. //Your doctor now has a computer program that uses many other factors. This computer program says you should start an antidepressant.”

[b]Multiple selections were allowed, so the sum of proportions can be >100%.

Transparency of AI Use

Participants were first asked how important it was to know when AI played a (1) small or (2) big role in their mental health treatment or diagnosis. Most participants found it somewhat or very important to know whether AI played a small (450/500, 90%) or big (474/500, 94.8%) role in their mental health treatment or diagnosis, although participants tended to report it was very important based on whether AI played a big (365/500, 73%) versus small (253/500, 50.6%) role. This pattern remained consistent with a specific scenario regarding the use of an AI program in prescribing antidepressants, with many participants stating it was very important (355/500, 71%) or somewhat important (114/500, 22.8%) that their mental health care professional informed them regarding the AI’s involvement in this prescribing decision.

Explainability and Performance

In comparing AI that was not explainable (ie, it could not describe why it made a given diagnosis), participants were generally uncomfortable even with stated AI performance accuracies of 90% (383/500, 76.6% somewhat or very uncomfortable) and 98% (289/500, 57.8% somewhat or very uncomfortable).

Responsibility and AI

Participants answered a series of questions regarding who was responsible in the event of a medical error when AI was used in conjunction with their mental health treatment; answer options included the following: the mental health professional who made the decision, the company that made the computer program, the hospital or clinic that bought the computer program, the government agency that approved the computer program, someone else, no one, and don't know. In the case where AI was used in collaboration with a single mental health professional, most participants (>80%) reported that the mental health professional would be the one responsible if a medical error (eg, wrong diagnosis or unnecessary treatment) occurred for both a specific (ie, sleep disorder) and a general scenario.

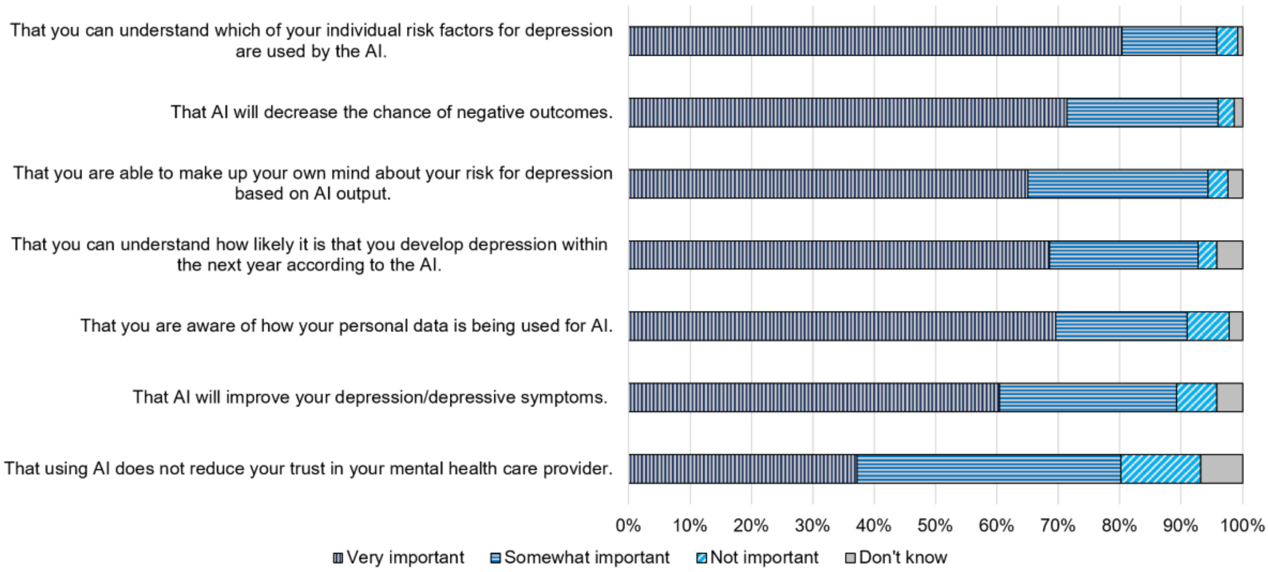
Participants were more divided on who had responsibility for ensuring an AI program for mental health care was safe, with the plurality stating the company that created the program was responsible (229/500, 45.8%), followed by the mental health professional (135/500, 27%) and then the health system (73/500, 14.6%), with fewer than 10% of the participants selecting each of the remaining answer options.

Effect of AI on Trust in Mental Health Care Professionals

Most participants (265/500, 53%) said that if an AI program that was accurate 80% of the time in detecting health issues related to sleeping, eating, and concentrating disagreed with their mental health professional, it would make them question the health professional's assessment. Notably, nearly 30% (137/500) of participants said they "did not know" how such information would change their view of their mental health professional's assessment, with the remaining 17% (85/500) stating it would not change.

To better understand what participants valued the most related to AI for their mental health, we asked the importance of various ethical constructs, based on MITRE's ethical framework for consumer-generated health information [43], as they pertained to an example AI program used to support treatment for depression (Figure 3). Over 80% (range 80.2%-96%) of participants found each of the constructs somewhat or very important. Notably, the highest proportion of participants (402/500, 80.3%) viewed explainability and transparency, "explainability," to be very important. Participants tended to perceive decreasing the risk of negative outcomes as slightly more important than improving symptoms. Participants found AI *not* reducing trust in their mental health professional as the least important trait by comparison, although 37.2% (186/500) and 43% (215/500) rated this trait as very and somewhat important, respectively.

Figure 3. Importance of various values related to artificial intelligence (AI) use for mental health.



Qualitative Results

Participants provided free-text responses describing themes related to nuanced aspects of AI's performance, human-AI dynamics, and further values or concerns pertaining to AI. Free-text responses were mandatory, but some participants simply stated they had no additional concerns (165/500, 33%), or they did not provide sufficient detail for their responses to be categorized (7/500, 1.4%). Of the 1000 responses (2 per participant), 97 involved >1 code, so percentages listed subsequently reflect the proportion of total codes observed. On the basis of the results listed in Table 1 (sociodemographic differences in the perceived benefits of AI for mental health care), quotes shown also provide the patients' gender, race, and self-rated literacy for context.

AI Performance

Table 4 provides the detailed codes, proportion of occurrence, and examples related to AI performance. Participants described issues related to AI's accuracy, biases in AI data or biases that may occur in the use of AI, potential errors AI may commit, and how these errors may affect the quality of care. Participants expressed mixed opinions regarding whether AI would improve or degrade the quality of mental health care.

Table 4. Detailed qualitative codes related to artificial intelligence (AI) performance.

Detailed code	Count (n=188, 17.9%), n (%)[a]	Example quotes
AI performance and error: accuracy	96 (51.1)	• "I think the main concern would be potential for getting an inaccurate diagnosis or the wrong treatment. It would definitely take time to trust the reliability. Programming errors, for example, could potentially lead to fatal outcomes for patients" (White woman with self-rated adequate health literacy).
Bias in data and use	50 (26.6)	• "I am concerned that the algorithms/data set that was utilized to train AI would be biased. For instance, if more white people seek mental health care, and AI is trained on their data, would AI be as good at diagnosing mental health conditions in people of color?" • "It will be biased, sexist and racist. It will rely on old ideas of mental health care and not use current information. It will be used to ignore or bully patients" (White nonbinary individual with self-reported inadequate health literacy).
Risk of harm	29 (15.4)	• "I think there is far too much on the line when it comes to mental health that it's risky to rely on AI for it. I think my main concern would be being over-diagnosed and having to be put in a psychiatric ward. I think that would be horrific" (White woman with self-rated adequate health literacy).
Care quality	13 (6.9)	• "I think that it might lessen the quality standard of hired healthcare professionals since the expertise of AI system could become more important" (Black man with self-reported adequate health literacy). • "It would improve the quality of work" (White woman with self-rated adequate health literacy).

[a]Percentages do not add to 100%, as "no additional concerns" and "indeterminable" codes are included in the count.

AI and Humans: Superior, Inferior, or Simply Better Together

Table 5 presents participant feedback related to human-AI dynamics. Participants described worry that AI may not be able to replicate things done by humans (AI capabilities, human reasoning and communication, and the importance of human connection). By contrast, some pointed out ways in which AI may offer advantages to human cognition (AI capabilities). They also provided feedback on how AI and humans may (or may not) work together (human-AI collaboration and overreliance on AI), with many noting that AI should be overseen by humans and not work autonomously in mental health care applications. Finally, a few participants expressed concerns regarding how AI may take away jobs from humans.

Table 5. Detailed qualitative codes related to human–artificial intelligence (AI) dynamics.

Detailed code	Count (n=297. 28.3%), n (%)[a]	Example quote
AI capabilities	78 (26.3)	• "A software program's understanding of mental health will never be as nuanced as a real humans. There will always be less common variables that AI systems aren't programmed to take into account. I am concerned that with AI driven healthcare, patient[s] with more unusual backgrounds, experiences and symptoms will not have access to human professional[s] who can more fully consider their circumstances" (Asian woman with self-rated adequate health literacy). • "AI is better at playing human games than humans are - it's already World Go Champion and World Chess Champion..." (Black man with self-reported adequate health literacy)
Human-AI collaboration	39 (13.1)	• "While I think that AI will be useful in mental health scenarios, I think that oversight should still be done, just like how I would prefer a few doctors to confirm a diagnosis. I think that it will be good at detecting some trends that can help move people towards better help, but that the help itself should be a joint effort and more personalized" (White nonbinary individual with self-reported adequate health literacy). • "Overall, the use of AI is to assist the doctor in providing an accurate diagnosis. It improves the reliability of the diagnosis" (Asian man with self-rated adequate health literacy).
Overreliance on AI	30 (10.1)	• "The biggest ethical concern I can think of is a mental health professional being completely reliant on AI without taking a closer look into how the program works" (White woman with self-rated adequate health literacy)
Human reasoning and communication	54 (18.2)	• "That it can't pick up on the subtleties of some symptoms that a human can" (White man with self-rated inadequate health literacy).
Importance of human connection	90 (30.3)	• "Human connection and understanding are crucial in mental health diagnosis and treatment. I finally found a doctor that made me feel understood, heard, and cared for. It resulted in an effective treatment for my major depression and suicidal ideology after many years. AI can't do that" (White woman with self-rated adequate health literacy).
Jobs	6 (2.0)	• "As someone who will be working in healthcare in the new future, I am concerned that AI in health scenarios will take away jobs from real people who put in all the work to be working there" (White woman with self-rated adequate health literacy).

[a]Percentages do not add to 100%, as "no additional concerns" and "indeterminable" codes are included in the count.

Additional Values and Concerns

Respondents expressed further values and concerns beyond performance and human-AI dynamics, many of which were also covered in the closed-ended survey responses, including trust, transparency, privacy, responsibility, and cost (Table 6).

Table 6. Detailed qualitative codes related to additional values and concerns regarding the use of artificial intelligence (AI) for mental health.

Detailed code	Count (n=222, 21.2%), n (%)[a]	Example quote
Privacy	98 (44.1)	• "I would be concerned about the company that owns the AI and if they could share the data with third parties. I'd also be concerned about what happens if someone admitted suicidal thoughts" (White woman with self-rated adequate health literacy).
Transparency	22 (9.9)	• "I would worry about how my data is used to make AI based decisions. I would also wonder about the type of data being used" (White man with self-rated adequate health literacy).
Ethics	23 (10.4)	• "One worry is that AI might be used to diagnose and treat mental health conditions without a person's consent. This could lead to people being treated for conditions they do not have, or not receiving treatment for conditions they do have..." (White man with self-rated inadequate health literacy).
Trust	26 (11.7)	• "I am just not comfortable with any machine diagnosing and treating any symptoms of mine, bottom line" (White man with self-rated adequate health literacy).
Appropriate use	19 (8.6)	• "Yes, that a manipulative enough person could sway the machine into getting what they want rather than what they need" (White man with self-rated adequate health literacy).
Responsibility	14 (6.3)	• "I wonder who would be held liable in the event a patient dies or experiences bad side effects due to the diagnosis or advice of AI. Would it be the AI itself or would the specialist also be held accountable?" (White woman with self-rated adequate health literacy)
Cost	20 (9.0)	• "I think if the purpose is to provide better and more thorough health care, then it is a good endeavor. If the purpose is to decrease the costs of providing health care while maximizing profits, the project is specious. That's why I can see it as a diagnostic tool to help healthcare providers come to a more accurate and thorough diagnosis. But I think it's about maximizing profits and all stages of the healthcare process" (Asian woman with self-rated adequate health literacy)

[a]Percentages do not add to 100%, as "no additional concerns" and "indeterminable" codes are included in the count.

Discussion

Principal Findings

This is one of the first studies to explore public perspectives of the use of AI for mental health–related applications. Our results expand upon other works studying public perceptions of AI for non–mental health care applications and raise important considerations regarding patient involvement in AI use for their mental health [31,44]. We also focused on various applications of AI to mental health care, differentiating our results from previous user-centered design studies that have elicited participant perceptions of a single, specific AI tool under development. Our study highlights the nuances of patients' perspectives regarding AI for mental health care, revealing that their comfort with AI use depends on the purpose of the AI (tasks it performs), use process (when it is used and what factors drive predictions), and performance of the AI (how well it works and what happens when it is wrong) [45].

Perceived Benefits of AI for Mental Health Care (RQ 1)

Just under half (245/497, 49.3%) of the respondents in our study reported that they thought AI would make mental health care better or somewhat better. This is similar to a study conducted in Germany where 53% of patients reported positive or very positive attitudes toward AI, but not specifically for mental health [46]. Participants in other studies asking perceptions regarding more specific applications of AI (eg, a radiology image interpretation study in Saudi Arabia and a pregnancy and postpartum exploration in Spain) had stronger positive attitudes toward AI [47,48]. This is consistent with qualitative studies that have found patients with specific health challenges more readily connected with AI's potential benefits. Our study also found that women were associated with lower perceived benefits of AI for mental health, while lower self-rated health literacy and Black or African American race were each associated with more positive perceptions. The previously cited study conducted in Germany had similar findings related to lower perceived benefit among women [46]. However, it is interesting that this result remained consistent for mental health care applications, in light of the fact that women have reported lower levels of stigma regarding mental health care than men [49,50]. Our study also detected an interesting paradox that those having lower self-reported health literacy had more positive perceptions toward AI for mental health, although this finding warrants further replication and investigation. Regardless of patient perceptions, the inclusivity of patient-facing information regarding AI, ensuring those of various levels of literacy and numeracy may equitably comprehend its functions, remains critically important. Those of African American and Black race in the United States consistently report greater stigma and lower levels of trust toward mental (and other) health institutions due to biases, discrimination, and systemic racism [51,52]. The greater perceived benefit of AI for mental health care may represent a view that AI can be more just and without the biases of humans. This notion also requires further exploration, particularly given that the biases of human can often be embedded into the AI because training data embody previous human behavior.

Concerns Regarding AI for Mental Health Care (RQ 2)

Participants in our study cited concerns consistent with previous work related to AI accuracy, risk of harm (eg, wrong diagnosis and inappropriate treatment) [44,53,54], decreased human communication and connection [35,44,48,53], and issues pertaining to confidentiality [35,44,48,54-57]. Issues related to privacy were also the most commonly mentioned concern in the qualitative feedback. Participants also qualitatively described concerns about the performance of AI and doubts in AI's ability to truly replicate human reasoning. In our study, participants expressed some concern related to rising costs, although, as found in other studies, this worry was less pronounced [35,58]. These results continue to stress the importance of contextualization for patients in terms of the following: the accuracy of AI, harms and how they are mitigated, and data use and protections. As described in previous studies, continuing to support human connection is particularly needed in mental health care applications given the importance of the patient–mental health care professional therapeutic relationship.

Comfort With AI Accomplishing Mental Health Care Tasks (RQ 3)

Our study provided further evidence that patients' comfort with AI varies based on what the AI does. People were the least comfortable with diagnosis delivery tasks, which lends further support to the importance of continuing to keep health professionals in the loop related to AI [35]. Similar to a previous study of pregnant people in Spain [47], patients were most comfortable with tasks that recommended general wellness strategies or talk therapy. These results also suggest that AI for tasks patients are less comfortable with may require greater care to explain and also emphasize how the AI works with the health professional.

Despite the rapid proliferation of chatbots [37,59], less than half of the participants (237/500, 47.4%) were comfortable sharing mental health information with a chatbot, which may simply signify that these types of tools should be used on an opt-in basis. Previous studies have suggested that it may be easier for someone to share these sensitive feelings with a computer or AI [60], and this may be true for certain people, but our findings did not universally support this assertion. It was also notable that approximately a quarter of the sample was not comfortable sharing mental health information with a mental health care professional. We acknowledge this may have been impacted by the types of mental health information listed in our survey, but it may also represent a continued stigma related to sharing mental health concerns.

Values Related to AI for Mental Health Care (RQ 4)

Findings in our study related to patients' values for AI for mental health care revealed challenges that AI integration may present to the patient–health professional relationship. In individual scenarios, patients overwhelmingly found mental health care professionals responsible for AI-related errors. While this does reflect similarity to the current standard of practice (ie, that health professionals are held responsible for medical errors even when computer systems are involved), future work should consider how this affects health professional well-being given the challenges related to burnout and shortages in trained mental health care workers. It also supports programs, such as AI programs falling under the US Food and Drug Administration's purview and the European Union's AI Act of 2023, where algorithms may be reviewed before use and subject to regulations [61]. Participants also noted that their trust in their mental health care professional would decrease if their assessment disagreed with AI. However, this was somewhat at odds with relatively fewer patients viewing issues with AI decreasing trust in their mental health care professional as "very important." It was also notable that approximately one-third (137/500, 27.4%) of participants said they "did not know" how this scenario would affect their trust in their mental health care professional, which seems to highlight that patients may still be wrapping their heads around feelings regarding emerging technologies, such as AI.

Patients desired a high degree of transparency related to AI use, with >90% of participants considering it important that they be told when AI played even a small role in their care. Participants also valued explainability, as most participants (289/500, 57.8%) were not comfortable with highly accurate AI that could not explain how it made its predictions and as "understanding individual risk factors" was rated as the most important value related to AI for mental health care applications. It is at best unclear what patients are typically told regarding when AI is used for their care; how explainable it is; and to what extent, if at all, they are informed what factors drive predictions regarding their care. These results suggest that patient values may be at odds with the current standard of practice for patient communication. Similar to the concerns previously described, participants also highly rated the importance of AI not leading to errors and helping with their mental health symptoms.

Implications and Future Challenges

Implications regarding our findings are organized topically and include concrete recommendations, which have been italicized for emphasis.

Navigating the Patient–Health Professional Relationship

The therapeutic relationship between a patient and their health professional is crucial in mental health care settings. Our study revealed issues that will need to be reconciled if AI is to be safely, transparently, and acceptably used for mental health care. From our results, it is clear that patients want mental health care professionals to be the ultimate decision makers, using AI to support (but not make) decisions when it is deemed safe and effective [36]. Participants also overwhelmingly viewed mental health care professionals as the people responsible if an error occurred related to treatment where AI had been used.

Future work should investigate shared regulations for AI responsibility, health professional competencies for AI use (Russell et al [62]), and interfaces that support shared decision-making when using AI.

Specifically, designs should support collaborative patient–health professional decision-making in a way that fosters trust instead of degrading it while also not creating undue burden for the health professional. Previous studies have described how clinicians should be able to contest AI, such as ignoring it (when it is not relevant or appropriate), trusting it when it is appropriate, or being able to uncover explanations to negotiate in borderline cases [63]. Such systems should be able to track health professional decisions in relation to the AI, possibly allowing health professionals to provide brief rationale that they may use in conversations with patients. In creating such systems, usability and model explainability will be critical. While these systems may be difficult to study in situ given the sensitivity of mental health conversations, *solutions may first be evaluated in realistic clinical simulation environments to ensure safety and usability prior to larger scale deployment.*

Communicating AI-Related Information to Patients

Participants in this study desired various information regarding the use of AI for their mental health care; that is, they wished to know when, for what, and why it was used; how accurate it was; and the risk factors that drove a decision. Even with highly predictive AI, patients were still not comfortable with AI that could not explain how it arrived at a result, and they qualitatively expressed concerns related to misdiagnosis and improper treatment that could result from AI use. They also reported "understanding individual risk factors" as the most important value. It is unclear as to how much if any "information" patient receive, and there seems to be a mismatch between the current deployment of AI for mental health care and patients' desires. *At minimum, we recommend promoting transparency in AI's use, the communication of its accuracy, and including individual risk factors to help patients and clinicians decide when AI may be appropriate for use in health-related applications. Communications should also address how potential biases (eg, in the training data) have been evaluated and mitigated in the resultant* AI *tool.*

Communicating the desired information to patients, however, is not straightforward, as concepts such as AI performance and process involve complex mathematical concepts. Furthermore, this desire for additional information regarding AI is also at odds with how patient communication has traditionally been practiced. When we consider other non–AI-based diagnostic or decision support tools (eg, magnetic resonance imaging, blood tests, and screening assessments), communicating information regarding how they work (ie, their process) or their performance is far from standard practice. AI seems to be held to a higher standard than other diagnostic tools related to transparency in performance. *Future work should consider not only communicating the process and performance of AI but also providing this information in the context with the performance of the existing approach to a given task.* This would allow health professionals and patients to determine whether the potential benefits of AI outweigh their concerns.

Providing patients with explanations of AI performance and factors driving prediction will require extensive study involving experts in human-centered, inclusive design working alongside AI developers, mental health care professionals, and patients. *There is a need for laboratory-based studies to understand what information regarding AI balances recognizing patient values, but also supports comprehension of important concepts and fostering appropriate trust.* These issues raise many questions for numeracy and data visualization experts regarding how information may be conveyed inclusively to patients with different needs so that the benefits of AI may be equitably realized.

Individuality and Autonomy

Our study detected differences in who may find AI beneficial and for what tasks they may be comfortable using it. *Future work should explore how we may respect individual autonomy with regard to the use of AI applications so that patients and health professionals may collaboratively make decisions about the appropriate application of AI to mental health care issues.*

Limitations

Our study was limited in that the sample was recruited using a web-based platform, which may not generalize to those with technology, literacy, and other barriers to web-based survey completion. The sample was representative of the adult US population in terms of age, gender, and race distributions (Prolific/academic researchers), leading to most of the survey respondents being White due to >70% due to the US demographic makeup (US Census Bureau [64]). Therefore, important perspectives from other racial and ethnic groups may be limited. Our results may also be subject to response bias, as those who had strong feelings regarding topics pertaining to AI, mental health, or their intersection may have been more likely to respond. For example, a higher-than-anticipated proportion of respondents (215/500, 43%) reported having a history of mental illness. We did attempt to control for this in RQ 1, and the history of mental illness was not found to significantly affect responses regarding participants' perceived benefits of AI for mental health care. We, however, did not have a mechanism for evaluating how preexisting attitudes of AI may have affected responses, and this may be addressed with validated measures of AI attitudes in future surveys (Schepman and Rodway [65]). Given the scope of this paper, we were unable to assess sociodemographic differences (eg, based on race and literacy) across all outcomes. Future work may use this or other data sets to provide a more nuanced picture of differing views related to specific questions regarding AI for mental health.

Conclusions

Our study found that approximately half (245/497, 49.3%) of the US adults surveyed perceived some benefit for the use of AI in mental health care applications. These perceived benefits were lower among women but higher among Black or African American participants and those with lower self-rated health literacy. Participants also expressed nuanced differences in the types of tasks they would be comfortable with AI completing, showing the greatest discomfort with AI handling clinical diagnosis, diagnosis delivery, and the recommendation of medication. Those surveyed valued high-performing AI that could explain individual risk factors driving predictions. In general, participants were concerned that AI may mean a loss of human connection, and they perceived humans as the ultimate decision makers, with AI serving as an additional data point when appropriate. Qualitative feedback also revealed participants' deep-seated fears regarding the use of AI for their mental health care. These findings stress the importance of working with patients and mental health care professionals to understand whether and how AI may be safely, ethically, and acceptably implemented for mental health care applications.

Acknowledgments

This work was supported by the National Institute for Mental Health (R41MH124581-S1; principal investigator: JP)

Data Availability

Data are available upon reasonable request to the corresponding author.

Authors' Contributions

NB and MRT jointly conceptualized the study design and questionnaire materials with input from all other authors. JK, PMD, and ZR completed the data analysis and were advised by NB and MRT. NB drafted the manuscript with substantive input from all other authors.

Conflicts of Interest

AH is a consultant at Progyny, Inc, co-founder and medical advisor at Iris Ob Health, Inc, and a writer and contributor at Uptodate.com. MRT is the co-founder of Iris OB Health and a consultant at Boston Scientific. All other authors declare no conflicts of interest.

Multimedia Appendix 1
Full survey battery, including exact questions and answers.
[DOCX File , 26 KB - mental_v11i1e58462_app1.docx]

References

1. Preparing for the future of artificial intelligence. Executive Office of the President, National Science and Technology Council Committee on Technology. 2016 Oct. URL: https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf [accessed 2023-01-03]
2. Topol E. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. Hachette, UK: Basic Books; 2019.
3. Triantafyllidis AK, Tsanas A. Applications of machine learning in real-life digital health interventions: review of the literature. J Med Internet Res 2019 Apr 05;21(4):e12286 [FREE Full text] [doi: 10.2196/12286] [Medline: 30950797]
4. Ting DS, Liu Y, Burlina P, Xu X, Bressler NM, Wong TY. AI for medical imaging goes deep. Nat Med 2018 May;24(5):539-540. [doi: 10.1038/s41591-018-0029-3] [Medline: 29736024]
5. TrueFidelity CT\textbar GE healthcare. GE Healthcare. 2022. URL: https://www.gehealthcare.com/products/truefidelity [accessed 2023-01-03]
6. Caption AI. Caption Health. 2022. URL: https://captionhealth.com/technology [accessed 2023-01-03]
7. Home page. Viz. ai. 2022. URL: https://www.viz.ai/ [accessed 2023-01-03]
8. Home page. Aidoc. 2022. URL: https://www.aidoc.com/ [accessed 2023-01-03]
9. Enterprise imaging solutions. IBM Watson Health. 2022. URL: https://www.ibm.com/watson-health/solutions/enterprise-imaging [accessed 2023-01-03]
10. AI-rad companion. Siemens. 2022. URL: https://www.siemens-healthineers.com/en-us/digital-health-solutions/digital-solutions-overview/clinical-decision-support/ai-rad-companion [accessed 2023-01-03]
11. Seal A, Bajpai R, Agnihotri J, Yazidi A, Herrera-Viedma E, Krejcar O. DeprNet: a deep convolution neural network framework for detecting depression using EEG. IEEE Trans Instrum Meas 2021;70:1-13. [doi: 10.1109/tim.2021.3053999]

12. Zhang Y, Wang S, Hermann A, Joly R, Pathak J. Development and validation of a machine learning algorithm for predicting the risk of postpartum depression among pregnant women. J Affect Disord 2021 Jan 15;279:1-8 [FREE Full text] [doi: 10.1016/j.jad.2020.09.113] [Medline: 33035748]
13. Flesia L, Monaro M, Mazza C, Fietta V, Colicino E, Segatto B, et al. Predicting perceived stress related to the COVID-19 outbreak through stable psychological traits and machine learning models. J Clin Med 2020 Oct 19;9(10):3350 [FREE Full text] [doi: 10.3390/jcm9103350] [Medline: 33086558]
14. Sundaram L, Bhat RR, Viswanath V, Li X. DeepBipolar: identifying genomic mutations for bipolar disorder via deep learning. Hum Mutat 2017 Sep;38(9):1217-1224 [FREE Full text] [doi: 10.1002/humu.23272] [Medline: 28600868]
15. Sheikh M, Qassem M, Kyriacou PA. Wearable, environmental, and smartphone-based passive sensing for mental health monitoring. Front Digit Health 2021;3:662811 [FREE Full text] [doi: 10.3389/fdgth.2021.662811] [Medline: 34713137]
16. Reale C, Novak LL, Robinson K, Simpson CL, Ribeiro JD, Franklin JC, et al. User-centered design of a machine learning intervention for suicide risk prediction in a military setting. AMIA Annu Symp Proc 2020;2020:1050-1058 [FREE Full text] [Medline: 33936481]
17. Adam G, Rampášek L, Safikhani Z, Smirnov P, Haibe-Kains B, Goldenberg A. Machine learning approaches to drug response prediction: challenges and recent progress. NPJ Precis Oncol 2020;4:19 [FREE Full text] [doi: 10.1038/s41698-020-0122-1] [Medline: 32566759]
18. Lin E, Lin CH, Lane HY. Precision psychiatry applications with pharmacogenomics: artificial intelligence and machine learning approaches. Int J Mol Sci 2020 Feb 01;21(3):969 [FREE Full text] [doi: 10.3390/ijms21030969] [Medline: 32024055]
19. Girosi. Emerging technologies and therapeutics reports: narrative review and evidence mapping of artificial intelligence in clinical care. RAND Corporation. 2021. URL: https://www.pcori.org/research-results/2020/emerging-technologies-and-therapeutics-reports-narrative-review-and-evidence-mapping-artificial-intelligence-clinical-care [accessed 2024-04-29]
20. Zhou Q, Chen ZH, Cao YH, Peng S. Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. NPJ Digit Med 2021 Oct 28;4(1):154 [FREE Full text] [doi: 10.1038/s41746-021-00524-2] [Medline: 34711955]
21. Plana D, Shung DL, Grimshaw AA, Saraf A, Sung JJ, Kann BH. Randomized clinical trials of machine learning interventions in health care: a systematic review. JAMA Netw Open 2022 Sep 01;5(9):e2233946 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.33946] [Medline: 36173632]
22. Walker SC, French B, Moore RP, Domenico HJ, Wanderer JP, Mixon AS, et al. Model-guided decision-making for thromboprophylaxis and hospital-acquired thromboembolic events among hospitalized children and adolescents: the CLOT randomized clinical trial. JAMA Netw Open 2023 Oct 02;6(10):e2337789 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.37789] [Medline: 37831448]
23. Lee TC, Shah NU, Haack A, Baxter SL. Clinical implementation of predictive models embedded within electronic health record systems: a systematic review. Informatics (MDPI) 2020 Sep;7(3):25 [FREE Full text] [doi: 10.3390/informatics7030025] [Medline: 33274178]
24. Benda NC, Das LT, Abramson EL, Blackburn K, Thoman A, Kaushal R, et al. "How did you get to this number?" Stakeholder needs for implementing predictive analytics: a pre-implementation qualitative study. J Am Med Inform Assoc 2020 May 01;27(5):709-716 [FREE Full text] [doi: 10.1093/jamia/ocaa021] [Medline: 32159774]
25. Matthiesen S, Diederichsen SZ, Hansen MK, Villumsen C, Lassen MC, Jacobsen PK, et al. Clinician preimplementation perspectives of a decision-support tool for the prediction of cardiac arrhythmia based on machine learning: near-live feasibility and qualitative study. JMIR Hum Factors 2021 Nov 26;8(4):e26964 [FREE Full text] [doi: 10.2196/26964] [Medline: 34842528]
26. Baxter SL, Bass JS, Sitapati AM. Barriers to implementing an artificial intelligence model for unplanned readmissions. ACI open 2020 Jul;4(2):e108-e113 [FREE Full text] [doi: 10.1055/s-0040-1716748] [Medline: 33274314]
27. Nong P, Adler-Milstein J. Socially situated risk: challenges and strategies for implementing algorithmic risk scoring for care management. JAMIA Open 2021 Jul;4(3):ooab076 [FREE Full text] [doi: 10.1093/jamiaopen/ooab076] [Medline: 34522847]
28. Salwei ME, Anders S, Novak L, Reale C, Slagle J, Harris J, et al. Preventing clinical deterioration in cancer outpatients: human centered design of a predictive model and response system. J Clin Orthod 2022 Jun 01;40(16_suppl):e13567. [doi: 10.1200/jco.2022.40.16_suppl.e13567]
29. Abbasgholizadeh Rahimi S, Cwintal M, Huang Y, Ghadiri P, Grad R, Poenaru D, et al. Application of artificial intelligence in shared decision making: scoping review. JMIR Med Inform 2022 Aug 09;10(8):e36199 [FREE Full text] [doi: 10.2196/36199] [Medline: 35943793]
30. Walsh CG, McKillop MM, Lee P, Harris JW, Simpson C, Novak LL. Risky business: a scoping review for communicating results of predictive models between providers and patients. JAMIA Open 2021 Oct;4(4):ooab092 [FREE Full text] [doi: 10.1093/jamiaopen/ooab092] [Medline: 34805776]

31. Khullar D, Casalino LP, Qian Y, Lu Y, Chang E, Aneja S. Public vs physician views of liability for artificial intelligence in health care. J Am Med Inform Assoc 2021 Jul 14;28(7):1574-1577 [FREE Full text] [doi: 10.1093/jamia/ocab055] [Medline: 33871009]
32. Public law 114--255. 114 th Congress. Senate and House of Representatives of the United States of America. 2021. URL: https://www.govinfo.gov/content/pkg/PLAW-114publ255/pdf/PLAW-114publ255.pdf [accessed 2024-04-29]
33. Dai T, Tayur S. Designing AI ‐ augmented healthcare delivery systems for physician buy ‐ in and patient acceptance. Prod Oper Manag 2022 Dec 01;31(12):4443-4451. [doi: 10.1111/poms.13850]
34. Ghassemi M, Mohamed S. Machine learning and health need better values. NPJ Digit Med 2022 Apr 22;5(1):51 [FREE Full text] [doi: 10.1038/s41746-022-00595-9] [Medline: 35459793]
35. Khullar D, Casalino LP, Qian Y, Lu Y, Krumholz HM, Aneja S. Perspectives of patients about artificial intelligence in health care. JAMA Netw Open 2022 May 02;5(5):e2210309 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.10309] [Medline: 35507346]
36. Yarborough BJ, Stumbo SP. Patient perspectives on acceptability of, and implementation preferences for, use of electronic health records and machine learning to identify suicide risk. Gen Hosp Psychiatry 2021;70:31-37 [FREE Full text] [doi: 10.1016/j.genhosppsych.2021.02.008] [Medline: 33711562]
37. Shan Y, Ji M, Xie W, Lam KY, Chow CY. Public trust in artificial intelligence applications in mental health care: topic modeling analysis. JMIR Hum Factors 2022 Dec 02;9(4):e38799 [FREE Full text] [doi: 10.2196/38799] [Medline: 36459412]
38. Frankel J. How artificial intelligence could help diagnose mental disorders. The Atlantic. 2016 Aug 23. URL: https://www.theatlantic.com/health/archive/2016/08/could-artificial-intelligence-improve-psychiatry/496964/ [accessed 2023-09-12]
39. Hudson KL, Collins FS. The 21st century cures act - a view from the NIH. N Engl J Med 2017 Jan 12;376(2):111-113 [FREE Full text] [doi: 10.1056/NEJMp1615745] [Medline: 27959585]
40. Chew LD, Bradley KA, Boyko EJ. Brief questions to identify patients with inadequate health literacy. Fam Med 2004 Sep;36(8):588-594. [Medline: 15343421]
41. Reading Turchioe M, Harkins S, Desai P, Kumar S, Kim J, Hermann A, et al. Women's perspectives on the use of artificial intelligence (AI)-based technologies in mental healthcare. JAMIA Open 2023 Oct;6(3):ooad048 [FREE Full text] [doi: 10.1093/jamiaopen/ooad048] [Medline: 37425486]
42. Academic researchers. Prolific. 2022. URL: https://www.prolific.com/academic-researchers [accessed 2023-12-08]
43. Granger E, Skopac J. An ethical framework for the use of consumer generated data in health care. In: Proceedings of the 2021 Health Datapalooza and National Health Policy Conference. 2021 Presented at: HDNHPC '21; February 16-18, 2021; Virtual Event URL: https://academyhealth.confex.com/academyhealth/2021hdpnhpc/meetingapp.cgi/Paper/44315 [doi: 10.1377/forefront.20111121.015214]
44. Esmaeilzadeh P, Mirzaei T, Dharanikota S. Patients' perceptions toward human-artificial intelligence interaction in health care: experimental study. J Med Internet Res 2021 Nov 25;23(11):e25856 [FREE Full text] [doi: 10.2196/25856] [Medline: 34842535]
45. Benda NC, Novak LL, Reale C, Ancker JS. Trust in AI: why we should be designing for APPROPRIATE reliance. J Am Med Inform Assoc 2021 Dec 28;29(1):207-212 [FREE Full text] [doi: 10.1093/jamia/ocab238] [Medline: 34725693]
46. Fritsch SJ, Blankenheim A, Wahl A, Hetfeld P, Maassen O, Deffge S, et al. Attitudes and perception of artificial intelligence in healthcare: a cross-sectional survey among patients. Digit Health 2022;8:20552076221116772 [FREE Full text] [doi: 10.1177/20552076221116772] [Medline: 35983102]
47. Oprescu AM, Miró-Amarante G, García-Díaz L, Rey VE, Chimenea-Toscano A, Martínez-Martínez R, et al. Towards a data collection methodology for responsible artificial intelligence in health: a prospective and qualitative study in pregnancy. Inf Fusion 2022 Jul;83-84:53-78. [doi: 10.1016/j.inffus.2022.03.011]
48. Bahakeem BH, Alobaidi SF, Alzahrani AS, Alhasawi R, Alzahrani A, Alqahtani W, et al. The general population's perspectives on implementation of artificial intelligence in radiology in the western region of Saudi Arabia. Cureus 2023 Apr;15(4):e37391 [FREE Full text] [doi: 10.7759/cureus.37391] [Medline: 37182053]
49. Corrigan PW, Watson AC. The stigma of psychiatric disorders and the gender, ethnicity, and education of the perceiver. Community Ment Health J 2007 Oct;43(5):439-458. [doi: 10.1007/s10597-007-9084-9] [Medline: 17876705]
50. Chatmon BN. Males and mental health stigma. Am J Mens Health 2020;14(4):1557988320949322 [FREE Full text] [doi: 10.1177/1557988320949322] [Medline: 32812501]
51. Novacek DM, Hampton-Anderson JN, Ebor MT, Loeb TB, Wyatt GE. Mental health ramifications of the COVID-19 pandemic for Black Americans: clinical and research recommendations. Psychol Trauma 2020 Jul;12(5):449-451 [FREE Full text] [doi: 10.1037/tra0000796] [Medline: 32525370]
52. Eylem O, de Wit L, van Straten A, Steubl L, Melissourgaki Z, Danışman GT, et al. Stigma for common mental disorders in racial minorities and majorities a systematic review and meta-analysis. BMC Public Health 2020 Jun 08;20(1):879 [FREE Full text] [doi: 10.1186/s12889-020-08964-3] [Medline: 32513215]
53. Wu C, Xu H, Bai D, Chen X, Gao J, Jiang X. Public perceptions on the application of artificial intelligence in healthcare: a qualitative meta-synthesis. BMJ Open 2023 Jan 04;13(1):e066322 [FREE Full text] [doi: 10.1136/bmjopen-2022-066322] [Medline: 36599634]

54. Richardson JP, Curtis S, Smith C, Pacyna J, Zhu X, Barry B, et al. A framework for examining patient attitudes regarding applications of artificial intelligence in healthcare. Digit Health 2022;8:20552076221089084 [FREE Full text] [doi: 10.1177/20552076221089084] [Medline: 35355806]
55. Nelson CA, Pérez-Chada LM, Creadore A, Li SJ, Lo K, Manjaly P, et al. Patient perspectives on the use of artificial intelligence for skin cancer screening: a qualitative study. JAMA Dermatol 2020 May 01;156(5):501-512 [FREE Full text] [doi: 10.1001/jamadermatol.2019.5014] [Medline: 32159733]
56. Paprica PA, de Melo MN, Schull MJ. Social licence and the general public's attitudes toward research based on linked administrative health data: a qualitative study. CMAJ Open 2019;7(1):E40-E46 [FREE Full text] [doi: 10.9778/cmajo.20180099] [Medline: 30718354]
57. Thenral M, Annamalai A. Challenges of building, deploying, and using AI-enabled telepsychiatry platforms for clinical practice among urban Indians: a qualitative study. Indian J Psychol Med 2021 Jul;43(4):336-342 [FREE Full text] [doi: 10.1177/0253717620973414] [Medline: 34385728]
58. Harnkham N. Artificial intelligence in medicine: measuring respondents' trust in artificial intelligence for better patient care. William Howard Taft University. 2023. URL: https://scholar.google.com/citations?view_op=view_citation&hl=en&user=vGOKlycAAAAJ&citation_for_view=vGOKlycAAAAJ:u5HHmVD_uO8C [accessed 2024-04-29]
59. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and opinions of patients about mental health chatbots: scoping review. J Med Internet Res 2021 Jan 13;23(1):e17828 [FREE Full text] [doi: 10.2196/17828] [Medline: 33439133]
60. Miner A, Chow A, Adler S, Zaitsev I, Tero P, Darcy A, et al. Conversational agents and mental health: theory-informed assessment of language and affect. In: Proceedings of the 4th International Conference on Human Agent Interaction. 2016 Presented at: HAI '16; October 4-7, 2016; Biopolis, Singapore p. 123-130 URL: https://dl.acm.org/doi/10.1145/2974804.2974820 [doi: 10.1145/2974804.2974820]
61. EU AI Act: first regulation on artificial intelligence. European Parliament. URL: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence [accessed 2023-09-12]
62. Russell RG, Lovett Novak L, Patel M, Garvey KV, Craig KJ, Jackson GP, et al. Competencies for the use of artificial intelligence-based tools by health care professionals. Acad Med 2023 Mar 01;98(3):348-356. [doi: 10.1097/ACM.0000000000004963] [Medline: 36731054]
63. Sivaraman V, Bukowski L, Levin J, Kahn J, Perer A. Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing System. 2023 Presented at: CHI '18; April 23-28, 2023; Hamburg, Germany p. 1-18 URL: https://dl.acm.org/doi/full/10.1145/3544548.3581075 [doi: 10.1145/3544548.3581075]
64. 2020 census illuminates racial and ethnic composition of the country. United States Census Bureau. URL: https://www.census.gov/library/stories/2021/08/improved-race-ethnicity-measures-reveal-united-states-population-much-more-multiracial.html [accessed 2024-04-29]
65. Schepman A, Rodway P. The General Attitudes towards Artificial Intelligence Scale (GAAIS): confirmatory validation and associations with personality, corporate distrust, and general trust. Int J Hum Comput Interact 2022 Jun 14;39(13):2724-2741. [doi: 10.1080/10447318.2022.2085400]

Abbreviations

AI: artificial intelligence
OR: odds ratio
RQ: research question

Edited by J Torous; submitted 15.03.24; peer-reviewed by T Khinvasara, S Emani; comments to author 25.04.24; revised version received 26.06.24; accepted 14.07.24; published 18.09.24.

Please cite as:
Benda N, Desai P, Reza Z, Zheng A, Kumar S, Harkins S, Hermann A, Zhang Y, Joly R, Kim J, Pathak J, Reading Turchioe M
Patient Perspectives on AI for Mental Health Care: Cross-Sectional Survey Study
JMIR Ment Health 2024;11:e58462
URL: https://mental.jmir.org/2024/1/e58462
doi:10.2196/58462
PMID:

Original Paper

Digital Phenotypes for Early Detection of Internet Gaming Disorder in Adolescent Students: Explorative Data-Driven Study

Kwangsu Cho[1], PhD; Minah Kim[2,3], MD, PhD; Youngeun Cho[4], PhD; Ji-Won Hur[5], PhD; Do Hyung Kim[6], BSc; Seonghyeon Park[1], MSc; Sunghyun Park[2], MD; Moonyoung Jang[2,3], MS, MD; Chang-Gun Lee[6], PhD; Jun Soo Kwon[2,3], MD, PhD

[1]3R Innovation Research Center, Seoul, Republic of Korea
[2]Department of Neuropsychiatry, Seoul National University Hospital, Seoul, Republic of Korea
[3]Department of Psychiatry, Seoul National University College of Medicine, Seoul, Republic of Korea
[4]Department of Artificial Intelligence, Hanyang University, Ansan, Republic of Korea
[5]School of Psychology, Korea University, Seoul, Republic of Korea
[6]Department of Computer Science and Engineering, Seoul National University, Seoul, Republic of Korea

Corresponding Author:
Jun Soo Kwon, MD, PhD
Department of Neuropsychiatry
Seoul National University Hospital
101 Dahak-no
Jongno-gu
Seoul, 03080
Republic of Korea
Phone: 82 2 2072 2972
Email: kwonjs@snu.ac.kr

Related Article:

This is a corrected version. See correction statement: https://mental.jmir.org/2024/1/e60568

Abstract

Background: Limited awareness, social stigma, and access to mental health professionals hinder early detection and intervention of internet gaming disorder (IGD), which has emerged as a significant concern among young individuals. Prevalence estimates vary between 0.7% and 15.6%, and its recognition in the *International Classification of Diseases, 11th Revision* and *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition* underscores its impact on academic functioning, social isolation, and mental health challenges.

Objective: This study aimed to uncover digital phenotypes for the early detection of IGD among adolescents in learning settings. By leveraging sensor data collected from student tablets, the overarching objective is to incorporate these digital indicators into daily school activities to establish these markers as a mental health screening tool, facilitating the early identification and intervention for IGD cases.

Methods: A total of 168 voluntary participants were engaged, consisting of 85 students with IGD and 83 students without IGD. There were 53% (89/168) female and 47% (79/168) male individuals, all within the age range of 13-14 years. The individual students learned their Korean literature and mathematics lessons on their personal tablets, with sensor data being automatically collected. Multiple regression with bootstrapping and multivariate ANOVA were used, prioritizing interpretability over predictability, for cross-validation purposes.

Results: A negative correlation between IGD Scale (IGDS) scores and learning outcomes emerged (r_{166}=–0.15; *P*=.047), suggesting that higher IGDS scores were associated with lower learning outcomes. Multiple regression identified 5 key indicators linked to IGD, explaining 23% of the IGDS score variance: stroke acceleration (β=.33; *P*<.001), time interval between keys (β=–0.26; *P*=.01), word spacing (β=–0.25; *P*<.001), deletion (β=–0.24; *P*<.001), and horizontal length of strokes (β=0.21; *P*=.02). Multivariate ANOVA cross-validated these findings, revealing significant differences in digital phenotypes between potential IGD and non-IGD groups. The average effect size, measured by Cohen *d*, across the indicators was 0.40, indicating a moderate effect. Notable distinctions included faster stroke acceleration (Cohen *d*=0.68; *P*=<.001), reduced word spacing (Cohen *d*=.57; *P*=<.001), decreased deletion behavior (Cohen *d*=0.33; *P*=.04), and longer horizontal strokes (Cohen *d*=0.34; *P*=.03) in students with potential IGD compared to their counterparts without IGD.

Conclusions: The aggregated findings show a negative correlation between IGD and learning performance, highlighting the effectiveness of digital markers in detecting IGD. This underscores the importance of digital phenotyping in advancing mental health care within educational settings. As schools adopt a 1-device-per-student framework, digital phenotyping emerges as a promising early detection method for IGD. This shift could transform clinical approaches from reactive to proactive measures.

(JMIR Ment Health 2024;11:e50259) doi:10.2196/50259

KEYWORDS

adolescents; digital biomarkers; digital phenotyping; digital psychiatry; early detection; IGD; internet gaming disorder; pediatric psychiatry; proactive medicine; secondary school; universal screening

Introduction

Overview

Internet gaming disorder (IGD) is characterized by persistent and repetitive engagement in online gaming. Studies consistently show that individuals with IGD experience dysfunctions in academic functioning, executive functioning, and inhibitory control, which are crucial aspects for adolescent students [1-5]. Moreover, adolescents with IGD tend to have impaired self-esteem, interpersonal relationships, and daily functioning [6]. Consequently, IGD is included in Section III of the *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition, Text Revision* (*DSM-5-TR*) [7], pending formal inclusion as a disorder after further investigation. In 2019, the World Health Organization recognized IGD as a mental disorder [8].

IGD usually emerges in early adolescence [9] and significantly impacts adolescents worldwide, particularly those in secondary schools [10]. Its prevalence estimates range between 0.7% and 15.6%, with South Korea reporting around 5.9% of affected adolescents [11]. In response, the French government enacted legislation in July 2018 prohibiting the use of smartphones in schools for students from kindergarten to ninth grade. Additionally, approximately 77% of US public schools prohibit cell phone use in classrooms [12], and Korean schools either collect students' smartphones or require them to turn off their devices when entering classrooms [13].

Therefore, early detection and intervention in at-risk young individuals affected by IGD are crucial due to these associations because early identification makes it possible to provide appropriate care for adolescents [14-17]. However, many cases are undiagnosed and untreated due to several factors, including limited awareness of IGD as a mental disorder, social stigma, and limited access to mental health professionals [18].

Digital phenotyping has not currently been developed for IGD cases, although it is promising for evaluating, diagnosing, predicting, and monitoring mental illnesses [19]. This innovative approach involves the collection and analysis of data from various digital sources, such as smartphones, wearables, and tablets, to assess an individual's mental health and behaviors. It encompasses passive sensor data collection, including keystrokes, screen interactions, GPS location, social media use, voice tone, and physical activity. Some studies have even captured emotional states through keyboard and touch-stroking activities [20], with associations between keyboarding and various types of emotions identified [21,22]. Moreover, passive data, such as geolocation and accelerometer information, have proven successful in classifying negative symptoms in psychotic disorders [20,23]. Leveraging these sensor data, a handful of studies suggest the potential of digital phenotyping for screening IGD.

Digital phenotyping could play a pivotal role in complementing present clinical approaches by offering a proactive and continuous method of monitoring mental health for an early IGD detection role in several ways.

Unlike symptom-driven practices that rely on noticeable manifestations for patients or caregivers, digital phenotyping can detect early indicators that may not be immediately apparent in traditional assessments. By proactively identifying potential IGD concerns before they become overtly symptomatic, it supports a proactive approach to mental health care. This shift from reactive to proactive strategies aligns with the goal of early intervention and prevention of IGD. Additionally, in contrast to traditional methods, which rely on one-on-one interviews or 1-time questionnaires that may lack reliability due to social biases and limited truthful responses, digital phenotyping allows for the continuous, objective, and nonintrusive collection and analysis of adolescents' mental health patterns. This approach is free from temporal or spatial constraints and offers near–real-time, unbiased data collection [8-10]. As emphasized by Torous et al [24], leveraging adolescents' device engagement may enhance its acceptability and accessibility for monitoring IGD and other mental health conditions, proving cost-effective and efficient compared to traditional clinic visits. Moreover, identifying potential signs of IGD through personal devices enables early recognition, even in individuals unaware of meeting diagnostic criteria. This approach may aid in developing self-directed interventions, providing immediate mental health feedback, and facilitating timely treatment.

Digital phenotyping may play a role as a universal mental health screening tool, especially in school settings, assessing all children and adolescents irrespective of visible symptoms. Unlike the traditional practices that rely on observable symptoms noted by teachers and caregivers, the digital phenotyping approach may identify subtler indicators [19,25], enabling early detection of mental health issues [26] while students use their devices in schools. Then, subsequent diagnostic assessments can be used upon identification of concerns to make early intervention strategies [26].

The paradigm shift toward digital education in schools (eg, 1 device per student) may spur digital phenotyping methods for universal mental health screening, representing an unprecedented opportunity to apply digital phenotyping for adolescent mental health care. This is particularly evident in the context of the COVID-19 pandemic, where the ubiquity of smart devices among students has become a fundamental element of their classroom engagement. In certain countries, such as South Korea, all individual students from 3rd to 12th grade have been equipped with smart devices for both in-person classroom use and remote classroom use, initiating the gradual replacement of traditional textbooks from 2025 [27].

Privacy and Data Security

The research, which focused on nonclinical adolescent students grappling with mental health challenges, adopted a restricted yet meaningful data collection approach. Ensuring the security and privacy of data constituted a top priority in this study. Adherence to protocols overseen by the institutional review board and compliance with legal regulations pertaining to privacy, anonymization, and data security were essential components within public school settings under government supervision. Every aspect of data collection, storage, and processing within the cloud environments, automatically managed by the Dr. Simon module of the focuspang artificial intelligence system [28], met the requisite ethical and legal compliance standards.

This unified consensus underscored the critical importance of protecting student information throughout the data collection process without interfering with teaching and learning activities.

Consequently, data collection in these settings was restricted to minimize the acquisition of information typically gathered in clinical settings, such as birth date, race, socioeconomic status, explicit content involving words, facial expressions, voice recordings, and GPS details.

This Study

This data-driven study aimed to identify digital phenotypes for early detection of potential IGD among adolescent students in schools. Using sensor data from student tablet devices, the goal is to seamlessly integrate these digital markers into daily school routines, enabling timely identification and intervention for potential IGD cases. Emphasizing the need for seamless integration into adolescents' daily routines in schools, the study for future mental health care aims to facilitate the timely identification of potential IGD cases within the academic framework.

The exploration of leveraging these devices for mental health care, specifically targeting IGD, forms the core objective of this study. Sensor data were gathered from student tablets while they learned Korean literature and mathematics using the tablets to investigate the viability of integrating digital phenotyping as a means of early detection.

From the screening or early detection point of view, this study underscored the differentiation between primary and secondary tests for detecting IGD. Primary tests for screening prioritize sensitivity to inclusively capture individuals with IGD, whereas secondary tests focus on specificity to accurately identify those without the disorder. Screening tests serve as a gauge for disease probability and do not offer definitive diagnoses, necessitating further evaluation through subsequent diagnostic procedures for individuals with positive results [26].

It should be noted that in this initial exploratory phase, we opted for a statistical approach instead of machine learning, prioritizing interpretability rather than predictability for future research [29]. Linear regression models can provide clear insights into the ways variables affect outcomes, enabling the drawing of significant conclusions. They act as crucial initial steps in distinguishing the digital phenotyping variations between individuals with and without IGD.

This study can make several novel contributions. Initially, the study breaks new ground by pioneering the identification of digital markers linked to IGD. This identification process has the potential to significantly contribute to early detection and prompt intervention for IGD concerns. Such contributions are crucial in preventing the escalation of IGD, especially considering that many cases remain undiagnosed and untreated. Various factors, including limited awareness of IGD as a mental disorder, social stigma, and restricted access to mental health professionals, contribute to the untreated nature of many cases, making early detection and intervention paramount.

Additionally, with the digital transformation of schools, this study may extend traditional limited clinical contributions within school settings and provide psychiatrists with a more comprehensive understanding of how IGD manifests in the context of school environments. This contextual insight can inform tailored interventions and support strategies.

In addition, this study's focus on digital phenotyping as a tool for mental health assessment introduces a novel approach in psychiatry. The emphasis on seamlessly integrating digital markers into daily school routines showcases a practical approach to monitoring mental health. This integration could offer a continuous and nonintrusive method for tracking behavioral patterns, providing valuable insights for psychiatrists. If successful, it could pave the way for the integration of digital phenotyping into routine psychiatric assessments for a broader range of mental health conditions.

In summary, the research can contribute valuable insights and methods for early detection, intervention, and monitoring of IGD, thereby advancing psychiatry and offering practical tools for mental health care professionals.

Methods

Ethical Considerations

Our research protocol obtained approval from the Seoul Metropolitan Office of Education, ensuring compliance with Korean privacy and data security regulations. Additionally, it was approved by Korea University Institutional Review Board (KUIRB-2023-0159-01), aligning with the principles of the Helsinki Declaration. Before their involvement, comprehensive information about the research purpose, methods, and data collection was provided to students and their parents or legal guardians to obtain their informed consent. According to guidelines, the data collected did not include information that could potentially identify or discriminate against participants, such as birth date, ethnicity, income, or social status. Finally, as a token of appreciation for their participation, a book gift certificate worth ₩30,000 (about US $23) was provided to their parents or legal guardians.

Participants and Their Recruitment

Between September and October 2023, middle school students in Seoul, Republic of Korea, were recruited through an official website approved by the Seoul Metropolitan Office of Education, South Korea. Initially, 935 students voluntarily participated in the study. Exclusions were made for participants with neurodevelopmental conditions, central nervous system disorders, significant medical issues affecting psychological symptoms, or an inability to use this study's tablet computers.

After 8 (0.9%) of the 935 participants dropped out due to technical problems, the final number of participants was 927, and their average age was approximately 13 years, falling within the 13- to 14-year-old range. A total of 53% (89/168) of the participants were female, while 48% (79/168) were male.

Using the IGD Scale–Short-Form (IGDS9-SF) [30] for screening purposes, of the 927 participants, 85 (9.2%) were identified to be in the potential IGD group, while 842 (90.8%) were in the non-IGD group. To ensure balance for analysis, 83 participants without IGD were randomly selected from the pool. This approach aimed to mitigate potential biases introduced by uneven group distributions, providing a more robust foundation for our analysis. Among the total 168 participants, 89 (53%) were female, and 79 (47%) were male (Table 1).

Table 1. Demographic information of participants with their internet gaming disorder (IGD) scores.

Group and demographics	Participants, n (%)	IGD score, mean (SD)
Non-IGD (n=83)		
Female	56 (68)	3.6 (6.8)
Male	27 (33)	8.1 (9.1)
Total	83 (100)	5.3 (8.0)
Potential IGD (n=85)		
Female	33 (39)	22.7 (4.6)
Male	52 (61)	23.8 (4.3)
Total	85 (100)	23.4 (4.4)
Total (n=168)		
Female	89 (53)	8.7 (11.2)
Male	79 (47)	16.1 (11.4)
Total	168 (100)	12.2 (11.8)

Procedure

Before the main session, the individual participants were individually provided with tablet computers designated for studying Korean literature and mathematics. They practiced the system interface and found no usability or technical difficulties.

The main session consisted of two 45-minute segments separated by a 15-minute break where the participants were engaged in learning tasks while sensor data were collected in the background. In the first segment, they were individually asked to respond to 20 questions in Korean literature and, in the second, 20 questions in math. Their answers to 40 questions were used as their learning performance scores.

The participants were free to use their on-screen keyboard and touchscreens to respond to the questions. Although there was no time limit, all the questions were answered during the segments. The data gathering system was designed to efficiently oversee and orchestrate each step of the learning activities, including the presentation of questions and capturing participants' responses. Simultaneously, this system seamlessly recorded passive sensor data from the tablet, ensuring the collection of additional information related to user interaction and behaviors during the learning sessions. This comprehensive approach allowed for the simultaneous tracking of both academic engagement and the corresponding sensor data.

Following the main session, the participants were asked to fill out the questionnaires, including the IGDS9-SF [30].

IGD Scale and Screening Criterion

The IGD Scale (IGDS), introduced by Lemmens et al [31], served as a screening tool for IGD, encompassing 9 subfactors aligned with the provisional IGD criteria of the *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition* (*DSM-5*). Of the 2 versions of the scale, a 27-item and a 9-item version, this research used the Korean-adapted 9-item version of the scale, IGDS9-SF, validated to have reliability comparable to the original [32]. Each item assessed experiences over the preceding year on a Likert-type scale ranging from 0 to 5 points, yielding a total score range of 0-45. The internal consistency of this study, measured by Cronbach α, yielded a value of .84. A comprehensive review of 21 studies using the IGDS9-SF [30] reported that the scale demonstrated satisfactory internal consistency, robust criterion validity, and consistency across sex and age groups. While previous research [33] commonly used a diagnostic threshold of 21, this study opted for a lower threshold of 18, representing a 10% reduction. This adjustment came at the potential expense of a slightly increased type I error rate [26]. Monacis et al [33] demonstrated that using a cutoff score of 21, the sensitivity was 0.860 and the specificity was 0.861. When using a cutoff score of 18, the sensitivity was 0.818, and the specificity was 0.887.

Sensor Data Collection

While the participants were engaged in their learning tasks for middle-school mathematics and Korean literature on the given tablet PCs, digital marker data were collected from 2 passive sensors, the on-screen keyboard sensor and the touchscreen sensor, of the tablet computer. Students placed their tablets on the desktop for their typing and keyboarding, deliberately minimizing any potential noises that could arise from device movements. The device was a Samsung Galaxy Tab S7 FE with Android 11 that was identical to the one the participants used in their regular classes. Data collection occurred continuously in the background at a rate of 24 Hz, equivalent to 42.67 milliseconds, using a built-in data collection module called Dr. Simon while students engaged in learning sessions on tablets. The Dr. Simon module, integrated into the focuspang artificial intelligence system, autonomously managed all data collection and learning activities [28]. This module used various sensors, including eye-tracking, voice, stylus pen, touch, keyboarding, and other smart sensors. The research measures encompassed factors such as the frequency of key presses, angles, speed, acceleration, length, duration, path, and pressure.

Data Analysis for Interpretability over Predictability

Z-Score Transformation

The original measurement units of the indicators, which included self-rating, frequency, duration in milliseconds, length in pixels, pressure force, and ratio, were transformed using *z*-scores so that each indicator had a mean of 0 and SD of 1. As a result of the *z*-score transformation, indicators derived from different distributions could be compared.

Combined Analysis of Multiple Regression and Multivariate ANOVA

Multiple regression examines relationships between multiple predictors and a single outcome, whereas multivariate ANOVA (MANOVA) addresses multiple outcome variables at once. Together, they support a comprehensive analysis of the relationships between multiple predictors and outcomes simultaneously. Additionally, both techniques assist in adjusting for confounding variables: multiple regression controls potential confounders in independent and dependent variable relationships, while MANOVA evaluates the overall impact of independent variables on multiple dependent variables, accounting for their intercorrelations.

Bootstrapping Procedure

Bootstrapping, a statistical resampling technique, repeatedly samples with replacement from observed data to estimate the sampling distribution of a statistic. It is a method for robust estimates and inferences without relying on strict assumptions about the data distribution. A common approach to finding key indicators or measurements was to conduct multiple regression with a stepwise procedure. However, stepwise regression was regarded as unreliable and inadequate [34,35]. Due to this, multiple regression using a bootstrap approach was applied in this study. Using bootstrapping, 1000 sample sets were selected using the Mersenne Twisyer random number generator [34]. Furthermore, stratified sampling [36] was used to maintain the subgroup proportions of the sex types (female vs male) and IGD groups (potential IGD vs non-IGD).

Results

IGD and Learning Outcomes

A negative correlation was found between the scores on the IGDS and the learning performance scores on a 40-question learning assessment, indicated by r_{166}=–0.15, with a significance level of P<.05. This finding underscores an inverse relationship: higher IGDS scores were associated with lower academic performance.

Multiple Regression for Identification of Digital Phenotypes

This research used both multiple regression and MANOVA methods together to strengthen interpretative robustness. Initially, multiple regression was applied to examine how individual phenotypes relate to IGD scores. Subsequently, MANOVA was used to investigate how the classification into IGD groups might impact various digital phenotypes, allowing for a broader assessment of these connections. The use of both methods serves as a form of cross-validation. When the results from both multiple regression and MANOVA align, it bolsters the consistency of the findings, reinforcing the established relationships identified through diverse analyses.

First, initial multiple regression models were built to pinpoint indicators that effectively identify IGD. Additionally, bootstrapping was used to generate a sufficient data set even with limited data, enabling the estimation of a population’s mean. The pool of potential indicators comprised 38 candidates, encompassing various aspects of keyboard usage. This included the frequencies of specific keys, such as delete, space, backspace, keyboard switch, enter, and shift. Additionally, the pool incorporated metrics related to stroke behaviors, such as frequencies, means, or SDs of total length, horizontal length, vertical length, duration, pressure, number of points, horizontal direction, vertical direction, horizontal size, vertical size, speed, and acceleration. Further metrics involved means and SDs of keypress-to-keypress duration, pause duration before pressing enter, and release-to-press duration, among others. The assessment also considered the number of correct answers and incorrect responses.

After addressing multicollinearity issues, the final regression model selected 5 indicators (Table 2). The regression yielded statistically significant results ($F_{5,\ 162}$=9.63; mean square error=2.09; P<.001), indicating that these indicators collectively accounted for R=0.47, r^2=23% of the variance in IGDS.

Table 2. Correlation and multiple regression results between the Internet Gaming Disorder Scale and its indicators.

Variable	Correlation results					Multiple regression results		
	X1	X2	X3	X4	X5	β[a]	*t* test (*df*=162)	*P* value[b]
X1. Stroke acceleration						0.33	3.81	<.001
r	1	0.25	–0.38	–0.27	0.52			
P value	—[c]	.001	<.001	<.001	<.001			
X2. Time interval between keys						–0.26	–3.18	.01
r	0.25	1	–0.26	–0.53	0.19			
P value	.001	—	<.001	<.001	.01			
X3. Word spacing						–0.25	–3.28	<.001
r	–0.38	–0.26	1	0.19	–0.18			
P value	<.001	<.001	—	.01	.02			
X4. Deletion						–0.24	–2.89	<.001
r	–0.27	–0.53	0.19	1	–0.24			
P value	<.001	<.001	.01	—	.002			
X5. Horizontal length of strokes						0.21	2.59	.02
r	0.52	0.19	–0.18	–0.24	1			
P value	<.001	.01	.02	.002	—			

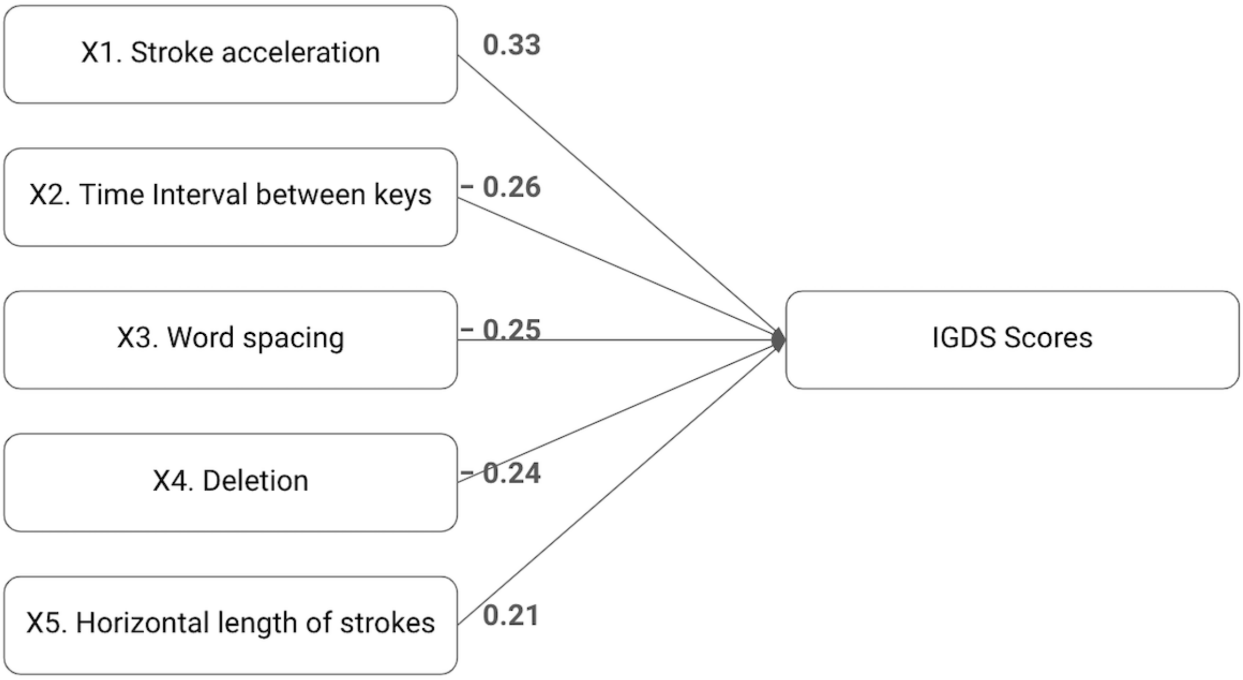
[a]β: standardized regression coefficients.

[b]All zero-order correlations are significant (*P*<.05).

[c]Not applicable.

As shown in Table 2 and Figure 1, the IGDS scores, the dependent measure, were positively associated with the first indicator: stroke acceleration (X1). Stroke acceleration (X1) was linked to higher IGDS scores. The stroke acceleration was the speed change rate while stroking on the touch screen. When this speed increased by 1, the IGDS score increased by 0.33, assuming all other factors remained constant. Conversely, all the other indicators—time interval between keys (X2), word spacing (X3), deletion (X4), and horizontal length of strokes (X5)—showed negative associations with the IGDS scores. The time interval between keys measures how quickly the participant typed, with faster typing tending to indicate higher IGDS scores. Word spacing refers to how much space there is between letters, words, or sentences while typing. A decrease in word spacing was associated with higher IGDS scores. Deletion represents removing characters during typing. For every 1-unit increase in deletion frequency, there was an expected decrease of 0.24 units in IGDS scores, meaning lower deletion frequency is related to lower IGDS scores. Lastly, an increase in the horizontal length of strokes indicated higher IGDS scores.

Figure 1. Multiple regression model based on digital phenotypes. IGDS: Internet Gaming Disorder Scale.



In summary, the β coefficients, serving as effect sizes, offer insights into the direction and strength of relationships between each independent variable (X1-X5) and the dependent variable, the IGDS scores. Notably, all independent variables exhibit statistically significant relationships with the IGDS scores. The magnitude of the β coefficients indicates moderate relationships, with larger absolute values signifying stronger effects.

MANOVA, Cross-Validation Through Group Comparison

To cross-validate the outcomes of the regression analysis, a comparison between the potential IGD and non-IGD groups was conducted following the method outlined by Cronbach and Meehl [37]. Using MANOVA as the methodological approach, the study used the groups as the independent variable to assess their influence on the 5 indicators, treated as dependent variables (Table 3).

Table 3. Multivariate ANOVA comparison between the internet gaming disorder (IGD) and non-IGD groups.

Dependent variable	IGD, mean (SD)	Non-IGD, mean (SD)	Sum of square (*df*)	Mean squares	*F* test (*df*=163)	*P* value	Cohen *d*
X1. Stroke acceleration	–0.28 (0.85)	0.40 (1.12)	19.32 (1)	19.32	19.51	<.001	0.68
X2. Time interval between keys	0.12 (1.04)	0.04 (0.95)	0.31 (1)	0.31	0.31	.58	0.09
X3. Word spacing	0.21 (1.01)	–0.36 (0.99)	13.67 (1)	13.67	13.65	<.001	0.57
X4. Deletion	0.11 (1.03)	–0.22 (0.95)	4.43 (1)	4.43	4.52	.04	0.33
X5. Horizontal length of strokes	–0.03 (1.03)	0.34 (1.16)	6.02 (1)	6.02	4.98	.03	0.34

The multivariate test yielded a substantial Wilks λ value of 0.799, accompanied by $F_{5,162}$=8.164, with a significance level of *P*<.001. The corresponding effect size, measured through η-squared, revealed a moderate magnitude of 0.23.

Upon conducting univariate ANOVAs, noteworthy disparities (*P*<.01) surfaced between the IGD and non-IGD groups concerning stroke acceleration (X1; *P*<.001), word spacing (X3; *P*<.001), deletion (X4; *P*=.04), and horizontal length of strokes (X5; *P*=.03). Specifically, potential IGD students displayed faster stroke acceleration, lower word-spacing tendencies, decreased deletion behavior, and longer horizontal strokes compared to non-IGD students.

However, no significant difference was observed in the time interval (X2; *P*=.48) between the groups. The divergence between the outcomes of multiple regression and MANOVA could stem from their inherent methodological dissimilarities: multiple regression focuses on predicting a single dependent variable using multiple predictors, while MANOVA evaluates the collective impact of 1 or more independent variables on multiple dependent variables simultaneously. This variation in approach might lead to discrepancies in determining significance [38]. Notably, considering that multicollinearity was controlled in the multiple regression, it appears to have limited influence on these findings.

Discussion

Overview

The aim of this exploratory study was to find digital indicators to be used for early detection of potential IGD among adolescent students in their learning activities. The surge in digital education trends—1 device per student—in public schools may present an unprecedented avenue for deploying digital phenotyping as a universal mental health screening tool in the management of young individuals' mental health and academic progress. Thus, digital phenotyping holds the potential to disrupt the detrimental cycle wherein school experiences and IGD mutually exacerbate each other [39,40].

Principal Findings

An initial investigation into the connection between IGD and academic performance uncovered a negative correlation, indicating that higher IGDS scores were associated with lower learning performances.

The main analyses, the multiple regression and MANOVA procedures with bootstrapping revealed 5 markers: stroke acceleration, the time interval between keys, word spacing, deletion, and the horizontal length of strokes. These findings suggest that the identified digital markers for IGD are reliable across multiple methods and may offer valuable insights into understanding how IGD operates in a learning context, connecting clinically useful knowledge for addressing the issue of IGD in educational settings.

The identified indicators seem to suggest that adolescent students with IGD tend to engage in learning activities rapidly and roughly rather than meticulously and accurately. This type of speedy behavior characterizes these students. This aligns with previous research by Kim et al [2], who found that children displaying gaming addiction are more likely to exhibit impulsive behavior, defined as a fast reaction without thinking or conscious judgment [2]. Similarly, Şalvarlı and Griffiths [41] conducted a comprehensive review of 33 empirical studies involving a total of 18,128 participants and reported that, with 1 exception, all 32 studies showed a positive association between impulsivity and IGD.

The tendency of students with IGD to act quickly and impulsively may negatively impact their learning outcomes. Impulsive students are more prone to making errors due to the speed-accuracy trade-off theory [42], which suggests that acting quickly can lead to reduced accuracy. Several neurophysiological studies have also linked IGD to impulsivity, impaired response inhibition, and dysfunctional attentional bias [2,3]. Moreover, Lee et al [4] found that patients with IGD have impaired reading ability.

Along with the negative correlation between the number of correct answers as learning outcomes and the IGDS score, all of the digital indicators—the stroke acceleration (r=–0.41), the time interval between keys (r=–0.30), word spacing (r=0.42), deletion (r=0.16), and horizontal length of strokes (r=–0.29)—were significantly correlated with the number of correct answers at P<.05. These correlations collectively further reinforce the notion that impulsive and speedy behaviors in IGD students may impact their learning outcomes.

The study concludes by suggesting a practical application of digital phenotyping in schools, particularly due to the recent digitalization of education. Smart devices are now available in in-person classes as well as in remote learning settings [27,43]. Using these devices, digital markers for IGD can be used to diagnose and improve students' mental health while they are in school. Digital phenotypes offer a potential solution to break the vicious cycle between learning and IGD [39,40,44,45].

Limitations

While the findings of this study suggest the potential use of digital markers for screening IGD during learning activities in schools, it is essential to approach the results with caution, considering several limitations.

First, before accepting digital markers as direct proxies for IGD, further research and validation are imperative. The study used an inductive or data-driven approach due to the limited theoretical frameworks for digital phenotyping in mental health. Consequently, the identified indicators require scrutiny for psychiatric validity. An additional constraint in this study lies in the lack of confirmed IGD diagnoses following the early detection or screening phase. Future research should integrate a comprehensive diagnostic process to validate the presence of IGD.

Second, it is crucial to note that in this study, IGD was treated as a unitary construct. However, Wang et al [46] identified at least 2 subgroups of people with IGD, each exhibiting different brain functional connectivity patterns and distinct psychiatric symptom profiles. Therefore, future phenotypic studies need IGD as a multifaceted and complex system.

Third, the IGDS questionnaire served as the standard for recognizing digital phenotypes. However, concerns about the credibility of psychiatric disorders lacking objective measures and depending on subjective self-reporting have prompted neuroscientist Insel [47] to characterize digital phenotyping as a "new behavioral science."

Fourth, it is crucial to test phenotypes across diverse age groups, particularly involving younger individuals. Younger individuals often display more pronounced and rapid responses to various stimuli, making them essential for understanding the initial effects of certain phenotypes or interventions. Studying phenotypes in younger age groups can offer insights into the early onset of certain traits or conditions, allowing for proactive interventions or preventive measures. This is particularly vital in fields such as health care or education, where early identification of certain phenotypes can enable tailored interventions to support healthy development or address potential issues before they escalate.

Fifth, digital phenotyping faces various challenges, with data quality standing out among them. Typically relying on data collected from smartphones or smartwatches in natural environments, these studies encounter distinct patterns due to device movements, differences among individuals, and variations within the same individual across multiple observations. This can lead to data gaps and irregularities. However, in this specific study, data maintained a high standard due to consistent collection methods with minimal missing data and disturbances. For example, placing tablets on students' desks helped minimize disruptions from their movements. Furthermore, all participants underwent similar conditions, spending about similar time on 2 similar tasks, thereby mitigating variations between individuals and within individuals.

Conclusions

The findings establish a correlation between IGD and academic performance, revealing significant distinctions in digital phenotypes between potential IGD and non-IGD cohorts. This highlights the potential use of specific digital markers for early identification of potential IGD students. However, it is imperative to emphasize the necessity for further research and validation to bolster the reliability of these digital markers and deepen our comprehension of the intricate nature of IGD.

Finally, the study proposes a practical application of digital phenotyping, considering the prevalence of smart devices not only in traditional classrooms but also in remote learning environments [27,47]. By leveraging these devices, digital markers for IGD can be deployed to diagnose and enhance students' mental well-being during their time in school. Essentially, digital phenotypes present a promising avenue to disrupt the detrimental cycle between learning and IGD.

Acknowledgments

This work was made possible with financial support from multiple organizations. Specifically, it was funded by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant from the Ministry of Science and ICT (MSIT) of Korea (project 2021-0-0268), Artificial Intelligence Innovation Hub (Artificial Intelligence Institute, Seoul National University). Additional financial support came from the Bio and Medical Technology Development Program through the National Research Foundation (NRF) of Korea, which was funded by MSIT (project 2021M3A9E408078412). Moreover, the Alchemist Project (1415187369 and 20025696) was funded by the Ministry of Trade, Industry, and Energy (MOTIE, Korea), and a portion of the support also came from NRF funded by MSIT (project 2023R1A2C3003007).

The authors extend their deep gratitude to the Seoul Metropolitan Office of Education for their outstanding support and collaboration. As the Office spearheaded the Digital Buddy (Di-Bud) project aimed at digitally transforming schools, their assistance was invaluable in gathering essential digital data. These data are vital for schools to break the destructive cycle connecting internet gaming disorder with academic performance and student well-being.

Furthermore, we would like to thank the Korea Institute of Patent Information and the National Information Society Agency of Korea for their generous support. Their contributions have been pivotal in enabling our study, enhancing its quality, and ensuring its successful completion.

Conflicts of Interest

The authors, KC and CGL, have a conflict of interest as they are shareholders of 3R Innovation Inc, which owns the Dr. Simon product.

References

1. Dong G, Potenza MN. A cognitive-behavioral model of internet gaming disorder: theoretical underpinnings and clinical implications. J Psychiatr Res 2014;58:7-11 [FREE Full text] [doi: 10.1016/j.jpsychires.2014.07.005] [Medline: 25062755]
2. Kim M, Lee TH, Choi JS, Kwak YB, Hwang WJ, Kim T, et al. Neurophysiological correlates of altered response inhibition in internet gaming disorder and obsessive-compulsive disorder: perspectives from impulsivity and compulsivity. Sci Rep 2017;7:41742 [FREE Full text] [doi: 10.1038/srep41742] [Medline: 28134318]
3. Kim M, Lee TH, Choi JS, Kwak YB, Hwang WJ, Kim T, et al. Dysfunctional attentional bias and inhibitory control during anti-saccade task in patients with internet gaming disorder: an eye tracking study. Prog Neuropsychopharmacol Biol Psychiatry 2019;95:109717. [doi: 10.1016/j.pnpbp.2019.109717] [Medline: 31351161]
4. Lee TH, Kim M, Kwak YB, Hwang WJ, Kim T, Choi JS, et al. Altered eye-movement patterns during text reading in obsessive-compulsive disorder and internet gaming disorder. Front Behav Neurosci 2018;12:248 [FREE Full text] [doi: 10.3389/fnbeh.2018.00248] [Medline: 30405372]
5. Zhou Z, Zhou H, Zhu H. Working memory, executive function and impulsivity in internet-addictive disorders: a comparison with pathological gambling. Acta Neuropsychiatr 2016;28(2):92-100. [doi: 10.1017/neu.2015.54] [Medline: 26400106]
6. Teng Z, Pontes HM, Nie Q, Xiang G, Griffiths MD, Guo C. Internet gaming disorder and psychosocial well-being: a longitudinal study of older-aged adolescents and emerging adults. Addict Behav 2020;110:106530. [doi: 10.1016/j.addbeh.2020.106530] [Medline: 32683173]
7. Diagnostic and Statistical Manual of Mental Disorders: DSM-5-TR, 5th Edition. Washington, DC: American Psychiatric Association; 2022.

8. Mestre-Bach G, Fernandez-Aranda F, Jiménez-Murcia S. Exploring internet gaming disorder: an updated perspective of empirical evidence (from 2016 to 2021). Compr Psychiatry 2022;116:152319 [FREE Full text] [doi: 10.1016/j.comppsych.2022.152319] [Medline: 35526417]
9. García-García E, Martínez-Gimeno ML, Benítez-Andrades JA, Miranda-Gómez J, de Dios Zapata-Cornejo E, Escobar-Aguilar G. Substance use and addictive behavior in Spanish adolescents in secondary school. Healthcare (Basel) 2021;9(2):186 [FREE Full text] [doi: 10.3390/healthcare9020186] [Medline: 33572308]
10. Chen KH, Oliffe JL, Kelly MT. Internet gaming disorder: an emergent health issue for men. Am J Mens Health 2018;12(4):1151-1159 [FREE Full text] [doi: 10.1177/1557988318766950] [Medline: 29606034]
11. Feng W, Ramo DE, Chan SR, Bourgeois JA. Internet gaming disorder: trends in prevalence 1998-2016. Addict Behav 2017;75:17-24 [FREE Full text] [doi: 10.1016/j.addbeh.2017.06.010] [Medline: 28662436]
12. Percentage of public schools with various safety and security measures: selected years, 1999-2000 through 2019-20. National Center for Education Statistics. 2022. URL: https://nces.ed.gov/programs/digest/d21/tables/dt21_233.50.asp [accessed 2023-06-14]
13. School bans on mobile phones violate students' human rights. The Korea Times. 2021. URL: https://www.koreatimes.co.kr/www/nation/2022/07/181_318152.html [accessed 2023-06-14]
14. Arango C, Díaz-Caneja CM, McGorry PD, Rapoport J, Sommer IE, Vorstman JA, et al. Preventive strategies for mental health. Lancet Psychiatry 2018;5(7):591-604. [doi: 10.1016/S2215-0366(18)30057-9] [Medline: 29773478]
15. Ricci V, De Berardis D, Maina G, Martinotti G. Internet gaming disorders and early onset psychosis in young people: a case study and clinical observations. Int J Environ Res Public Health 2023;20(5):3920 [FREE Full text] [doi: 10.3390/ijerph20053920] [Medline: 36900931]
16. Keyes CLM. Promoting and protecting mental health as flourishing: a complementary strategy for improving national mental health. Am Psychol 2007;62(2):95-108. [doi: 10.1037/0003-066X.62.2.95] [Medline: 17324035]
17. Levitt JM, Saka N, Romanelli LH, Hoagwood K. Early identification of mental health problems in schools: the status of instrumentation. J School Psychol 2007;45(2):163-191. [doi: 10.1016/j.jsp.2006.11.005]
18. Gill PJ, Saunders N, Gandhi S, Gonzalez A, Kurdyak P, Vigod S, et al. Emergency department as a first contact for mental health problems in children and youth. J Am Acad Child Adolesc Psychiatry 2017;56(6):475-482.e4 [FREE Full text] [doi: 10.1016/j.jaac.2017.03.012] [Medline: 28545752]
19. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. NPJ Digit Med 2019;2:88 [FREE Full text] [doi: 10.1038/s41746-019-0166-1] [Medline: 31508498]
20. Ghosh S, Hiware K, Ganguly N, Mitra B, De P. Emotion detection from touch interactions during text entry on smartphones. Int J Hum Comput Stud 2019;130:47-57. [doi: 10.1016/j.ijhcs.2019.04.005]
21. Lee PM, Tsui WH, Hsiao TC. The influence of emotion on keyboard typing: an experimental study using auditory stimuli. PLoS One 2015;10(6):e0129056 [FREE Full text] [doi: 10.1371/journal.pone.0129056] [Medline: 26065902]
22. Nahin AFMNH, Alam JM, Mahmud H, Hasan K. Identifying emotion by keystroke dynamics and text pattern analysis. Behav Inf Technol 2014;33(9):987-996. [doi: 10.1080/0144929x.2014.907343]
23. Narkhede SM, Luther L, Raugh IM, Knippenberg AR, Esfahlani FZ, Sayama H, et al. Machine learning identifies digital phenotyping measures most relevant to negative symptoms in psychotic disorders: implications for clinical trials. Schizophr Bull 2022;48(2):425-436 [FREE Full text] [doi: 10.1093/schbul/sbab134] [Medline: 34915570]
24. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018;21(3):116-119 [FREE Full text] [doi: 10.1136/eb-2018-102891] [Medline: 29871870]
25. Nisenson M, Lin V, Gansner M. Digital phenotyping in child and adolescent psychiatry: a perspective. Harv Rev Psychiatry 2021;29(6):401-408. [doi: 10.1097/HRP.0000000000000310] [Medline: 34313626]
26. Maxim LD, Niebo R, Utell MJ. Screening tests: a review with examples. Inhal Toxicol 2014;26(13):811-828 [FREE Full text] [doi: 10.3109/08958378.2014.955932] [Medline: 25264934]
27. Korea Education and Research Information Service (KERIS). Chapter 18: Korea. In: Country Digital Education Ecosystems and Governance: A Companion to Digital Education Outlook 2023. Paris: OECD Publishing; 2023.
28. Dr. Simon. 3R Innovation Inc. URL: https://drsimon.ai [accessed 2023-06-14]
29. McCain K. Explanation and prediction. In: Understanding How Science Explains the World. Cambridge, UK: Cambridge University Press; 2022:43-54.
30. Poon LYJ, Tsang HWH, Chan TYJ, Man SWT, Ng LY, Wong YLE, et al. Psychometric properties of the Internet Gaming Disorder Scale-Short-Form (IGDS9-SF): systematic review. J Med Internet Res 2021;23(10):e26821 [FREE Full text] [doi: 10.2196/26821] [Medline: 34661543]
31. Lemmens JS, Valkenburg PM, Gentile DA. The Internet Gaming Disorder Scale. Psychol Assess 2015;27(2):567-582. [doi: 10.1037/pas0000062] [Medline: 25558970]
32. Kim BN, Ko H. Psychometric properties of the nine-item Korean Internet Gaming Disorder Scale: Short Form. Cyberpsychol Behav Soc Netw 2020;23(12):854-859. [doi: 10.1089/cyber.2020.0227] [Medline: 32960098]

33. Monacis L, de Palo V, Griffiths MD, Sinatra M. Validation of the Internet Gaming Disorder Scale—Short-Form (IGDS9-SF) in an Italian-speaking sample. J Behav Addict 2016;5(4):683-690 [FREE Full text] [doi: 10.1556/2006.5.2016.083] [Medline: 27876422]
34. Derksen S, Keselman HJ. Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. Br J Math Stat Psychol 2011;45(2):265-282. [doi: 10.1111/j.2044-8317.1992.tb00992.x]
35. Kuo L, Mallick B. Variable selection for regression models. Sankhya Ser B 1998;60(1):65-81 [FREE Full text]
36. Wilcox RR, Keselman HJ. Modern robust data analysis methods: measures of central tendency. Psychol Methods 2003;8(3):254-274. [doi: 10.1037/1082-989X.8.3.254] [Medline: 14596490]
37. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull 1955;52(4):281-302. [doi: 10.1037/h0040957] [Medline: 13245896]
38. Nelson LR, Zaichkowsky LD. A case for using multiple regression instead of ANOVA in educational research. J Exp Educ 1979;47(4):324-330. [doi: 10.1080/00220973.1979.11011701]
39. Samaha M, Hawi N. Internet gaming disorder and its relationships with student engagement and academic performance. In: Research Anthology on Game Design, Development, Usage, and Social Impact. Hershey, PA: IGI Global; 2023:1704-1720.
40. Wright JM. The effects of video game play on academic performance. Mod Psychol Stud 2011;17(1):37-44 [FREE Full text] [doi: 10.1037/e568882012-005]
41. Şalvarlı Ş, Griffiths MD. The association between internet gaming disorder and impulsivity: a systematic review of literature. Int J Ment Health Addiction 2019;20:92-118 [FREE Full text] [doi: 10.1007/s11469-019-00126-w]
42. Zimmerman ME. Speed-accuracy tradeoff. In: Kreutzer JS, DeLuca J, Caplan B, editors. Encyclopedia of Clinical Neuropsychology. New York, NY: Springer; 2011:2344-2344.
43. Broadband Commission for Sustainable Development, International Telecommunication Union, UNESCO, United Nations Children's Fund, Sepúlveda A. The Digital Transformation of Education: Connecting Schools, Empowering Learners. New York, NY: United Nations; 2020.
44. Kuss DJ, Lopez-Fernandez O. Internet addiction and problematic internet use: a systematic review of clinical research. World J Psychiatry 2016;6(1):143-176 [FREE Full text] [doi: 10.5498/wjp.v6.i1.143] [Medline: 27014605]
45. Xu J, Shen LX, Yan CH, Hu H, Yang F, Wang L, et al. Personal characteristics related to the risk of adolescent internet addiction: a survey in Shanghai, China. BMC Public Health 2012;12:1106 [FREE Full text] [doi: 10.1186/1471-2458-12-1106] [Medline: 23259906]
46. Wang HY, Cheng C. The associations between gaming motivation and internet gaming disorder: systematic review and meta-analysis. JMIR Ment Health 2022;9(2):e23700 [FREE Full text] [doi: 10.2196/23700] [Medline: 35175204]
47. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA 2017;318(13):1215-1216. [doi: 10.1001/jama.2017.11295] [Medline: 28973224]

Abbreviations

DSM-5: Diagnostic and Statistical Manual of Mental Disorders, 5th Edition
DSM-5-TR: Diagnostic and Statistical Manual of Mental Disorders, 5th Edition, Text Revision
IGD: internet gaming disorder
IGDS: Internet Gaming Disorder Scale
IGDS9-SF: Internet Gaming Disorder Scale–Short-Form
MANOVA: multivariate ANOVA

Edited by J Torous; submitted 18.07.23; peer-reviewed by R Orji, M Metcalf, NH Zainal; comments to author 30.10.23; revised version received 23.01.24; accepted 26.02.24; published 29.04.24.

Please cite as:
Cho K, Kim M, Cho Y, Hur JW, Kim DH, Park S, Park S, Jang M, Lee CG, Kwon JS
Digital Phenotypes for Early Detection of Internet Gaming Disorder in Adolescent Students: Explorative Data-Driven Study
JMIR Ment Health 2024;11:e50259
URL: https://mental.jmir.org/2024/1/e50259
doi:10.2196/50259
PMID:38683658

Review

Web-Based Forums for People Experiencing Substance Use or Gambling Disorders: Scoping Review

Annette Peart[1], BOccThy, MHlthSc, PhD; Freya Horn[1], BA, GDA-PSYCH; Rachel Petukhova[1], BPsych(Hons); Anthony Barnett[1], PhD; Dan I Lubman[1], BSc(Hons), MBChB, PhD

Eastern Health Clinical School, Monash University, Richmond, Victoria, Australia

Corresponding Author:
Annette Peart, BOccThy, MHlthSc, PhD
Eastern Health Clinical School
Monash University
Level 2, 110 Church St
Richmond, Victoria, 3121
Australia
Phone: 61 3 8413 8413
Email: annette.peart@monash.edu

Abstract

Background: For people experiencing substance use or gambling disorders, web-based peer-supported forums are a space where they can share their experiences, gather around a collective goal, and find mutual support. Web-based peer support can help to overcome barriers to attending face-to-face meetings by enabling people experiencing addiction to seek support beyond their physical location and with the benefit of anonymity if desired. Understanding who participates in web-based peer-supported forums (and how), and the principles underpinning forums, can also assist those interested in designing or implementing similar platforms.

Objective: This study aims to review the literature on how people experiencing substance use or gambling disorders, and their family, friends, and supporters, use and participate in web-based peer-supported forums. Specifically, we asked the following research questions: (1) What are the characteristics of people who use web-based peer-supported substance use or gambling-focused forums? (2) How do people participate in web-based peer-supported forums? (3) What are the key principles reportedly underpinning the web-based peer-supported forums? (4) What are the reported outcomes of web-based peer-supported forums?

Methods: Inclusion criteria for our scoping review were peer-reviewed primary studies reporting on web-based addiction forums for adults and available in English. A primary search of 10 databases occurred in June 2021, with 2 subsequent citation searches of included studies in September 2022 and February 2024.

Results: Of the 14 included studies, the majority of web-based peer-supported forums reported were aimed specifically for, or largely used by, people experiencing alcohol problems. Results from the 9 studies that did report demographic data suggest forum users were typically women, aged between 40 years and early 50 years. Participation in web-based peer-supported forums was reported quantitatively and qualitatively. The forums reportedly were underpinned by a range of key principles, mostly mutual help approaches and recovery identity formation. Only 3 included studies reported on outcomes for forum users.

Conclusions: Web-based peer-supported forums are used by people experiencing addiction in a number of ways, to share information and experiences, and give and receive support. Seeking web-based support offers an alternative approach to traditional face-to-face support options, and may reduce some barriers to engaging in peer support.

(JMIR Ment Health 2024;11:e49010) doi:10.2196/49010

KEYWORDS

web-based forums; peer support; substance use; gambling; scoping; review method; review methodology; forum; forums; substance abuse; addiction; addictive; addictions; peer-based; peer support

Introduction

People with substance use disorders or experiencing gambling problems can feel shame and stigma, contributing to increased social isolation and delayed help-seeking [1-4]. Peer support, which involves the sharing of experiences, knowledge, support, and practical help among people with lived experiences of similar issues [5,6], has a long history in substance use and gambling recovery [7,8]. Peer support has been particularly effective in overcoming shame and stigma, creating spaces built on shared experiences where people can connect safely and learn about help-seeking [5,7].

Peer support has traditionally been accessed through face-to-face meetings, such as 12-step or within therapeutic communities [7]. However, many people face barriers to accessing face-to-face peer support, including geographical distance, regional and rural service gaps, and fear of stigma, as well as insufficient time to travel to and attend meetings, amid the general demands of work, family, and life. The expansion of the digital world means that opportunities to connect with peers have grown. Web-based peer support can help to overcome barriers to attending face-to-face meetings by enabling people experiencing addiction to seek support beyond their physical location and with the benefit of anonymity if desired [9].

Increasingly, web-based peer support is available in web-based discussion communities or forums hosted on social media platforms or websites. Web-based forum users initiate discussions by starting a thread, responding to users' threads, or scrolling or searching past threads for content of interest. Forums typically only require an email address and self-selected username to post a discussion. Web-based forums, therefore, provide anonymity, access to peer support at any hour and from any location, and the option for both synchronous and asynchronous discussions [10-12].

For a range of health issues, engagement in web-based forums provides benefits for forum users, including improvements in mood, connectedness, and social support, and access to practical support and advice [13-15]. For people experiencing substance use or gambling disorders, web-based forums are a space where they can share their experiences, gather around a collective goal, and find mutual support [12,16]. This type of sharing, particularly with peers who have lived or living experience of substance use or gambling disorders is a key principle of recovery [17].

While there are limited reviews specifically on substance use or gambling web-based forums, a recent systematic review looked at digital recovery support services (D-RSS) led by substance use specialists or peers for people with a substance use disorder [18]. Ashford et al [18] identified 22 studies of various web-based services, including peer-based communities such as recovery social networking sites and web-based forums, and nonpeer-based interventions such as mobile text messaging and digital module-based learning. The authors found that while the evidence of the effectiveness of D-RSS for improving recovery-related outcomes was currently lacking, these services have high use and could overcome accessibility and availability barriers. Ashford et al [18] also called for further exploration of how, why, and to what extent people participate in D-RSS. Given the diversity of services identified in the systematic review, focusing on a subset, such as web-based peer-supported forums, may provide clearer information on the use and utility of these services. Additionally, by including gambling-focused digital services, which were omitted from the review by Ashford et al [18], we can broaden the focus to include web-based peer-supported forums on any addiction. Understanding who participates in the web-based peer-supported forums (and how), and the principles underpinning forums, can also assist those interested in designing or implementing similar platforms.

We conducted a scoping review on how people experiencing substance use or gambling disorders, or their family, friends, and supporters, use and participate in web-based peer-supported forums. Specifically, we asked the following research questions: (1) What are the characteristics of people who use web-based peer-supported substance use or gambling-focused forums? (2) How do people participate in web-based peer-supported forums? (3) What are the key principles reportedly underpinning the web-based peer-supported forums? (4) What are the reported outcomes of web-based peer-supported forums?

Methods

Scoping reviews can assist in summarizing findings from a heterogeneous body of knowledge and identifying gaps in the literature [19]. Our work was structured around the 5 stages of the Arksey and O'Malley [20] framework and informed by Levac et al's [21] refinements to this framework.

We used a 2-step search strategy. First, in June 2021, we searched MEDLINE (OVID), PsycINFO, CINAHL, Emcare (OVID), AMED, Web of Science, Scopus, Central Register of Controlled Trials, Informit, and Sociological Abstracts using the terms: addictive behavior, substance-related disorders, gambling, web-based social networking, and term variations (eg, web-based, internet, community, dependence, and addiction). The complete search strategy for MEDLINE can be found in Multimedia Appendix 1. An academic librarian reviewed and refined our search strategy. Second, in September 2022 and February 2024, we performed a citation search of our original included studies to obtain more recent studies.

Our inclusion criteria were as follows: (1) peer-reviewed primary studies, (2) reporting on web-based addiction-focused forums, (3) web-based forums designed for adults (aged 18 years and older), and (4) those available in English. In relation to the second inclusion criterion, the studies needed to report results based on data derived from forum posts or metrics. We excluded gray literature and full-text conference proceedings. Studies where reference was made to web-based forums, but data were not derived from the forums, were excluded; for example, studies where data were about participant perceptions of forums in general. We also excluded studies where addiction was not the focus of the forum, or interventions were apps, rather than forums. No restrictions on the date of publication were used.

We used Covidence (Veritas Health Innovation), a web-based collaboration software platform, to facilitate screening study titles and abstracts, and again for full-text review. Two independent authors (AP, RP, FH, or AB) reviewed the titles and abstracts of studies for inclusion or exclusion. Two independent authors (AP, RP, FH, or AB) then extracted the data from the full text of the included studies. We resolved conflicts on study selection by consensus and team discussion if needed.

Using the research questions as a guide, we developed a data extraction template in Covidence. We extracted the following data from the included studies: study characteristics (eg, country of origin, year, and study type); web-based forum characteristics (eg, principles, models or theories reported, forum user characteristics, use, and participation); and outcomes reported. After each author independently charted the data, we discussed the results and updated the form through an iterative process. A formal assessment of methodological quality was not part of this scoping review.

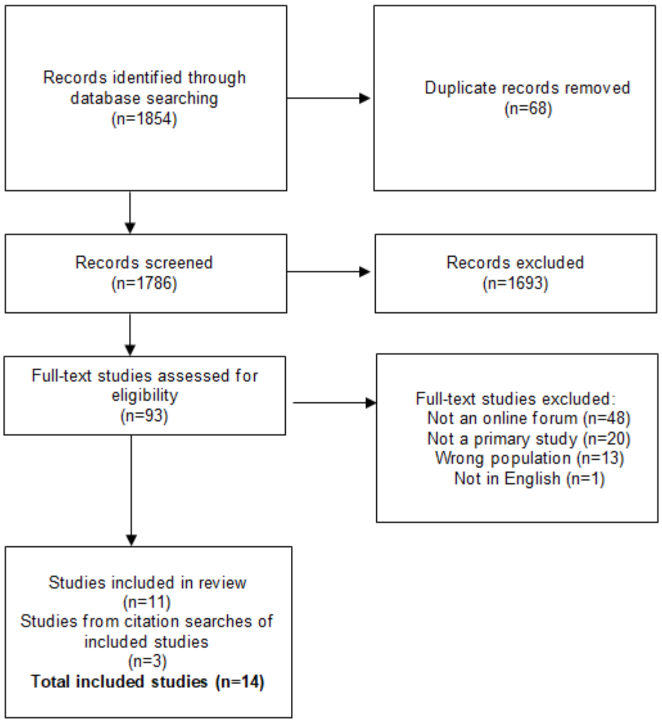
We used Microsoft Word to visually display our data extraction as tables, and for the analysis, we discussed descriptive findings as a group. We grouped the studies according to our research questions and summarized the key findings numerically and thematically. We met regularly throughout the project, incorporating reflexive elements to consider how we analyzed the data and where data best fit to answer our research questions.

Results

Selection of Sources

Our initial search terms generated 1854 records from the databases. We removed 68 duplicates, leaving 1786 to be screened according to title and abstract, of which 1693 were excluded. After that screening, 93 full-text studies were reviewed, and 82 were excluded, leaving 11 studies included as a result of the database search. The citation search of included studies resulted in the addition of 2 further studies in September 2022, and 1 further study in February 2024. The final number of included studies was 14 (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow chart.



Characteristics of Sources

Of the 14 included studies, 5 studies originated in the United Kingdom, 3 studies in each of the United States and Australia, 1 study in each of Canada and Russia, and 1 study not confined to a specific country, reporting on a range of web-based peer-supported forums. The majority of web-based peer-supported forums reported in the studies were aimed specifically for, or largely used by, people experiencing alcohol problems. Two studies were not specific to any substance; however, these reported that more than half of the people using the web-based forum were seeking alcohol support (306/343, 91.6% alcohol [22] and 79/123, 65% alcohol [23]). Two forums focused on people with opioid concerns (1 codeine-specific [24] and 1 for pregnant women with opioid use problems [25]), and 2 focused on gambling concerns [26,27]. Two web-based peer-supported forums were the focus of 5 studies: Hello Sunday Morning ([HSM] Australia, 3 studies) [28-30] and Soberistas (United Kingdom, 2 studies) [12,31]. For 3 studies, the authors did not name the web-based forums they reported on [11,25,27]. Table 1 outlines the characteristics of the included studies.

Table 1. Characteristics of included studies.

Authors (date)	Country	Source description of web-based forum	Primary substance or behavior	Reported purpose of forum	Forum moderation	Sample size
Bergman et al (2017) [23]	United States	In The Rooms, a recovery social network site with 430,000 registered users	Substance use	For people in or seeking substance use recovery, including recovery resources and recovery-based discussion boards	Moderated	123 participants
Black et al (2020) [28]	Australia	Hello Sunday Morning (HSM), free website and app with over 40,000 participants	Alcohol	A community for facilitating action and reflection, at individual and collective levels, rather than information provision	Not specified	24 participants
Bradley and James (2021) [26]	United Kingdom	"My Journal" on Gambling Therapy website	Gambling	For people to post about life before, during, and after gambling problems	Moderated	First posts from 2298 threads
Carah et al (2017) [29]	Australia	HSM (refer to Black, et al [28])	Alcohol	As above [28]	Not specified	13,878 blog posts from 7890 registered users
Chambers et al (2017) [31]	United Kingdom	Soberistas, web-based mutual aid group with 1828 paid members and 2000 browsers	Alcohol	For people trying to resolve their problematic drinking patterns	Moderated	31 participants
Colditz et al (2023) [32]	No specific country	r/StopDrinking, subreddit	Alcohol	For people who are trying to abstain from alcohol use	Moderated	1460 direct responses to posts
Kirkman et al (2018) [30]	Australia	HSM (refer to Black, et al [28])	Alcohol	As above [28]	Not specified	1917 participants
Lee and Cooper (2019) [24]	United Kingdom	Mumsnet social media forum	Codeine	A parenting website with alcohol and other drug sections for people seeking support	Moderated	25 threads comprising 757 individual posts
Liang et al (2021) [25]	United States	Unnamed web-based health community with a long-standing history and active user participation	Opioids during pregnancy	Web-based health community	Not specified	200 posts
Lyytikäinen (2016) [33]	Russia	Alcoholics Anonymous, "newcomers" subsection of an asynchronous web-based forum comprising over 10,000 posts	Alcohol	For people to discuss problems related to drinking and their recovery journeys	Not specified	10 most recent threads, including 617 posts by more than 35 members
Mudry and Strong (2013) [27]	Canada	Unnamed web-based support forum of 3253 members	Gambling	Support for people concerned about problem gambling	Moderated	1791 posts of 11 members
Sanger et al (2019) [11]	United Kingdom and United States	Five unnamed web-based support groups ranging in size	Alcohol	Support groups using non–12-step philosophies	Moderated	25 participants
Schwebel and Orban (2022) [22]	United States	Harm reduction, Abstinence, and Moderation Support (HAMS), a private forum-based and social media support group	Substance use	Originally a support group for people to change their alcohol use, expanded to other substances	Moderated	343 participants
Sinclair et al (2017) [12]	United Kingdom	Soberistas [31]	Alcohol	As above [31]	Moderated	432 participants

What Are the Characteristics of People Who Use Peer-Based Web-Based Forums?

Reporting of the characteristics of web-based peer-supported forum users in the included studies varied. Due to the use of either anonymous posts or web-scraping for data collection, the authors of 5 of the studies provided no demographic information [24,26,27,29,32]. Results from studies that did report demographic data (n=9 [11,12,22,23,25,28,30,31,33]) suggest forum users were typically women, aged between 40 years and early 50 years (reported median age range 41.5-50.8 years). In the 9 papers that reported on gender [11,12,22,23,25,28,30,31,33], more than half of the sampled forum users were women: this ranged from 57% women [23] to 94% women [12]. In studies where information on education or socioeconomic status was reported (n=3 [11,12,28]), most forum users held higher education qualifications or were classified as being of moderate to high socioeconomic status. In 3 studies, the authors reported on race [11,22,23], and the majority of participants were reported to be White (>90%).

Current substance use of sampled web-based peer-supported forum users was reported in 6 studies [12,23,25,28,30,31]; however, the information provided varied. Several authors reported measures that indicated high levels of current use [25,28,30]. For example, in 2 studies reporting on the HSM forum, 42% [28] and 54% [30] of forum users' Alcohol Use Disorders Identification Test [34] responses indicated high-risk or dependent alcohol use at the time of data collection. Similarly, Liang et al [25] reported that 75% of forum users met one or more criteria for an opioid use disorder.

In 3 studies [12,23,31], the authors did not report the severity of substance use problems of web-based peer-supported forum users, instead providing abstinence-related data. The proportion of abstinent and nonabstinent forum users across these 3 papers varied widely. According to Sinclair et al [12], 53% of forum users were currently drinking alcohol or had consumed alcohol in the last month and 18% were abstinent for a year or more. Chambers et al [31] reported that only 23% were current or recent (<1 month) alcohol consumers and 39% were abstinent for a year or more. Bergman et al [23] reported that the majority of forum users were abstinent for a year or more (65%), and only 13% were current or recent alcohol consumers. The forum featured in Bergman et al [23], "InTheRooms" (ITR) comprised a particularly large proportion of long-term abstinent forum users; 21% were abstinent for 5-9 years and 26% were abstinent for 10 or more years. The average duration of continuous abstinence for forum users in the study by Bergman et al [23] was 7.3 (SD 9.3) years.

How Do People Participate in Peer-Based Web-Based Forums?

Participation in the web-based peer-supported forums, in terms of patterns of use and information sought, varied between the 14 studies. Some authors described forum participation quantitatively such as the number of users (n=7 [12,25-27,29,30,33]), number of posts (n=7 [24-27,29,32,33]), length of time per visit (n=3 [22,23,30]), and frequency of use (n=2 [22,23]).

In 11 studies, how web-based peer-supported forum users participated was also reported qualitatively. These qualitative features are presented in Table 2. Web-based forum users participated in a range of ways, by sharing information and personal experiences, and seeking support (Table 2).

Table 2. Qualitative data reported in the included studies.

Authors (year)	Name of the web-based forum	Data analysis methods reported	Qualitative data reported
Black et al (2020) [28]	Hello Sunday Morning (HSM; alcohol)	Thematic analysis of semistructured interviews	HSM attracted people actively seeking help and people not yet seeking help. Forum users viewed HSM in a positive, nonthreatening manner and liked the anonymity and convenience of the mobile format. They joined out of curiosity or desire for a challenge. Forum users liked the support and normalization of experiences through the blogging feature. They were motivated by the goal-setting and self-monitoring components enabled by challenges and weekly check-ins.
Carah et al (2017) [29]	HSM	Text analysis of blog posts, grouping expressions together as related concepts	Forum users' expressions changed over time. In the first month, they set goals, and described current drinking practices, hopes and anxieties, and early efforts to change. After the first month, forum users reported on change efforts and challenges, and reflected on their place in a drinking culture. They evaluated their efforts to change and presented "findings" and "theorised" them to advise others.
Chambers et al (2017) [31]	Soberistas (alcohol)	Grounded theory techniques to analyze in-depth interviews	Key stages of engagement, through which forum users' identities were constructed and adjusted to support recovery. The most linear and commonly discussed engagement involved transitions through "lurking," "participating," "leading," then "moving on"; coinciding with forum users' journey from problematic use to "secure in sobriety."
Colditz et al (2023) [32]	r/StopDrinking (alcohol)	Constant comparative method	Emotional support included expressions of encouragement, emotional alignment, sympathy, or empathy, in relation to an original post, sobriety-related accomplishment, challenge, or some other quality of the narrative. Appraisal support normalized the original poster's experiences, negatively appraised drinking behavior, and positively appraised recovery behaviors and outcomes. Informational support included fact-based information, informed opinions, and instructions.
Lee and Cooper (2019) [24]	Mumsnet (codeine)	Thematic analysis of threads	Forum users created posts to request help in relation to usually, but occasionally their relative's, problems with codeine use and self-reported addiction. Positive and negative descriptions of side effects, problems experiencing withdrawal, and failed attempts to discontinue use were reported. Advice was provided about formal health services or informal approaches, and often anecdotal advice about how to taper or use cold turkey techniques. Arguments and challenges to advice were not uncommon. Shame and stigma were often associated with posts and forum users often wanted to keep codeine use hidden in their lives.
Liang et al (2021) [25]	Unnamed (opioids during pregnancy)	Thematic analysis of posts	The following six themes highlighted self-management support needs: (1) clarity on the impact on pregnancy, (2) clinically validated information on how to reduce dosage, (3) guidelines on safe pain management during pregnancy, (4) information on local child protection procedures, (5) strategies for obtaining support from offline systems, and (6) emotional support for those experiencing negative emotions.
Lyytikäinen (2016) [33]	AA[a]	Content analysis of posts	Forum users gave each other mutual support in going through phases of change. Many started to adopt the philosophy of AA, model the AA life story, and acquire new self-understanding of a sick person with a chronic disease. By engaging, forum users acquired a sense of agency, and being in charge of their lives. The forum created a web-based space where users collectively acted according to AA values, which supported them to do so offline as well.
Mudry and Strong (2013) [27]	Unnamed free support forum (gambling)	Discourse analysis of posts	The following six common discourses were used in the forum: (1) shame and guilt, (2) causality, (3) nature of gambling, (4) gambling as an addiction or illness, (5) control and responsibility, and (6) recovery as a process.
Sanger et al (2019) [11]	Five unnamed web-based support groups (alcohol)	Thematic analysis of semistructured interviews	Most important benefit of groups was finding "someone like me." Forums provided support without requiring users to follow a set program for recovery. Forum users respected others' rights to choose their goal for sobriety and how they achieved it.
Sinclair et al (2017) [12]	Soberistas (alcohol)	Coding and summarizing of free-text survey responses	Anonymity, ability to be honest, source of trusted information, and ongoing support were reasons for continued membership.

[a]AA: Alcoholics Anonymous.

Across the 11 studies reporting on these qualitative features of participation, web-based peer-supported forums were a source of information and advice for people in similar circumstances. For example, Liang et al [25] reported that forum users were seeking information on opioid use during pregnancy and advice on self-managing their use. Forum users obtained information about the adverse effects of opioid use during pregnancy, self-managed withdrawal, continued safe use, and child protection and health systems. Similarly, Sinclair et al [12] examined the Soberistas forum where users shared information on alcohol, health, and well-being. Information sharing was reported as useful for all forum users, even passive ones. Chambers et al [31] found "lurking" behavior was common in the Soberistas forum and reported that users who "lurk" without posting were still able to gain important information. However, web-based peer-supported forums could also contain potentially inaccurate advice or misinformation. For example, Lee and Cooper [24] reported that when one forum user suggested using cannabis to help with codeine addiction, this comment was swiftly negated by another user who posted, "be careful with rubbish advice." While 9 of the 14 papers included forums with moderators, it was unclear how active moderators were regarding posts involving misinformation. Moderator duties were described as providing content, feedback, or support [22,26,27] or removing spam posts or posts that violated community guidelines [12,23,26,27,31].

Many web-based peer-supported forum users also sought emotional support. Lyytikäinen observed that Russian Alcoholics Anonymous forum users made posts that described their current situation and asked for support, and other forum users then offered guidance or motivation (eg, "Stay and recover with us") [33]. In other studies, forum users shared their personal circumstances. For example, from Lee and Cooper's study: "My partner has been getting lots and lots of codeine in over-the-counter Nurofen Plus over the years by going to different pharmacies" [24]. Mudry and Strong [27] found that senior forum users shared past experiences of recovery to support others, and this sharing also established their seniority, authority, and legitimacy within the group.

What Are the Key Principles Reportedly Underpinning Peer-Based Web-Based Forums?

In 11 of the 14 included studies, the authors reported key principles underpinning the web-based peer-supported forums (Table 3). Behavior change models appeared to underpin the HSM forums [28-30]. Bergman et al [23] and Schwebel and Orban [22] referred to the Social Identity Model of Recovery, as described by Best et al [35]. Carah et al's [29] study of a gambling-focused web-based forum referred to the importance of relationships to help control or stop gambling. Web-based peer-supported forums reportedly founded through mutual aid included the Soberistas forum, described as nonprescriptive, nonjudgmental, and nonreligious [12,31] and the Russian Alcoholics Anonymous forum [33]. Liang et al [25] focused on self-management support, and Mudry and Strong [27] referenced learning and support through others in a community of practice model [36] although the name of the web-based forum was not reported in both studies. Finally, data were analyzed in the subreddit StopDrinking reported by Colditz et al [32] according to House's [37] conceptual model of social support.

Table 3. Key principles of web-based peer-supported forums reported in the primary studies.

Name of web-based forum	Authors (date)	Key forum principles or approaches
Alcoholics Anonymous (AA)	Lyytikäinen (2016) [33]	Phase model of therapeutic change embedded within AA movement.
Gambling Therapy: "My Journal"	Bradley and James (2020) [26]	Forum as a beneficial source of support, fostering growth of relationships to help attempts to control or stop gambling.
Harm reduction, Abstinence, and Moderation Support (HAMS)	Schwebel and Orban (2023) [22]	Recovery through exposure to relatable role models, enhancing motivation and strengthening identity and self-efficacy.
Hello Sunday Morning (HSM)	Carah et al (2017) [2]; Black et al (2020) [28]; Kirkman et al (2018) [30]	Core principles of HSMs behavioral change model are mindfulness and community, to promote reflection on the user's relationship with alcohol.
r/StopDrinking	Colditz et al (2023) [32]	Social support: emotional, appraisal, informational, instrumental
In The Rooms	Bergman et al (2017) [23]	Recovery through exposure to relatable role models, enhancing motivation and strengthening identity and self-efficacy. Most resources were grounded in 12-step mutual help philosophy.
Soberistas	Sinclair et al (2017) [12]; Chambers et al (2017) [31]	Web-based mutual aid that is nonprescriptive, nonreligious, and nonjudgmental.
Unnamed	Liang et al (2021) [25]	Self-management support for opioid use during pregnancy.
Unnamed	Mudry and Strong (2013) [27]	Community of practice.

What Are the Reported Outcomes of Web-Based Peer-Supported Forums?

In 4 of the 14 studies, authors used quantitative methods to report on outcomes related to the use of web-based peer-supported forums. The outcomes reported related to participation, perceived benefits of participation, and alcohol consumption.

Colditz et al [32] used mixed methods to characterize the social support provided on a StopDrinking recovery forum hosted on Reddit. Qualitative content analysis of 1386 responses to posts was undertaken to identify the type of social support provided:

emotional, appraisal, or informational. The linguistic characteristics of these responses were quantified based on text length, complexity, and sentiment variables. Emotional support was coded as most common, and these responses were significantly shorter, less complex, and more positive than responses without emotional support, indicating that this type of response was a quick and easy way to exchange support among participants who could benefit from brief encouragement.

Bergman et al [23] examined the participation in ITR, primarily for people in, or working toward recovery, with a focus on abstinence, and surveyed 123 ITR users to examine their participation and perceived benefits of participation [23]. Participation was measured using ordinal scales to assess past-90-day ITR log-in frequency and intensity. The ITR users also reported their level of agreement with four statements on perceived benefit from participation: (1) enhanced recovery motivation, (2) enhanced recovery self-efficacy, (3) reduced craving, and (4) strengthened recovery identity. The ITR users engaged on average 30 minutes per day several times each week. Engagement was generally endorsed as helpful, particularly with respect to increased abstinence or recovery motivation and self-efficacy. Compared with ITR users who reported being abstinent for at least 1 year, those abstinent for less than 1 year showed similar rates of engagement with activities and similar levels of perceived benefit.

Schwebel and Orban [22] extended the study by Bergman et al [23] by examining the participation in a private, forum-based support group for changing alcohol use (Harm Reduction, Abstinence, and Moderation Support; HAMS), while not with an abstinence focus. For this study, 343 HAMS users were surveyed to examine their participation and perceived benefits of participation, measured through a modified version of Bergman et al [23] scales for participation. Perceived benefit was measured on a 5-point response scale assessing changing substance use behavior, craving, substance use behavior change self-efficacy, and substance use change motivation. The forum users most frequently reported visiting HAMS via Facebook daily, and up to 30 minutes per day. Most users somewhat or strongly agreed that HAMS helped them feel better about changing use, increased motivation for changing use, and increased self-efficacy for reaching or maintaining their substance use goals.

Kirkman et al [30] reported on the registration data of 1917 HSM users who signed up for 3 months of abstinence. To determine whether alcohol consumption changes were associated with participation, users completed the Alcohol Use Disorders Identification Test [34] at baseline and completion of the 3-month period. The HSM users who reported hazardous and harmful consumption levels, and engaged in the program, reported a significant decrease in alcohol consumption, achieving low-risk consumption levels 4 months after starting the program. Those who reported high-risk or dependent consumption levels before HSM engagement experienced the biggest reduction. These reductions in risk were maintained by forum users 7 months after starting HSM.

Discussion

Principal Findings

We identified 14 primary studies reporting on the use of web-based peer support for people experiencing addiction, for a variety of substances and behaviors, published between 2013 and 2023. While not all studies reported demographic data, of the 9 that did, the majority of web-based forum users were reported as middle-aged women participating in forums focusing on alcohol use. Participation in web-based peer-supported forums was reported both quantitatively (eg, number of forum users, length of time per visit, number of posts) and qualitatively (eg, information sharing, seeking support, sharing experiences). The web-based peer-supported forums were reportedly underpinned by a range of key principles, mostly mutual help approaches and recovery identity formation. Only 4 included studies reported on outcomes for forum users; however, these studies were observational: it is not possible to draw conclusions about the impact of participation in forums.

Web-based peer support can help to overcome barriers to attending traditional, face-to-face forums such as 12-step meetings, by enabling people experiencing addiction improved access to seek support beyond their physical location and with the benefit of anonymity [9]. For example, for people in regional and rural locations, web-based services such as web-based counseling are increasingly being used for people who have difficulty engaging with or accessing face-to-face services [38]. These benefits may extend to web-based peer-supported forums for people experiencing addiction, as demonstrated recently in an Australian mental health web-based forum, where participants described the importance of connection through peer support [39]. Further research in the field of addiction is warranted, particularly for people who are geographically isolated.

Although two-thirds of clients receiving alcohol and other drug treatment in Australia are men [40], our finding that women (particularly those aged between 40 years and early 50 years) were disproportionately represented among forum users is consistent with broader gender trends in help-seeking and life responsibilities. Women are more likely to seek health information, and are more likely to do so via the web, compared with men [41-44]. Women also experience a range of factors that restrict their access to formal addiction support [45,46], including caring responsibilities and social stigma [47]. Our finding adds to a growing body of literature that suggests nontraditional alcohol and other drug digital services (such as telehealth, telephone-based, and web-based interventions) may be filling a service gap for women [48,49]. Women's overrepresentation may indicate that, due to the reduced cost and increased availability and accessibility of web-based peer-supported forums, receiving web-based addiction support may be more feasible than accessing traditional, offline services.

We found inconsistent reporting of web-based peer-supported forum participation and use. Forums were regularly reported as a source of information and support for people in similar circumstances. Information sharing was reported as useful for all forum users, even "passive" ones ("lurkers"). However, there was limited information distinguishing between the different ways users engaged with the forum. This nuance has been captured by other studies investigating nonaddiction forums. For example, in a recent publication on the role of group dynamics in shaping social support through web-based health communities, James et al [50] presented a model that focuses on active and passive use within such communities [50]. In their model, information sharing was a measure of active use, and information consumption was a measure of passive use. James et al [50] hypothesized that these 2 activities encompassed what people "do" in communities such as web-based forums [50]. Their emerging "web-based health community social support model" outlines how information consumption and sharing predict received social support through the forum. For administrators of web-based peer-supported forums, measuring the frequency of information consumption (eg, reading posts) and information sharing may be more useful than the number of forum users, posts, threads, and "likes."

There was a mix of key principles underpinning the web-based peer-supported forums included in this review. The most frequent approaches reported were behavioral change, social identity, and mutual aid. The Social Identity Model of Recovery describes recovery as a process of change in a person's social identity from being defined by membership of a group of people whose norms and values center on substance misuse, to membership of a group of people whose norms and values encourage recovery [35]. This socially embedded process is reflected in the included studies, particularly through the peer support element of web-based forums. The study included in this review by Colditz et al [38] highlighted the importance of emotional support to participants in an abstinence-based drinking forum, whereby although emotional support was provided in short text responses, even brief expressions of encouragement were valuable. Seeking peer support on the web, through sharing experiences and knowledge, for example, is gaining traction among people experiencing addiction [6]. This concept of sharing, particularly with people who have lived or living experience of addiction is a key principle of recovery [17]. Further research incorporating the Social Identity Model of Recovery within web-based peer-supported forums would assist in understanding how this socially embedded process occurs.

In line with other reviews of web-based support for people experiencing addiction, evidence on the benefits and effectiveness of web-based peer-supported forums remains lacking [18,51]. Face-to-face peer support remains a more evidence-based approach [52]. Recent guidelines to qualitatively analyze web-based support forums provide practical and methodological issues to consider when undertaking forum research [53]. However, there remain conceptual, theoretical, and methodological considerations, such as a lack of clarity around definitions of web-based forums (or communities), the dimensions of participation, and the need for experimental designs [18,53]. For example, questions remain in relation to when people engage with web-based peer-supported forums, for how long, and how web-based forums are positioned within and outside the treatment system. Understanding what works, for whom, how, and in what web-based contexts, requires further investigation.

Limitations

Our scoping review has some limitations. First, our analysis is limited to those web-based peer-supported forums described in the studies identified via our search strategy. This means that some web-based forums may be overrepresented (eg, HSM, Soberistas, and alcohol-focused forums), while other forums may not be reported in the literature. Our findings were also limited to papers available in English, with our resulting sample featuring forums from only 5 countries, of which the majority (4/5) were Western, high-income, and English-speaking countries (Australia, Canada, the United Kingdom, and the United States). It is, therefore, unclear whether these findings are generalizable to other countries and cultures. In addition, text in a study title or abstract may not have referred to our specific search terms, and therefore, may have been missed. Since our research questions were based on peer-reviewed literature, we did not include gray literature or conference proceedings. Additionally, following Arksey and O'Malley's [20] methodology for scoping reviews, we did not include quality appraisals as well. Papers from the computer science field are likely to be underrepresented. Due to the diverse and limited literature on this topic, we did not attempt to restrict our search to research on the effectiveness of web-based peer-supported forums. Instead, using the strengths of the scoping review approach, we brought together a heterogeneous body of literature that included descriptions and participation in web-based forums, as well as changes in substance use and other measures of effectiveness.

Conclusions

Web-based peer-supported forums are used by people experiencing addiction in a number of ways, to share information and experiences, and give and receive support. Seeking web-based support offers an alternative approach to traditional face-to-face support options, and may reduce some barriers to engaging in peer support. Further research will assist forum users and forum administrators to articulate and optimize the benefits of web-based forum participation.

Conflicts of Interest

None declared

Multimedia Appendix 1
Search string for MEDLINE.
[DOCX File , 15 KB - mental_v11i1e49010_app1.docx]

Multimedia Appendix 2
PRISMA-ScR Checklist.
[PDF File (Adobe PDF File), 101 KB - mental_v11i1e49010_app2.pdf]

References

1. Castaldelli-Maia JM, Bhugra D. Analysis of global prevalence of mental and substance use disorders within countries: focus on sociodemographic characteristics and income levels. Int Rev Psychiatry 2022;34(1):6-15 [FREE Full text] [doi: 10.1080/09540261.2022.2040450] [Medline: 35584016]
2. Hing N, Russell A, Nuske E, Gainbury S. The stigma of problem gambling: causes, characteristics and consequences. Victorian Responsible Gambling Foundation. 2015. URL: https://responsiblegambling.vic.gov.au/resources/publications/the-stigma-of-problem-gambling-causes-characteristics-and-consequences-351/ [accessed 2024-05-17]
3. Room R. Stigma, social inequality and alcohol and drug use. Drug Alcohol Rev 2005;24(2):143-155 [FREE Full text] [doi: 10.1080/09595230500102434] [Medline: 16076584]
4. Wakeman SE, Rich JD. Barriers to medications for addiction treatment: how stigma kills. Subst Use Misuse 2018;53(2):330-333 [FREE Full text] [doi: 10.1080/10826084.2017.1363238] [Medline: 28961017]
5. Eddie D, Hoffman L, Vilsaint C, Abry A, Bergman B, Hoeppner B, et al. Lived experience in new models of care for substance use disorder: a systematic review of peer recovery support services and recovery coaching. Front Psychol 2019;10:1052 [FREE Full text] [doi: 10.3389/fpsyg.2019.01052] [Medline: 31263434]
6. Manning V, Savic M, Lubman D. Peer support programs. In: Haber PS, Riordan BC, editors. Guidelines for the Treatment of Alcohol Problems. 4th Edition. Sydney: The University of Sydney: Specialty of Addiction Medicine, Faculty of Medicine and Health; 2021:173.
7. Tracy K, Wallace SP. Benefits of peer support groups in the treatment of addiction. Subst Abuse Rehabil 2016;7:143-154 [FREE Full text] [doi: 10.2147/SAR.S81535] [Medline: 27729825]
8. White WL. Peer-based addiction recovery support: history, theory, practice, and scientific evaluation. Great Lakes Addiction Technology Transfer Center and Philadelphia Department of Behavioral Health and Mental Retardation Services. 2009. URL: https://www.naadac.org/assets/2416/whitew2009_peer-based_addiction_recovery_support.pdf [accessed 2024-05-17]
9. Understanding the cost of addiction in Australia. Rethink Addiction. 2022. URL: https://www.rethinkaddiction.org.au/the-cost-of-addiction [accessed 2024-05-17]
10. Dilkes-Frayne E, Savic M, Carter A, Kokanović R, Lubman DI. Going online: the affordances of online counseling for families affected by alcohol and other drug issues. Qual Health Res 2019;29(14):2010-2022 [FREE Full text] [doi: 10.1177/1049732319838231] [Medline: 30931830]
11. Sanger S, Bath PA, Bates J. 'Someone like me': user experiences of the discussion forums of non-12-step alcohol online support groups, June 2019. Addict Behav 2019;98:106028 [FREE Full text] [doi: 10.1016/j.addbeh.2019.106028] [Medline: 31302313]
12. Sinclair JMA, Chambers SE, Manson CC. Internet support for dealing with problematic alcohol use: a survey of the soberistas online community. Alcohol Alcohol 2017;52(2):220-226 [FREE Full text] [doi: 10.1093/alcalc/agw078] [Medline: 28182197]
13. Erfani SS, Blount Y, Abedin B. The influence of health-specific social network site use on the psychological well-being of cancer-affected people. J Am Med Inform Assoc 2016;23(3):467-476 [FREE Full text] [doi: 10.1093/jamia/ocv170] [Medline: 26911816]
14. Litchman ML, Rothwell E, Edelman LS. The diabetes online community: older adults supporting self-care through peer health. Patient Educ Couns 2018;101(3):518-523 [FREE Full text] [doi: 10.1016/j.pec.2017.08.023] [Medline: 28947360]
15. Smith-Merry J, Goggin G, Campbell A, McKenzie K, Ridout B, Baylosis C. Social connection and online engagement: insights from interviews with users of a mental health online forum. JMIR Ment Health 2019;6(3):e11084 [FREE Full text] [doi: 10.2196/11084] [Medline: 30912760]
16. Coulson NS. Sharing, supporting and sobriety: a qualitative analysis of messages posted to alcohol-related online discussion forums in the United Kingdom. J Subst Use 2013;19(1-2):176-180 [FREE Full text] [doi: 10.3109/14659891.2013.765516]
17. Best DW, Lubman DI. The recovery paradigm - a model of hope and change for alcohol and drug addiction. Aust Fam Physician 2012;41(8):593-597. [Medline: 23145400]
18. Ashford RD, Bergman BG, Kelly JF, Curtis B. Systematic review: digital recovery support services used to support substance use disorder recovery. Human Behav and Emerg Tech 2019;2(1):18-32 [FREE Full text] [doi: 10.1002/hbe2.148]
19. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
20. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005;8(1):19-32 [FREE Full text] [doi: 10.1080/1364557032000119616]
21. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]

22. Schwebel FJ, Orban DG. Online support for all: examining participant characteristics, engagement, and perceived benefits of an online harm reduction, abstinence, and moderation focused support group for alcohol and other drugs. Psychol Addict Behav 2023;37(2):228-234 [FREE Full text] [doi: 10.1037/adb0000828] [Medline: 35266792]
23. Bergman BG, Kelly NW, Hoeppner BB, Vilsaint CL, Kelly JF. Digital recovery management: characterizing recovery-specific social network site participation and perceived benefit. Psychol Addict Behav 2017;31(4):506-512. [doi: 10.1037/adb0000255] [Medline: 28206780]
24. Lee E, Cooper RJ. Codeine addiction and internet forum use and support: qualitative netnographic study. JMIR Ment Health 2019;6(4):e12354 [FREE Full text] [doi: 10.2196/12354] [Medline: 31021328]
25. Liang OS, Chen Y, Bennett DS, Yang CC. Identifying self-management support needs for pregnant women with opioid misuse in online health communities: mixed methods analysis of web posts. J Med Internet Res 2021;23(2):e18296 [FREE Full text] [doi: 10.2196/18296] [Medline: 33538695]
26. Bradley A, James RJE. Defining the key issues discussed by problematic gamblers on web-based forums: a data-driven approach. Int Gambl Stud 2020;21(1):59-73 [FREE Full text] [doi: 10.1080/14459795.2020.1801793]
27. Mudry TE, Strong T. Doing recovery online. Qual Health Res 2013;23(3):313-325 [FREE Full text] [doi: 10.1177/1049732312468296] [Medline: 23208202]
28. Black N, Loomes M, Juraskova I, Johnston I. Engagement in a novel internet intervention for alcohol reduction: a qualitative study of user motivations and experiences. Cyberpsychol Behav Soc Netw 2020;23(4):225-233 [FREE Full text] [doi: 10.1089/cyber.2019.0289] [Medline: 32083488]
29. Carah N, Meurk C, Angus D. Online self-expression and experimentation as 'reflectivism': using text analytics to examine the participatory forum hello sunday morning. Health (London) 2017;21(2):119-135 [FREE Full text] [doi: 10.1177/1363459315596799] [Medline: 26216897]
30. Kirkman JJL, Leo B, Moore JC. Alcohol consumption reduction among a web-based supportive community using the hello sunday morning blog platform: observational study. J Med Internet Res 2018;20(5):e196 [FREE Full text] [doi: 10.2196/jmir.9605] [Medline: 29773530]
31. Chambers SE, Canvin K, Baldwin DS, Sinclair JMA. Identity in recovery from problematic alcohol use: a qualitative study of online mutual aid. Drug Alcohol Depend 2017;174:17-22 [FREE Full text] [doi: 10.1016/j.drugalcdep.2017.01.009] [Medline: 28282522]
32. Colditz JB, Chu KH, Hsiao L, Barrett E, Kraemer KL, Pedersen SL. Characterizing online social support for alcohol use disorder: a mixed-methods approach. Alcohol Clin Exp Res (Hoboken) 2023;47(11):2110-2120 [FREE Full text] [doi: 10.1111/acer.15187] [Medline: 38226760]
33. Lyytikäinen L. Mutual support and recovery in the Russian alcoholics anonymous online community. Nord Stud Alcohol Drugs 2017;33(2):151-172 [FREE Full text] [doi: 10.1515/nsad-2016-0013]
34. Saunders JB, Aasland OG, Babor TF, de la Fuente JR, Grant M. Development of the Alcohol Use Disorders Identification Test (AUDIT): who collaborative project on early detection of persons with harmful alcohol consumption—II. Addiction 1993;88(6):791-804 [FREE Full text] [doi: 10.1111/j.1360-0443.1993.tb02093.x] [Medline: 8329970]
35. Best D, Beckwith M, Haslam C, Haslam SA, Jetten J, Mawson E, et al. Overcoming alcohol and other drug addiction as a process of social identity transition: the social identity model of recovery (SIMOR). Addict Res Theory 2015;24(2):111-123 [FREE Full text] [doi: 10.3109/16066359.2015.1075980]
36. Wenger E. Knowledge management as a doughnut: shaping your knowledge strategy through communities of practice. Ivey Bus J 2004;68:1-8 [FREE Full text]
37. House JS. Work Stress and Social Support. Boston MA: Addison-Wesley Publishing Company; 1981.
38. Haylock PAC, Carter A, Savic M, Lubman DI. Regional and rural clients' presenting concerns and experiences of care when engaging with an online substance use counseling service. Addict Res Theory 2022;30(5):330-339 [FREE Full text] [doi: 10.1080/16066359.2022.2039911]
39. Smith-Merry J, Goggin G, Campbell A, McKenzie K, Ridout B, Baylosis C. Social connection and online engagement: insights from interviews with users of a mental health online forum. JMIR Ment Health 2019;6(3):e11084 [FREE Full text] [doi: 10.2196/11084] [Medline: 30912760]
40. Alcohol and other drug treatment services in Australia annual report. Australian Institute of Health and Welfare. Canberra, Australia: Australian Institute of Health and Welfare; 2021. URL: https://www.aihw.gov.au/reports/alcohol-other-drug-treatment-services/alcohol-other-drug-treatment-services-australia/contents/about [accessed 2024-05-17]
41. Garde EL, Manning V, Lubman DI. Characteristics of clients currently accessing a national online alcohol and drug counselling service. Australas Psychiatry 2017;25(3):250-253 [FREE Full text] [doi: 10.1177/1039856216689623] [Medline: 28541729]
42. Hallyburton A, Evarts LA. Gender and online health information seeking: a five survey meta-analysis. J Consum Health Internet 2014;18(2):128-142 [FREE Full text] [doi: 10.1080/15398285.2014.902268]
43. Nikoloudakis IA, Vandelanotte C, Rebar AL, Schoeppe S, Alley S, Duncan MJ, et al. Examining the correlates of online health information-seeking behavior among men compared with women. Am J Mens Health 2018;12(5):1358-1367 [FREE Full text] [doi: 10.1177/1557988316650625] [Medline: 27193765]

44. Stern MJ, Cotten SR, Drentea P. The separate spheres of online health. J Fam Iss 2011;33(10):1324-1350 [FREE Full text] [doi: 10.1177/0192513x11425459]
45. Barnett ER, Knight E, Herman RJ, Amarakaran K, Jankowski MK. Difficult binds: a systematic review of facilitators and barriers to treatment among mothers with substance use disorders. J Subst Abuse Treat 2021;126:108341 [FREE Full text] [doi: 10.1016/j.jsat.2021.108341] [Medline: 34116826]
46. Rizzo D, Mu T, Cotroneo S, Arunogiri S. Barriers to accessing addiction treatment for women at risk of homelessness. Front Glob Womens Health 2022;3:795532 [FREE Full text] [doi: 10.3389/fgwh.2022.795532] [Medline: 35252964]
47. Nichols TR, Welborn A, Gringle MR, Lee A. Social stigma and perinatal substance use services: recognizing the power of the good mother ideal. Cont Drug Prob 2020;48(1):19-37 [FREE Full text] [doi: 10.1177/0091450920969200]
48. Grigg J, Volpe I, Tyler J, Hall K, McPherson B, Lubman DI, et al. Ready2Change: preliminary effectiveness of a telephone-delivered intervention program for alcohol, methamphetamine and cannabis use problems. Drug Alcohol Rev 2022;41(2):517-527 [FREE Full text] [doi: 10.1111/dar.13363] [Medline: 34343370]
49. Grigg J, Manning V, Cheetham A, Youssef G, Hall K, Baker AL, et al. A latent class analysis of perceived barriers to help-seeking among people with alcohol use problems presenting for telephone-delivered treatment. Alcohol Alcohol 2023;58(1):68-75 [FREE Full text] [doi: 10.1093/alcalc/agac063] [Medline: 36448844]
50. James T, Calderon EDV, Bélanger F, Lowry PB. The mediating role of group dynamics in shaping received social support from active and passive use in online health communities. Info Manage 2022;59(3):103606 [FREE Full text] [doi: 10.1016/j.im.2022.103606]
51. Malinen S. Understanding user participation in online communities: a systematic literature review of empirical studies. Comp Human Behav 2015;46(1):228-238 [FREE Full text] [doi: 10.1016/j.chb.2015.01.004]
52. Kelly JF, Humphreys K, Ferri M. Alcoholics anonymous and other 12-step programs for alcohol use disorder. Cochrane Database Syst Rev 2020;3(3):CD012880 [FREE Full text] [doi: 10.1002/14651858.CD012880.pub2] [Medline: 32159228]
53. Smedley RM, Coulson NS. A practical guide to analysing online support forums. Qual Res Psychol 2018;18(1):76-103 [FREE Full text] [doi: 10.1080/14780887.2018.1475532]

Abbreviations

D-RSS: digital recovery support service
HAMS: Harm reduction, Abstinence, and Moderation Support
HSM: Hello Sunday Morning
ITR: InTheRooms

Edited by J Torous, YH Lin; submitted 15.05.23; peer-reviewed by K Durant, J Colditz, MP Gagnon; comments to author 26.01.24; revised version received 08.03.24; accepted 26.04.24; published 17.06.24.

Please cite as:
Peart A, Horn F, Petukhova R, Barnett A, Lubman DI
Web-Based Forums for People Experiencing Substance Use or Gambling Disorders: Scoping Review
JMIR Ment Health 2024;11:e49010
URL: https://mental.jmir.org/2024/1/e49010
doi:10.2196/49010
PMID:38885012

Review

Behavior Change Techniques Within Digital Interventions for the Treatment of Eating Disorders: Systematic Review and Meta-Analysis

Pamela Carien Thomas[1], BA, MSc; Kristina Curtis[2], BA, MSc, PhD; Henry W W Potts[3], BA, MSc, PhD; Pippa Bark[4], BSc; Rachel Perowne[2], BSc, MSc; Tasmin Rookes[5], BSc, MSc; Sarah Rowe[1], BA, MEd, PhD

[1]Department of Epidemiology & Applied Clinical Research, Division of Psychiatry, University College London, London, United Kingdom

[2]Department of Clinical, Educational and Health Psychology, University College London, London, United Kingdom

[3]UCL Institute of Health Informatics, University College London, London, United Kingdom

[4]UCL Cancer Institute, University College London, London, United Kingdom

[5]UCL Research Department of Primary Care and Population Health, University College London, London, United Kingdom

Corresponding Author:
Sarah Rowe, BA, MEd, PhD
Department of Epidemiology & Applied Clinical Research
Division of Psychiatry
University College London
6th Floor, Wings A and B, Maple House
149 Tottenham Court Road
London, W1T 7NF
United Kingdom
Phone: 44 0207 679 9000
Email: s.rowe@ucl.ac.uk

Abstract

Background: Previous systematic reviews of digital eating disorder interventions have demonstrated effectiveness at improving symptoms of eating disorders; however, our understanding of how these interventions work and what contributes to their effectiveness is limited. Understanding the behavior change techniques (BCTs) that are most commonly included within effective interventions may provide valuable information for researchers and developers. Establishing whether these techniques have been informed by theory will identify whether they target those mechanisms of action that have been identified as core to changing eating disorder behaviors. It will also evaluate the importance of a theoretical approach to digital intervention design.

Objective: This study aims to define the BCTs within digital self-management interventions or minimally guided self-help interventions for adults with eating disorders that have been evaluated within randomized controlled trials. It also assessed which of the digital interventions were grounded in theory and the range of modes of delivery included.

Methods: A literature search identified randomized controlled trials of digital intervention for the treatment of adults with eating disorders with minimal therapist support. Each digital intervention was coded for BCTs using the established BCT Taxonomy v1; for the application of theory using an adapted version of the theory coding scheme (TCS); and for modes of delivery using the Mode of Delivery Ontology. A meta-analysis evaluated the evidence that any individual BCT moderated effect size or that other potential factors such as the application of theory or number of modes of delivery had an effect on eating disorder outcomes.

Results: Digital interventions included an average of 14 (SD 2.6; range 9-18) BCTs. *Self-monitoring of behavior* was included in all effective interventions, with *Problem-solving*, *Information about antecedents*, *Feedback on behavior*, *Self-monitoring of outcomes of behavior*, and *Action planning* identified in >75% (13/17) of effective interventions. *Social support* and *Information about health consequences* were more evident in effective interventions at follow-up compared with postintervention measurement. The mean number of modes of delivery was 4 (SD 1.6; range 2-7) out of 12 possible modes, with most interventions (15/17, 88%) being web based. Digital interventions that had a higher score on the TCS had a greater effect size than those with a lower TCS score (subgroup differences: χ^2_1=9.7; *P*=.002; I^2=89.7%) within the meta-analysis. No other subgroup analyses had statistically significant results.

Conclusions: There was a high level of consistency in terms of the most common BCTs within effective interventions; however, there was no evidence that any specific BCT contributed to intervention efficacy. The interventions that were more strongly informed by theory demonstrated greater improvements in eating disorder outcomes compared to waitlist or treatment-as-usual controls. These results can be used to inform the development of future digital eating disorder interventions.

Trial Registration: PROSPERO CRD42023410060; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=410060

(JMIR Ment Health 2024;11:e57577) doi:10.2196/57577

KEYWORDS

digital health; eHealth; mobile health; mHealth; mobile apps; smartphone; behavior change; behavior change technique; systematic review; eating disorders; disordered eating; binge eating; bulimia nervosa; mobile phone

Introduction

Background

The current eating disorder (ED) treatment model is falling short for patients [1], with a significant majority of people with EDs failing to get help [2]. This may be due to limited access to services [3] and the stigma and shame associated with their condition [4]. EDs have the highest mortality of any psychiatric disorder [5], and they may be long-lasting and may cause physical, emotional, and neurobiological damage if left untreated [6]. The COVID-19 pandemic has further compounded the problem, with a surge in urgent referrals and increased waiting time in an already underresourced system [7]. Action is urgently required to address this treatment gap [8]. A promising strategy that can improve access to evidence-based treatments is the development and implementation of digital interventions. Digital interventions refer to the use of digital technologies, such as mobile apps, websites, or virtual reality, to deliver health care or behavioral interventions.

Advantages of digital interventions include the ability to reach many people at minimal or no additional cost per person, and they can be used at an individual's convenience, at home, anonymously, and at a self-suited pace [9]. Shame and stigma may make people with EDs more likely to engage in digital interventions to achieve improvements in their symptoms [10,11], and evidence demonstrates that the demand for self-guided digital interventions is growing among people with EDs [11]. While digital self-management interventions are not the only solution to address the existing service gap, they can broaden the dissemination of evidence-based treatments and help more people get support for their condition [12].

Digital interventions for EDs have shown promising evidence in treating ED symptoms [13-15] with results sustained, or even improved, at follow-up [16]. However, our understanding of how these interventions work and what contributes to their effectiveness is limited [17], restricting the potential effectiveness and impact of digital ED interventions. It is widely recognized that digital health interventions should incorporate evidence-based methods and behavior change theory into their development [18]. Theory represents the accumulated knowledge of the mechanisms of action (MOAs; mediators) and moderators of change as well as the a priori assumptions about what human behavior is and what the influences on it are [19]. Using behavior change theory in designing digital health interventions may help pinpoint the factors influencing the target behavior, referred to as *MOAs* in behavioral science. These MOAs, such as knowledge and beliefs, are pathways through which interventions can impact behaviors. Designers can then connect these MOAs to practical elements called "behavior change techniques" (BCTs), which play a crucial role in transforming disordered behaviors into healthier target behaviors. While there are some dissenters regarding such systematization of practice, arguing for the importance of variability, there is general agreement of the value of better descriptions of interventions for clarity and replication [20]. This systematic approach has been applied in the development of effective digital health interventions in areas such as the treatment of addictive disorders, physical activity, and weight loss [21,22], as well as in more clinically oriented interventions, such as diabetes management [23-25]. Specific BCTs have been linked to improved clinical outcomes [26-28] and are a useful means of describing active components within complex digital interventions [29]. The integration of specific BCTs may optimize digital ED treatment interventions, helping achieve significant symptom improvement by addressing those factors (eg, food avoidance, dietary restriction, and body image concerns) that influence common ED behaviors (eg, bingeing and purging).

Objectives

This review aimed to gain insights from previous randomized controlled trials (RCTs) as to which BCTs may contribute to the effectiveness of digital ED interventions [30]. It focused on RCTs as they have the highest possible level of evidence compared to other study designs and can be used to make causal inferences [31]. It also assessed whether the interventions were grounded in theory, given that theory is a "necessary precursor to the development of effective interventions" [32].

We hypothesized that interventions that specifically targeted the behavioral and psychological aspects of ED via the use of relevant BCTs would be more likely to improve ED outcomes. We also hypothesized that the interventions informed by theory were more likely to be effective. Having multiple modes of delivery (eg, apps, video, and audio) may be associated with enhanced treatment outcomes [15] based on the idea that the diversity offered by multimedia formats might facilitate effectiveness through an enhanced and more engaging user experience [33].

Our specific research questions were as follows:

- Which BCTs are most frequently included in digital interventions for the treatment of EDs that have been

evaluated in RCTs? Which BCTs are most frequently associated with effective interventions?
- Are included BCTs informed by theory?
- Which modes of delivery have been adopted to deliver the BCTs?
- Was there evidence to suggest that specific BCTs, or related factors, moderated the intervention effect size?

Methods

Search Strategy

The searches were completed across the following databases between April 1, 2023, and June 30, 2023: MEDLINE (Ovid), Embase, PsycINFO, CINAHL, Emcare (Ovid), CENTRAL, Web of Science, and Scopus. The protocol was registered in the PROSPERO database (CRD42023410060). These findings are reported in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [34]. The search strategy was developed based upon previous similar systematic reviews of digital interventions and EDs [15,35] and in consultation with a specialist librarian at University College London. The search strategy included 2 main concepts based on EDs and types of digital intervention (web based or smartphone). It included a combination of Medical Subject Headings (MeSH) terms and free-text terms. The search was adapted for each database. A Cochrane RCT filter was applied to the search results within relevant databases [36]. Full details of the search strategy can be found in Multimedia Appendix 1.

The first reviewer (PT) initially screened all titles and abstracts for the first phase of the review, and a second reviewer (PB) screened a random 9.98% (375/3758) of the results within Covidence (Veritas Health Innovation). Both reviewers independently screened 100% (79/79) of articles in the final full-text screening stage. Results were compared, and any discrepancy was resolved by discussion. There was a good to excellent degree of interrater agreement (initial screening: κ=0.92 and final screening: κ=0.720).

Study Selection

Eligible studies were selected by applying the inclusion and exclusion criteria (Textbox 1).

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Adults in general population
- Self-management interventions and guided self-help interventions for individuals
- Included study participants who meet subthreshold and threshold criteria for an eating disorder
- Stand-alone digital intervention with minimal or some therapist support
- Outcome measure using the Eating Disorder Examination Questionnaire (EDE-Q)
- Randomized controlled trials

Exclusion criteria

- Interventions aimed at <16 years old
- Intervention aimed at health care professionals
- Intervention specific to relapse prevention and aftercare
- Intervention specific to eating disorder prevention
- Intervention aimed at obesity and weight management
- Telemedicine or teleconferencing
- Augmentation therapy (app as an *add-on*)
- Digital intervention with intensive levels of supplementary therapist support
- Group cognitive behavioral therapy; group therapy
- Technologies that have been superseded (ie, CD-ROM, vodcast, and SMS text messaging)
- Interventions that used mobile phones but did not involve apps (eg, were based solely on SMS text messaging or emails)
- No clear description of the intervention design (not possible to code for behavior change techniques)
- Qualitative studies
- Feasibility and acceptability studies as well as pilot studies
- No clear outcome measures (using the EDE-Q)

Data Extraction

The primary researcher (PT) extracted and coded the data for included studies, including author, year, country of origin, study and participant characteristics (number of participants, age, gender, ethnicity, diagnosis, inclusion and exclusion criteria, and dropout rates), and intervention characteristics (intervention description, therapist involvement, BCTs, modes of delivery, duration of treatment, follow-up, and key outcomes). Outcomes data for all the studies were independently extracted by 2 reviewers (PT and TR). Results were compared, and any disagreements were resolved by discussion. Where key data were missing, study authors were contacted for the missing information. A cutoff period of 4 weeks was provided.

Outcome Measures

The Eating Disorder Examination Questionnaire (EDE-Q) [37] was used as the primary outcome measure of interest, given that it is the National Institute for Health and Care Excellence "gold standard" measure of ED psychopathology and was used as the primary outcome measure in most of the included RCTs. It includes frequency data on key behavioral features of EDs in terms of number of episodes of the behavior (including bingeing and purging), making it a suitable outcome measure for this review [38]. Where reported, changes in the number of objective binge episodes (OBEs) after treatment were examined for consistency, providing complementary data on intervention effectiveness.

BCT Coding, Modes of Delivery, and Theory Coding Scheme

Each study was assessed for the presence of each of the 93 BCTs using the BCT Taxonomy v1 [30], assessing the number of BCTs in each digital intervention and the frequency of each BCT in the sample overall. The BCT Taxonomy is a hierarchically organized, common language tool for the classification of the *active ingredients* [30] required to bring about change in an intervention. The validity of this approach has been well established, and its reliability and value have been consistently demonstrated across multiple areas since its inception [39-41].

The modes of delivery used within each of the interventions to deliver the BCTs was assessed using relevant components from the Model of Delivery Ontology v2 [42]. If the modes of delivery were changed during the course of the study, the modes of delivery included within the initial study design were coded, as these were appropriate for the outcome measures used.

An adapted version of the theory coding scheme (TCS) [43] was used to evaluate the theoretical basis of the included studies. These adaptations were made in consultation with an experienced behavior change scientist (KC), on the basis that the coding scheme was originally developed for use in a different context and some of the criteria were not relevant. Hobbis and Sutton [44] justified the case for cognitive behavioral therapy (CBT) as an addition to the Theory of Planned Behavior–based interventions; hence, it was considered a valid theoretical basis when used to inform intervention design. All studies were independently coded against these frameworks by 2 reviewers (PT and RC), with any discrepancies resolved by discussion involving a third reviewer (KC). This meant the BCTs were double-blind coded by 2 reviewers across all studies. These results were compared, with a third reviewer involved where necessary to resolve any discrepancies. A briefing document was provided to the second reviewer in advance of coding, which included definitions and examples of BCTs, to ensure reliability. The coding was completed in 2 stages, with the second reviewer coding approximately 30% (5/17) papers first. The coding was compared between the 2 reviewers to identify any inconsistencies in applying the BCT framework, aiming to maximize consistency when reviewing the remaining 70% (12/17) of the papers.

For interventions to be included in the follow-up, they had to be assessed at least 8 weeks after the postintervention period. This time frame allows for a reasonable evaluation of sustained treatment effects and avoids coinciding posttreatment and follow-up evaluations across different studies.

Data Synthesis

The associations between BCTs and intervention effectiveness were analyzed. A brief narrative synthesis was used to organize and present the data within the text, with a summary of the information extracted from each study, including outcomes reported, BCTs, and other items provided in tabular form.

Frequency counts of the most commonly used BCTs were conducted for both *all interventions* and *effective interventions*, and the results were compared. The effectiveness of an intervention was determined by a statistically significant effect ($P<.05$) on ED behavior change (as measured by the EDE-Q 6.0). In studies with an active comparator, the pre-post outcome data for the intervention arm were examined independently to assess efficacy. These results were then considered in the context of the study design and compared with similar waiting list (WL) control studies. BCTs were considered effective if they were identified in at least 75% (13/17) of effective interventions [18]. A further division of effective interventions was completed based on whether they were effective at postintervention or follow-up.

Meta-Analytic Procedure

The purpose of this meta-analysis was to pool data across RCT studies regarding the effectiveness of digital interventions compared to waitlist control or treatment-as-usual (TAU) controls at postintervention and follow-up time points to explore what might be contributing to the overall effect sizes, primarily the contribution of any particular BCT. Studies with an active control group, such as face-to-face (F2F) therapy, bibliotherapy, another digital intervention, or day patient programs, as well as studies with missing (EDE-Q total) outcome data were excluded from the meta-analysis.

As a first stage, the meta-analysis procedure calculated pooled estimates of effect sizes (differences in EDE-Q total scores) at postintervention and follow-up time points for waitlist and TAU RCTs and presented these results as forest plots (using RevMan v. 5.4, The Cochrane Collaboration). Effects were based on means, SDs, and sample sizes reported within the studies. The primary outcome was EDE-Q behaviors (dietary restraint, weight concern, shape concern, and eating concern). As the included studies were RCTs, baseline values were not adjusted for across studies, as they would be expected to be similar across treatment and control groups. Due to substantial heterogeneity among the studies, which varied in design (eg, duration of treatment and level of therapist involvement), a random-effects model was used to estimate the weighted pooled effect for each outcome. This approach accounts for the distribution of the true effect across individual studies [45]. The I^2 statistic was used as a measure of heterogeneity, describing the percentage of variation across studies that was due to heterogeneity rather than chance [46]. Heterogeneity >60% was considered substantial [47] and suitable for subgroup analyses. Given that the EDE-Q primary outcome measure was continuous, the mean difference (MD) was used to describe the pooled outcome effects and the overall effect size (*z*-statistic) alongside its *P* value. Sensitivity analysis was completed to check for consistency of the effect size, and publication bias was explored using funnel plots (Multimedia Appendix 2 [13-15,48-56]).

It was then possible to complete subgroup analyses to identify whether there was evidence for any BCTs acting as moderators of effect size. A shortlist of BCTs were identified upfront according to the transdiagnostic theory of EDs by Fairburn et al [57,58]. This was to avoid post hoc analysis of multiple BCTs, which would increase the likelihood of finding significant results through chance. If any of these prespecified BCTs were identified in >75% of effective interventions, they were included in the subgroup analyses: *2.2. Feedback on behavior*, *2.3 Self-monitoring of behavior*, *2.4 Self-monitoring of outcome(s) of behavior*, *4.2 Information about antecedents*, *7.7 Exposure*, and *11.2 Reduce negative emotions*. Additional related concepts were also explored, including mode of delivery (<5 vs ≥5 out of 12 possible sessions), TCS (high vs low), degree of therapist support (none or minimal vs some), and duration of therapy (<8 weeks vs ≥8 weeks). These factors were considered as they could contribute to heterogeneity and impact effect size.

Risk-of-Bias Assessment

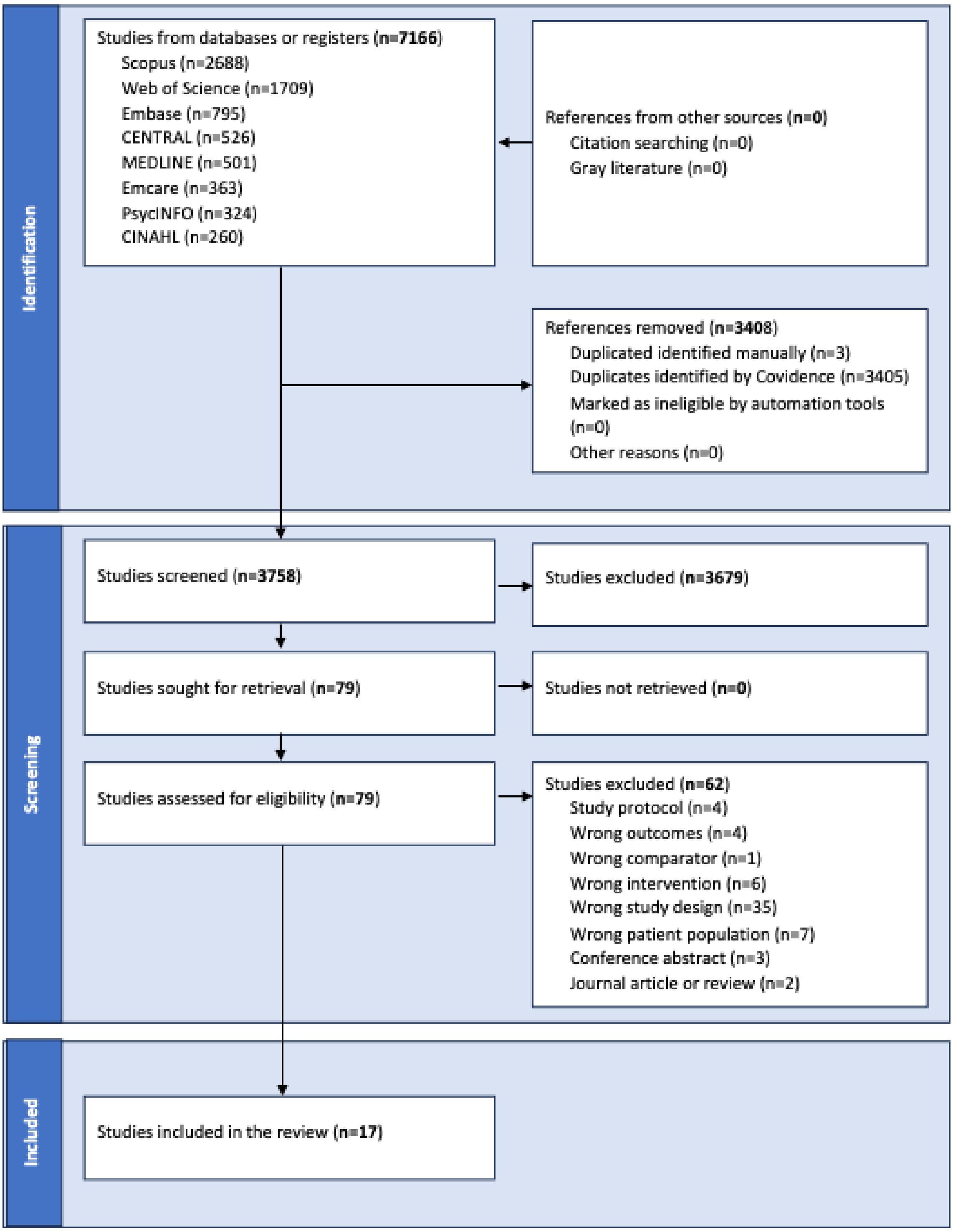
The revised Cochrane Risk-of-Bias tool for randomized trials was used for assessing risk of bias in RCTs with studies assessed against 6 domains [59] (Multimedia Appendix 3 [43]). Risk-of-bias analysis was completed for all articles by PT, with over 20% (4/17) of the articles also being independently assessed by a second reviewer (TR). Disagreements were resolved via discussion. There was a high level of interrater agreement (interrater reliability [IRR]=0.9).

Results

Included Studies

A PRISMA flow diagram (Figure 1) represents the literature search. A total of 17 RCT studies were identified for inclusion in this review.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram.



Of the 17 RCT studies identified, 12 (71%) included a WL comparator (or TAU), with 5 (21%) having *active* controls.

General Study Characteristics

The 17 studies included 12 (71%) parallel arm trials, 4 (24%) multiple-arm studies [13,48,49,60], and 1 (6%) cluster RCT [50]. Of these, 12 (71%) studies included active treatment compared to a WL control, informational control, or TAU, while 5 (29%) studies compared active treatments to other interventions, including F2F treatment [16], day patient care [61], and other digital treatment interventions [51,62,63].

A total of 9 (53%) studies included all or nearly all female participants (>95%); 5 (29%) studies included 5% to 10% male participants, and 2 (12%) studies included >10% male participants. Ethnicity was not mentioned in 12 (71%) of the 17 studies, with 2 (12%) mentioning nationality but not ethnicity and only 3 (18%) providing any ethnic breakdown. Mean age ranged from 22.1 years [50] to 43.2 years [16] across studies, with participants aged between 17.3 and 55.5 years. The total number of participants overall was 5254, with 1956 included in the meta-analysis (WL and TAU studies only). Inclusion and exclusion criteria were highly variable, with some studies having clear diagnostic criteria that had to be met, excluding participants with comorbidities or with previous experience of inference-based CBT, while others permitted individuals to participate without meeting any diagnostic criteria, provided they were aged >16 years and had access to the internet. One study allowed participants to receive other forms of psychological, medical, or other treatment for their ED, whether in the digital intervention treatment arm or control condition [52].

The studies took place in North America (2/17, 12%) [50,62], Europe (11/17, 65%; Switzerland, Germany, Sweden, Austria, and the Netherlands) [13,15,16,48,49,52-55,64,65], and Australia or New Zealand (4/17, 24%) [14,51,60,63]. The included studies are listed in Multimedia Appendix 4 [13-16,48-55,60-65].

Summary of Intervention Types and Outcomes

The ED diagnoses included 6 studies focusing on binge eating disorder and binge eating symptoms [15,16,51,52,60,63], 3 studies on bulimia or eating disorders not otherwise specified [13,64,65], and 8 studies concerning individuals with any ED symptoms [14,48-50,53-55,62]. The studies included a number of different interventions (Multimedia Appendix 5 [13-16,48-55,60-65]), with the most common being Salut BED or Salut BN (5/17, 29%) [13,15,16,64,65], Break Binge Eating or Break the Diet Cycle (4/17, 24%) [14,51,60,63], and Featback (2/17, 12%) [48,49].

Studies included internet and mobile-based digital interventions, frequently including messaging or email feedback or prompts. A total of 2 studies focused specifically on an app [14,62], 4 studies included blended internet and smartphone interventions [50,51,60,63], and 11 studies were internet-only interventions. Interventions lasted between 4 weeks and 12 months, with 11 interventions lasting ≤8 weeks and 6 interventions lasting >8 weeks [13,15,48-50,53]. Interventions varied in the number of modules, ranging from 4 to 11, which resulted in differences in the amount of content provided and allowed for varying timescales to complete these modules.

Only studies with digital interventions with no or relatively minimal levels of therapist support (eg, weekly emails) as well as interventions with *some* therapist support were included. This resulted in 4 studies with no therapist involvement [14,48,49,62], 7 studies with *minimal* therapist involvement [13,15,51,52,54,60,63], and 6 studies with *some* therapist involvement [16,50,53,55,61,64].

Outcome measures were most commonly the EDE-Q, although other measures such as the number of OBEs were also frequently reported. Dropout rates at postintervention measurement were between 6.7% and 58% for the digital intervention. They tended to be higher in the interventions with minimal or no support conducted in a community setting, such as those in which participants signed up and participated via an internet service [14,48,49,51,62]. However, design characteristics such as feedback on behavior or feedback on outcomes of behavior also seemed important [51].

The details of the digital interventions within the 17 studies, including their constituent BCTs, are described in Multimedia Appendix 5.

Study Outcomes at Postintervention and Follow-Up

A total of 11 (92%) of the 12 RCTs that compared a digital intervention to a WL or TAU control demonstrated a significant improvement in ED outcomes (as measured by the EDE-Q) for the digital intervention over the control condition at postintervention, except for the study by Aardoom et al [48]. The WL and TAU control studies that reported the number of binge eating episodes at the postintervention time point (11/12, 92%) [13-15,48,49,51,53-55,60] also reported a significant reduction in OBEs compared to the WL and TAU control condition. All WL and TAU studies that reported follow-up data (9/12, 75%) reported a significant reduction in ED outcomes (EDE-Q total and OBEs) compared to the control condition, including the study by Aardoom et al [48].

When the control condition was an active comparator, of traditional F2F treatment [16] or a day patient program [61], participants in the active comparator arm performed considerably better than the digital intervention at the postintervention time point, but results were comparable at follow-up in both studies. Where the active comparator was a similar digital health intervention, either broader in terms of functionality [51] or consisting of interactive versus static content [62,63], there were no significant differences observed in EDE-Q total outcomes or secondary outcome measures at the postintervention time point (and no follow-up data).

BCTs in Effective Interventions

A total of 38 (41%) out of 93 BCTs were identified across the clinical content of the interventions (Table 1). The mean number of BCTs per intervention was 14 (SD: 2.57, range 9-18). The following BCTs were reported in >75% (13/17) of effective interventions: *2.3 Self-monitoring of behavior*, *1.2 Problem-solving*, *4.2 Information about antecedents*, *2.2 Feedback on behavior*, *2.4 Self-monitoring of outcomes of behavior*, and *1.4 Action Planning*. *7.7 Exposure* and *11.2 Reduce negative emotions*, which had been predicted to be important, were identified 56% (9/16) and 38% (6/16) of effective interventions, respectively. *5.2 Behavioral practice/rehearsal* (10/16, 63%), *13.2 Framing/Reframing* (10/16, 63%), and *7.1 Prompts/Cues* (9/16, 56%) were present in >50% of effective interventions, suggesting they may also be important in supporting ED behavior change. The IRR was high (IRR=0.84).

BCTs were not identified from the following BCT categories in the taxonomy: *6. Comparison of the behavior*, *14. Schedules consequences*, or *16. Covert learning*. Only 3 studies included a component from the *10. Reward and Threat* category, the *10.4 Social Reward* component.

Table 1. Behavior change techniques included in the treatment interventions (by study[a])

Evaluation of eating disorder studies	[15]	[13]	[16]	[55]	[52]	[62]	[63]	[51]	[60]	[14]	[53]	[49]	[50]	[48]	[54]	[61]	[64]	AI[b] (n=17), n (%)	EI[c] (at post intervention; n=16), n (%)
2.3 Self-monitoring of behavior	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	17 (100)	16 (100)
1.2 Problem solving	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	16 (94)	15 (94)
4.2 Information about antecedents	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	16 (94)	15 (94)
2.2 Feedback on behavior	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓		✓	14 (82)	13 (81)
2.4 Self-monitoring of outcomes of behavior	✓				✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	14 (82)	13 (81)
1.4 Action planning	✓	✓			✓	✓	✓	✓	✓	✓	✓		✓			✓	✓	12 (71)	12 (75)
8.1 Behavioral practice or rehearsal	✓		✓			✓	✓	✓		✓			✓		✓	✓	✓	10 (59)	10 (63)
13.2 Framing or reframing	✓	✓				✓	✓	✓	✓	✓			✓		✓	✓		10 (59)	10 (63)
7.1 Prompts or cues				✓		✓	✓	✓	✓	✓	✓	✓		✓				9 (52)	9 (56)
7.7 Exposure	✓	✓	✓	✓			✓		✓				✓		✓		✓	9 (52)	9 (56)
3.2 Social support (practical)	✓		✓	✓	✓							✓		✓	✓		✓	8 (47)	7 (44)
4.1 Instructions on how to perform the behavior				✓	✓		✓	✓	✓	✓			✓	✓				8 (47)	7 (44)
5.1 Information about health consequences					✓		✓		✓		✓	✓		✓	✓	✓		8 (47)	7 (44)
2.7 Feedback on outcomes of behavior									✓		✓	✓	✓	✓	✓		✓	7 (41)	6 (38)
3.1 Social support (unspecified)	✓		✓									✓	✓	✓		✓	✓	7 (41)	6 (38)
8.2 Behavior substitution				✓				✓		✓	✓				✓	✓		6 (35)	6 (38)
11.2 Reduce negative emotions				✓				✓		✓			✓		✓		✓	6 (35)	6 (38)
8.4 Habit reversal				✓		✓		✓		✓			✓					5 (29.4%)	5 (31.3%)

Evaluation of eating disorder studies	[15]	[13]	[16]	[55]	[52]	[62]	[63]	[51]	[60]	[14]	[53]	[49]	[50]	[48]	[54]	[61]	[64]	AI[b] (n=17), n (%)	EI[c] (at post intervention; n=16), n (%)
5.6 Information about emotional consequences					✓				✓			✓	✓	✓				5 (29)	4 (25)
8.3 Habit formation				✓			✓						✓			✓		4 (25)	4 (25)
9.3 Comparison of future outcomes	✓		✓			✓			✓									4 (25)	4 (25)
1.1 Goal setting (behavior)				✓	✓	✓												1 (18)	3 (19)
1.3 Goal setting (outcome)				✓	✓	✓												1 (18)	3 (19)
9.2 Pros and cons	✓		✓			✓												1 (18)	3 (19)
12.4 Distraction								✓		✓			✓					1 (18)	3 (19)
15.4 Self-talk	✓	✓											✓					1 (18)	3 (19)
3.3 Social support (emotional)				✓								✓		✓				1 (18)	2 (13)
5.3 Information about social and environmental consequences												✓		✓	✓			1 (18)	2 (13)
10.4 Social reward						✓						✓		✓				1 (18)	2 (13)
5.4 Monitoring of emotional consequences	✓		✓															2 (12)	2 (13)
15.3 Focus on past success				✓		✓												2 (12)	2 (13)
15.1 Verbal persuasion about capability												✓		✓				2 (12)	1 (6)
1.5 Review behavioral goal						✓												1 (6)	1 (6)
1.7 Review outcome (goal)						✓												1 (6)	1 (6)
1.9 Commitment				✓														1 (6)	1 (6)

Evaluation of eating disorder studies	[15]	[13]	[16]	[55]	[52]	[62]	[63]	[51]	[60]	[14]	[53]	[49]	[50]	[48]	[54]	[61]	[64]	AI[b] (n=17), n (%)	EI[c] (at post intervention; n=16), n (%)
2.1 Monitoring of behavior by others without feedback											✓							1 (6)	1 (6)
4.4 Behavioral experiments		✓																1 (6)	1 (6)
13.4 Valued self-identity				✓														1 (6)	1 (6)

[a] Carrard et al [15] (2011), Ruwaard et al [13] (2013), de Zwaan et al [16] (2017), Strandskov et al [55] (2017), Wyssen et al [52] (2021), Tregarthen et al [62] (2019), Linardon et al [63] (2022a), Linardon et al [51] (2022b), Linardon et al [60] (2021b), Linardon et al [14] (2020), Melisse et al [53] (2023), Rohrbach et al [49] (2022), Fitzsimmons-Craft et al [50] (2020), Aardoom et al [48] (2016), Jacobi et al [54] (2012), Högdahl et al [61] (2023), Wagner et al [64] (2013).

[b] AI: All interventions.

[c] EI: Effective interventions.

Follow-up data (>8 weeks after postintervention) was available for 9 (53%) out of the 17 studies. In 2 of the studies, there was no data available for the control condition because participants received the intervention. However, since the outcome effects at postintervention were sustained at follow-up, these studies were still included in the analysis [13,52]. A total of 2 studies included an active comparator [16,61], showing improvements on the EDE-Q for the digital intervention arm at postintervention that were sustained or improved at follow-up; hence, they were included in the analysis. This analysis (Table 2) resulted in the following BCTs being identified in effective interventions at follow-up (in >75% of interventions): *2.2 Feedback on behavior*, *2.3 Self-monitoring of behavior*, *2.4 Self-monitoring of outcomes of behavior*, *4.2 Information about antecedents*, and *1.2 Problem-solving* (these all were the same at the postintervention time point). The BCTs of *3.2 Social support (practical)*, *3.1 Social support (unspecified)*, and *5.1 Information about health consequences* were more evident in the interventions that were effective at follow-up compared with the postintervention time point. These may be important in sustaining positive outcome effects; however, these findings are based on a small number of studies.

Definitions of the most common BCTs (included in at least 9/17, >50% of interventions), with examples of how they were implemented within the interventions, are included in Multimedia Appendix 6.

Table 2. Behavior change techniques included in effective treatment interventions at follow-up (by study).

Evaluation of eating disorder Studies	Carrard et al [15] (2011)	Ruwaard et al [13] (2013)	de Zwaan et al [16] (2017)	Wyssen et al [52] (2021)	Rohrbach et al [49] (2022)	Fitzsimmons-Craft et al [50] (2020)	Aardoom et al [48] (2016)	Jacobi et al [54] (2012)	Högdahl et al [61] (2023)	Effective (at follow-up; 9 studies had follow-up data), n (%)
2.3 Self-monitoring of behavior	✓	✓	✓	✓	✓	✓	✓	✓	✓	9 (100)
1.2 Problem solving	✓	✓	✓	✓	✓	✓	✓		✓	8 (89)
2.2 Feedback on behavior	✓	✓	✓	✓	✓	✓	✓	✓		8 (89)
4.2 Information about antecedents	✓	✓	✓	✓	✓	✓	✓		✓	8 (89)
2.4 Self-monitoring of outcomes of behavior	✓			✓	✓	✓	✓	✓	✓	7 (78)
3.1 Social support (unspecified)	✓		✓		✓	✓	✓		✓	6 (67)
3.2 Social support (practical)	✓		✓	✓	✓		✓	✓		6 (67)
1.4 Action planning	✓	✓		✓		✓			✓	5 (56)
5.1 Information about health consequences				✓	✓		✓	✓	✓	5 (56)
7.7 Exposure	✓	✓	✓			✓		✓		5 (56)
8.1 Behavioral practice or rehearsal	✓		✓			✓		✓	✓	5 (56)
13.2 Framing or reframing	✓	✓				✓		✓	✓	5 (56)

Theoretical Basis

Nearly all studies (16/17, 94%) reported some level of theoretical basis to their intervention design (Multimedia Appendix 3). Of those that did mention a theoretical basis, CBT and the transdiagnostic theory of EDs were most frequently reported [31,57], sometimes in combination with other theoretical approaches, including acceptance commitment therapy (ACT) and dialectical behavior therapy [51,52]. The description of this theoretical basis was often minimal within the studies; however, these approaches are generally well understood and accepted within ED treatment, and further literature was often referenced [57] to support their use.

Of the 17 studies, 13 (77%) mentioned a target construct as a predictor of behavior (eg, emotional regulation and body image concerns) and designed interventions that targeted these constructs to change ED behaviors. A total of 13 (77%) of the 17 studies reported how theory or predictors were used to select or develop BCTs. However, this was not often done explicitly; instead, interventions typically listed features alongside their theoretical constructs (eg, emotional regulation—access to an emotions tracker and body image concerns—an exercise to break avoidance patterns). Only 4 studies used theory or predictors to tailor interventions to participants [53,55,62,63] based on their specific eating-related concerns.

Modes of Delivery

The mean number of modes of delivery per intervention was 4 (SD: 1.6, range 2-7) out of 12 possible modes (Multimedia Appendix 3). All interventions included textual information, after which the most common mode of delivery was website (15/17, 88% of studies). Mobile apps were included in just 6 (35.3%) of the 17 studies. Of the 6 studies, 4 (24%) included both website and app modes of delivery [50,51,60,63] and 2 (12%) were app only [14,62].

Video and audio modes of delivery were identified in only 18% (3/17) and 29% (5/17) of the apps, respectively, suggesting rather limited use of multimedia functionality within the interventions, with a greater reliance upon textual information. In 47% (8/17) of the studies, an at-a-distance mode of delivery involving human interaction was included. This typically involved therapists providing weekly feedback on behaviors and assignments delivered via SMS text messaging (10/17, 59%) or email (9/17, 53%). Email was also used to *check-in* with participants to ensure engagement with the intervention. Although phone was used in 24% (4/17) of the interventions, this was usually only if the user was not engaging in the service at risk of dropout, rather than being part of the service.

It should be noted that Fitzsimmons-Craft et al [50] changed their study design after 1 year, based on performance in the first year, from a web-based intervention to an app-based intervention. Given that the outcomes at 1 year were used in the analysis, the app mode of delivery was not coded.

Risk of Bias

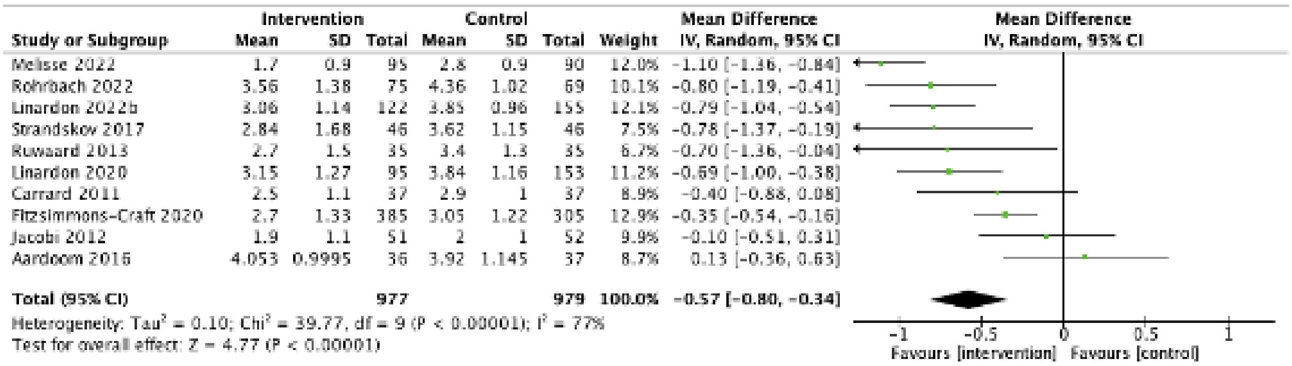
Most studies (15/17, 88%) reported an adequate method of randomization, frequently including computer-generated randomization sequence, although assessors were not always blind to treatment allocation. Most studies (16/17, 94%) reported adequate blinding of outcome assessment, either through the use of web-based self-report outcome assessments or through F2F or phone assessments, in which assessors were blind to treatment allocation. No studies reported blinding participants to the digital intervention, which would have been difficult to achieve. However, none discussed how this lack of blinding might have biased the self-reported outcomes.

The domain where studies scored lowest was in terms of missing outcome data (13/17, 77% studies), which was due to the relatively high attrition rates across studies. Some studies deviated from their analysis plan, including alternative statistical methods in their analysis [49]. Although these may have been justified, they introduced *some concerns* in how those studies had been analyzed and the data that were reported. There was also selective reporting of the results in 5 (29%) of the 17 studies, which put them at a higher risk of bias (Multimedia Appendix 3). When a subgroup analysis was conducted between studies with low or some concerns regarding bias and those at high risk of bias, no significant differences in outcomes were observed (Multimedia Appendix 2).

Results From Meta-Analysis

We used the MD of EDE-Q total scores as the primary estimate of effect size for each intervention. A total of 10 studies were included with WL or TAU control with EDE-Q outcome data at the postintervention time point. Although 12 studies had a WL or TAU control, one study was excluded due to missing outcome data on the dietary restraint subscale [60] and another was excluded as the control group was given the intervention at 4 weeks; hence, the study comparison at the postintervention time point was against 8 weeks versus 4 weeks active treatment [52]. The pooled effect sizes for the comparison between digital ED interventions and WL or TAU control groups was moderate and statistically significant in favor of the treatment group for ED psychopathology (MD=–0.57, 95% CI –0.080 to –0.39; Z=4.77; P<.001). Heterogeneity was high (I^2=77%), making it sensible to conduct subgroup analyses (Figure 2).

Figure 2. Forest plot showing the mean difference in outcomes (Eating Disorder Examination Questionnaire) for digital eating disorder interventions versus waiting list and TAU controls (at the postintervention time point). TAU: treatment as usual.



Sensitivity analysis was completed by removing studies one at a time to consider the impact on effect size, but this did not change the results significantly. There was no clear evidence of publication bias based on a relatively even distribution of studies around the summary estimate line (Multimedia Appendix 2).

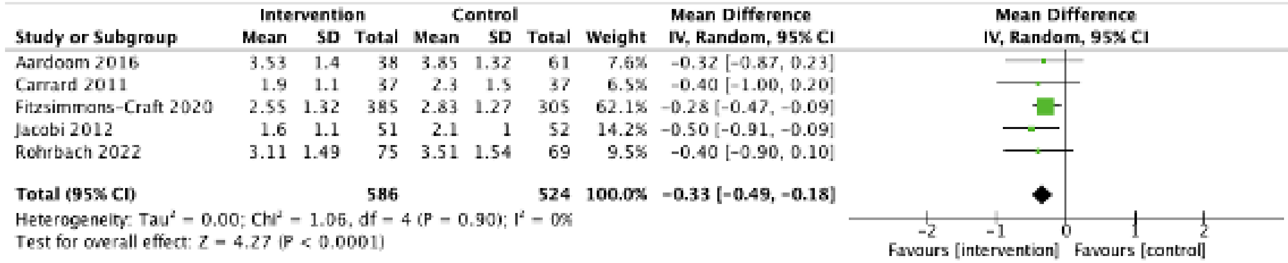
It should be noted that while data used for this meta-analysis did not demonstrate statistical significance for the studies by Jacobi et al [54] or Carrard et al [15], when baseline values were adjusted for (as in the original papers), outcomes significantly favored the interventions compared to the control at the postintervention time point in both studies (P<.001). Therefore, the interventions were considered to be effective at the postintervention time point. Baseline values were not adjusted for within the meta-analysis based on the assumption that randomized controlled studies should not have baselines differences.

Moderator and Subgroup Analyses

A total of 6 moderator analyses were conducted to investigate differences in EDE-Q total pooled effect size according to the presence or absence of BCTs in digital interventions. None of the subgroup analyses of BCTs explained any of the heterogeneity of effect sizes across the studies, suggesting that there were other factors that explained this heterogeneity (refer to the example in Multimedia Appendix 1). Heterogeneity within BCT subgroups was also moderate, confirming that there were likely to be other factors explaining this variability.

Digital ED interventions that had a higher score on the TCS had a greater effect size than those with a lower TCS score (Figure 3). Subgroup analyses showed that interventions that were more highly grounded in theory (high TCS mean=–0.86, 95% CI –1.06 to –0.66; I^2=37%) were significantly more effective than those that had a low theoretical basis (low TCS mean=–0.36, 95% CI –0.61 to –0.11; I^2=56%; subgroup differences: χ^2_1=9.7; *P*=.002; I^2=89.7%; Multimedia Appendix 2). This was the only statistically significant moderation effect that emerged from the subgroup analyses.

Figure 3. Forest plot showing the results of meta-analysis from theory coding scheme subgroup analysis. TAU: treatment as usual.



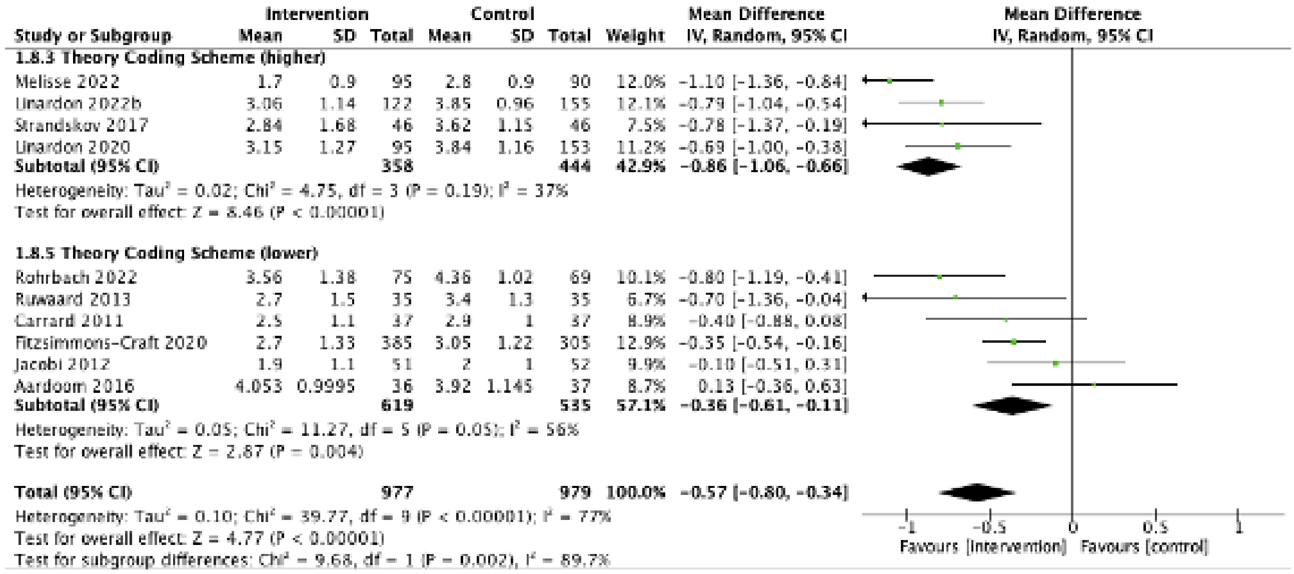
There were no significant differences across other subgroup analyses. All subgroup analyses are presented in Multimedia Appendix 2.

Follow-Up

Only 5 studies with WL and TAU control included EDE-Q outcome data at follow-up (>8 weeks) [15,48,50,54]. The results were significant, with reduction in ED psychopathology favoring the treatment arm (MD=–0.33, 95% CI –0.049 to –0.18) and an overall effect size of *z*=4.27 (*P*<.001). There was no heterogeneity (I^2=0%; Figure 4).

A total of 2 studies were considered high risk of bias, and the remaining 3 studies had *some concerns* due to missing data and selective reporting of the study data; hence, these data should be interpreted cautiously. Given the limited number of studies with outcome data at follow-up (and lack of heterogeneity), subgroup analyses were not completed.

Figure 4. Forest plot showing the mean difference in outcomes (Eating Disorder Examination Questionnaire) for digital eating disorder interventions versus waiting list and TAU controls at follow-up. TAU: treatment as usual.



Discussion

Principal Findings

There is good evidence to support the efficacy of digital interventions (mainly websites) for people with mild to moderate EDs, with 16 (94%) out of the 17 studies demonstrating efficacy at the postintervention time point, strengthening findings from previous reviews [15,35,56,66]. Effects appear to be maintained at follow-up, with some studies demonstrating continuous reduction in bingeing and purging symptoms with effect sizes similar to those observed in F2F treatment [16,61]. There were few studies on smartphone-based interventions (*apps*); hence, data on their effectiveness as self-management tools, or as guided interventions, remain limited, and further research is required.

Interventions included an average 14 (SD 2.6; range 9-18) BCTs, which compares favorably with other reviews of digital behavior change interventions [67,68], demonstrating that existing interventions already incorporate BCTs to help change ED behaviors. Across the various interventions reviewed, there was a high level of agreement regarding BCTs that were included, which were *Self-monitoring of behavior*, *Self-monitoring of outcomes of behavior*, *Information about antecedents*, *Problem-solving skills*, and *Feedback on behavior*. These are in line with the principles of CBT-ED and the transdiagnostic theory of EDs by Fairburn [57]. Other CBT-ED–related BCTs of *Exposure*, *Cognitive restructuring*, and *Reducing negative* *emotions* also ranked moderately highly, although there may be an opportunity to integrate these techniques further within digital interventions based on their relevance in treating patients with EDs. While *Prompts or cues* were present in just 56% (9/16) of effective interventions, these techniques may be important to facilitate user engagement within digital interventions [69], which is important if these interventions are to be effective for a greater number of people by reducing dropout rates.

Some effective interventions (3/16, 19%) included additional techniques that are often used in therapy, such as *Distraction* and *Pros and cons*; however, there was insufficient evidence to evaluate if these helped contribute to intervention effects. There were no BCTs in the categories of *Comparison of the behavior*, *Scheduled consequences*, or *Covert learning* across the digital ED interventions, and *Reward or Threat* techniques were rarely used. There is an opportunity to explore how these could be used, potentially learning from other areas of digital health behavior change and testing some of these techniques with potential users. At follow-up, it seemed that social support may be important in supporting a sustained outcome effect [15,48,50], achieved through personalized feedback, encouragement, and practical advice provided within the intervention. This enabled users to achieve greater self-awareness, improved coping skills, greater accountability, and the development of a more supportive social network to assist them in their recovery.

There was no indication that individual BCTs were responsible for differences in outcome effects. This may have been due to the limited number of studies, the small numbers of participants, other factors accounting for the heterogeneity, and the similarity of digital intervention characteristics. It is also because of the study designs, which did not facilitate direct comparison of intervention components across studies. A different approach to design involves using a factorial RCT guided by the Multiphase Optimization Strategy [70], which enables the simultaneous evaluation of multiple variables (eg, BCTs and modes of delivery) and their interactions, without the need for a large sample size. Most studies used CBT-based internet interventions (and some used the same or similar interventions, eg, Salut BN or Salut BED and Break Binge Eating); hence, it could be that there was insufficient variability in the BCTs across studies, making it difficult to detect an association between the most commonly reported BCTs and treatment outcomes. It is also most likely that a combination of BCT inclusion, dose, mode of delivery, and theoretical basis may be important for intervention effectiveness alongside other key design characteristics. Further studies are required to better understand how these factors interact to achieve their effects.

Nearly all studies (16/17, 94%) referred to a theoretical basis for their intervention design; however, they differed to the extent to which theory had been rigorously applied. Most interventions (16/17, 94%) were based on CBT, informed by the transdiagnostic theory of EDs by Fairburn [57], although some interventions also incorporated techniques from ACT and dialectical behavior therapy [51,55]. Interventions that were informed by theory seemed to have a greater effect size within the meta-analysis, consistent with what was hypothesized. They were designed to target those specific MOAs (eg, dietary restraint, body image concerns, and emotional dysregulation) that have been identified as important in changing ED behaviors. However, this result should be interpreted with caution due to the small number of studies, other factors that could explain this result, and the remaining heterogeneity within these subgroups requiring further explanation.

The mean number of modes of delivery per intervention was 4 (SD 1.6, range 2-7) out of 12 possible modes, with a heavy reliance upon textual information and a limited use of audio and video to deliver the BCTs. Nearly half of the interventions (8/17, 47%) included some degree of human interaction, delivered *at a distance*; however, there was no evidence that therapist involvement moderated effect size. There was also no evidence that an increased number of modes of delivery moderated any outcome effect. A key finding was that 2 (12%) of the 17 interventions were app only, suggesting that we require more evidence on app-only approaches with no or minimal therapist support. At the time of this review, the technology used across interventions was relatively homogeneous; hence, we focused on modes of delivery to capture differences in how the interventions were delivered. As technology evolves, it may be important to consider the type of technology used, such as artificial intelligence, as an additional moderating factor.

Comparison With Prior Work

These results strengthen findings from previous meta-analyses, which provide initial evidence for the effectiveness of digital interventions for reducing ED symptoms [15,35,71,72]. Loucas et al [66] found small effects in their review of internet-based treatments for EDs (n=20), but with the inclusion of more recent studies, small to moderate effects have consistently been demonstrated, with some participants showing significant improvement in ED behaviors at postintervention [53] and at follow-up [15].

Results are consistent with a previous systematic review of mobile health (mHealth) interventions for EDs [73] that concluded that mHealth interventions, either as a self-management tool or complementary to F2F therapy, had limited support. Previous qualitative research has highlighted the promise of such interventions, with high levels of interest in mobile apps and level of acceptability [11,74], although the number of RCTs that demonstrate efficacy remains limited [14,62]. Specific advantages have been identified by patients and clinicians, such as better supporting the real-time logging (food and mood), tracking and feedback to users, reminders to increase adherence to the intervention [17], and the opportunity for just-in-time interventions when an individual may be an elevated risk of engaging in an *unhealthy* behavior (eg, purging) [75]. Research to translate these ideas into effective ED apps that have a place in treatment is still ongoing.

Although this study found that increased levels of multimedia within the digital interventions did not mediate intervention effects, previous research [15] did find that studies with increased use of multimedia channels (audio, video, etc) were associated with greater improvement in ED symptoms. Barakat et al [76] performed a more robust analysis of multimedia channels, analyzing data from surveys returned by the study authors and incorporating additional components such as quizzes and homework, which provided a more detailed and accurate reflection of multimedia inclusion, despite including a range of study designs. Their findings need to be replicated by including more recent RCTs, especially given that they included older studies, some of which were based on now-obsolete technologies (ie, CD-ROM and vodcasts). Interactivity alone is unlikely to meaningfully affect key outcomes in internet-based interventions; instead, it will likely be a combination of interactivity and other design characteristics, such as the quality of intervention content, personalization, persuasive design, or therapeutic alliance principles [7], which are important determinants of outcomes. It could also be that certain populations, such as those with neurodiversity (eg, autism spectrum disorder [77]), benefit more from increased levels of multimedia within digital interventions.

Strengths

We only included studies using an RCT design, which has not been the case in previous reviews [15,35]. This is the first study to systematically review the BCTs within digital ED interventions, providing greater insights and a more comprehensive picture to inform intervention design and evaluation. Studies that included blended interventions or high levels of therapist support were excluded to allow a thorough analysis of the BCTs within digital interventions and how these may be specifically contributing toward symptom improvement. This study evaluated the effect of BCTs, modes of delivery, and theoretical underpinning on intervention outcomes quantitatively as well as narratively to enable a rigorous evaluation of the data. A large number of databases were searched to ensure that all relevant studies are included in this review, and we found 9 studies that were not included in previous similar reviews. This study has furthered our understanding of how to develop effective digital interventions, providing an opportunity to develop new or improved mHealth interventions for EDs that have the potential to be effective.

Most participants in these studies were recruited from within a community setting; hence, they should be reflective of those with ED in the population who may not currently be getting help from clinical services. This is especially important given the significant increase in demand for ED services since the pandemic [7] and a sustained move to the use of more digital services.

Limitations

Only 10 studies were suitable for inclusion in the meta-analysis, restricting the power required to detect significant moderating effects of BCTs. Some studies included a small number of participants; hence, it might be underpowered to demonstrate significant differences compared to control groups or significantly affect the meta-analytic findings due to low weight. CIs in several of the studies were relatively large, limiting the ability to find significant results across the pooled studies.

Given that the meta-analysis only examined differences in effect sizes between the digital interventions and control at postintervention and follow-up without including baseline values, it did not assess whether the observed differences were clinically meaningful. In addition, use of the EDE-Q may not have provided a clear picture of all changes in ED behaviors, as not all compensatory behaviors are adequately covered by the EDE-Q [78]. For example, it is possible that participants replaced purging with nonpurging compensatory behavior, such as excessive physical exercise, dieting, and fasting. Studies using self-reported measures of outcomes may not have accurately reflected actual outcomes being subject to self-reporting bias.

Dropout rates in some of the studies was high, varying from 6.7% to 58%. Although studies typically assessed differences in baseline characteristics between those who completed and those who dropped out and typically found minor or no differences, the proportion of participants across who did not complete treatment and provide postintervention assessments is a significant limitation.

We did not have access to the interventions; hence, the BCT coding was based on descriptions of interventions that were available in the public domain (ie, journal publications, supporting information, etc) and some discussion with authors. Studies often did not go into much detail about the theoretical basis upon which the interventions were developed; hence, we were limited in terms of the information that could be coded. While there was a high level of IRR (IRR=0.83), there was an element of subjectivity in how the BCT Taxonomy was interpreted and applied. The theoretical coding scheme used was abbreviated for the purposes of this review and has not been externally validated. There was limited follow-up data; hence, it was not possible to evaluate the effectiveness of the interventions over the longer term (nor complete subgroup analyses). Some BCTs may have helped specific user populations, but studies did not report on outcomes for specific populations, limiting our understanding of what worked for whom. While it is helpful to categorize interventions based on their BCTs, the BCT Taxonomy v1 may be inadequate for identifying all active ingredients that might be contributing to the effectiveness of an intervention, such as those included within ACT.

Further Work

Further research is needed to evaluate the effects of specific BCTs and combinations of BCTs to identify which are most crucial for improving outcomes in digital interventions for EDs. This research should also explore who benefits most from these techniques and which modes of delivery are most effective. A factorial experiment would allow different combinations of BCTs to be tested to see which combinations, as well as the effect of different modes of delivery, are the most effective [70]. Greater consistency in RCT design would be helpful to maximize the learnings that can be gained as to what is effective, as there continues to be considerable heterogeneity across study designs. Existing digital programs for EDs typically involve numerous strategies, techniques, or modules designed to target a range of behavior change mechanisms, such as restrictive eating, mood dysregulation, body image concerns, and low self-esteem deficits [16,50]. Therefore, further research is required into how to tailor interventions to better meet the needs of individual patients or user *clusters* [79]. Receiving intervention content that is not relevant to a user's symptom profile may lead to issues with motivation, engagement, and dropout [80]. One way in which this could be explored is via bandit trials, which are a type of adaptive intervention design that allow for personalized treatment allocation based on individual responses. Treatment outcomes across the different intervention options could be evaluated, with a further analysis to determine which treatment options are most effective for which individuals.

The BCT Taxonomy v1 has since evolved into an ontology [81], which could be applied to help identify any additional techniques, such as in ACT, which may not have been accounted for in this review. We did not analyze the dose of BCTs by coding the frequency of each BCT within interventions. This decision was made to avoid adding an additional layer of complexity to this review. Further research could explore whether there is an optimal dose for BCTs.

There is some evidence to suggest that some specific BCTs may improve the user experience and adherence to treatment, which could be explored via further qualitative research, including the way in which BCTs are translated in an intervention. It also may be worthwhile to get user feedback on those BCTs that have rarely been incorporated into digital ED interventions to establish if they may be beneficial to users. This should include those that are commonly used within therapy but are not widely implemented within digital interventions. Further work is required to understand how to leverage the benefits of mobile apps, such as enabling real-time data capture and the opportunity for just-in-time intervention [75] at the point of need.

These studies included minimal or no therapist support. The study by Aardoom et al [48] suggests that self-guided interventions can be effective with automated feedback, while some therapist involvement improves user satisfaction. In depression and anxiety, studies show that treatment programs with some level of guidance are more effective compared to those without some level of guidance [82,83]. More work is required to understand what level of support is optimal, how it benefits users, and the cost-effectiveness of additional support [84]. Research into what level and type of therapist interactions are sufficient to develop any therapeutic alliance within digital ED interventions requires further study, given that therapeutic alliance has been shown to be positively associated with treatment outcome in both F2F treatment [85] and internet-based treatment [86]. This includes research into the use of artificial intelligence chatbots and how they might support the establishment of an alliance [87].

There remains a lack of studies of digital ED interventions involving older people, men, and those belonging to sexual and ethnic minority groups [17]. There is evidence to suggest that ethnicities may have differing requirements from an ED intervention, and these populations may also be less likely to access treatment [72]. It is important that these groups are represented in future research on digital health interventions in EDs from the outset [35] to support the design and development of more accessible and inclusive digital interventions.

Further research is required to understand exactly where these interventions should fit in the treatment pathway to complement the work of ED therapists and health care professionals in this field. It is crucial that this research is translated into *real-world* interventions to offer more evidence-based apps to people with mild to moderate EDs [88,89]. However, it is important to recognize that these apps may not be suitable for everyone and that health care professional support may still be necessary at some stage.

Conclusions

There is increasing evidence for the effectiveness of digital interventions for the treatment of people with mild to moderate EDs, with improved outcomes at postintervention and sustained outcomes at follow-up time points. Effective digital ED interventions mostly used the same specific BCTs, informed by theory; however, there was no evidence that any 1 BCT contributed to improvements in ED behaviors. However, the presence of self-monitoring in 100% of effective interventions suggests that it may be important for enabling ED behavior change. There seems to be an opportunity for further refinement of BCTs within digital interventions to improve intervention effectiveness by applying learnings from what works in therapy and conducting factorial experiments.

The interventions that were informed by theory and where theory had been applied to identify mechanisms of change and select specific BCTs within the intervention had better outcomes. There was no evidence that increasing the number of modes of delivery had an impact on effect size. There were few studies that evaluated digital apps, indicating potential for the development of higher-quality, evidence-based apps to enhance access to treatment. Future interventions should be grounded in theory targeting those specific mechanisms of change which are important for improving individuals' ED behaviors.

Acknowledgments

The authors would like to thank William Henderson of University College London Library Services for his assistance in developing the search terms and reviewing the search strategy.

Authors' Contributions

PCT, KC, and SR conceived and designed the review question and search strategy. PCT ran the searches; did the data extraction, quality appraisal, and narrative synthesis; and led the write-up of the manuscript. PB, RP, and TR were independent raters for the screening, data extraction, coding, and quality appraisal stages. HWWP provided input into the design of the meta-analysis and reviewed the findings. All authors have seen and commented on the final version of the manuscript.

Conflicts of Interest

PCT is an honorary eating disorder researcher at Thrive Mental Wellbeing. PCT has received no funding for this project. HWWP has paid consultancies with Thrive Therapeutic Software and Flo Health. He supervises PhD students with relationships with Better Points Ltd, Patients Know Best and AstraZeneca.

Multimedia Appendix 1
Search strategy.
[DOCX File , 45 KB - mental_v11i1e57577_app1.docx]

Multimedia Appendix 2
Supplementary subgroup analyses from the meta-analysis.
[DOCX File , 8992 KB - mental_v11i1e57577_app2.docx]

Multimedia Appendix 3
Results of the theory coding scheme, modes of delivery, and risk-of-bias analyses.
[DOCX File , 816 KB - mental_v11i1e57577_app3.docx]

Multimedia Appendix 4
Summary of the study characteristics.
[DOCX File , 63 KB - mental_v11i1e57577_app4.docx]

Multimedia Appendix 5
Summary of the interventions including behavior change techniques and modes of delivery.
[DOCX File , 60 KB - mental_v11i1e57577_app5.docx]

Multimedia Appendix 6
Most common behavior change techniques, definitions, and implementation examples.
[DOCX File , 20 KB - mental_v11i1e57577_app6.docx]

Multimedia Appendix 7
PRISMA checklist.
[PDF File (Adobe PDF File), 61 KB - mental_v11i1e57577_app7.pdf]

References

1. Ignoring the alarms: how NHS eating disorder services are failing patients. Parliamentary and Health Service Ombudsman (PHSO). 2017. URL: https://www.ombudsman.org.uk/ [accessed 2024-04-29]
2. Eisenberg D, Nicklett EJ, Roeder K, Kirz NE. Eating disorder symptoms among college students: prevalence, persistence, correlates, and treatment-seeking. J Am Coll Health 2011 Nov;59(8):700-707 [FREE Full text] [doi: 10.1080/07448481.2010.546461] [Medline: 21950250]
3. Lives at risk: the state of NHS adult community eating disorder services in England. BEAT. 2019. URL: https://www.beateatingdisorders.org.uk/about-beat/policy-work/policy-and-best-practice-reports/lives-at-risk-the-state-of-nhs-adult-services/ [accessed 2024-04-29]
4. Ali K, Farrer L, Fassnacht DB, Gulliver A, Bauer S, Griffiths KM. Perceived barriers and facilitators towards help-seeking for eating disorders: a systematic review. Int J Eat Disord 2017 Jan;50(1):9-21. [doi: 10.1002/eat.22598] [Medline: 27526643]
5. Arcelus J, Mitchell AJ, Wales J, Nielsen S. Mortality rates in patients with anorexia nervosa and other eating disorders: a meta-analysis of 36 studies. Arch Gen Psychiatry 2011 Jul;68(7):724-731. [doi: 10.1001/archgenpsychiatry.2011.74] [Medline: 21727255]
6. Steinhausen HC. Outcome of eating disorders. Child Adolesc Psychiatr Clin N Am 2009 Jan;18(1):225-242. [doi: 10.1016/j.chc.2008.07.013] [Medline: 19014869]
7. Katzman DK. The COVID-19 pandemic and eating disorders: a wake-up call for the future of eating disorders among adolescents and young adults. J Adolesc Health 2021 Oct;69(4):535-537 [FREE Full text] [doi: 10.1016/j.jadohealth.2021.07.014] [Medline: 34489056]
8. Kazdin AE, Fitzsimmons-Craft EE, Wilfley DE. Addressing critical gaps in the treatment of eating disorders. Int J Eat Disord 2017 Mar;50(3):170-189 [FREE Full text] [doi: 10.1002/eat.22670] [Medline: 28102908]
9. Andersson G. Internet-delivered psychological treatments. Annu Rev Clin Psychol 2016;12:157-179. [doi: 10.1146/annurev-clinpsy-021815-093006] [Medline: 26652054]

10. Green M, Kroska A, Herrick A, Bryant B, Sage E, Miles L, et al. A preliminary trial of an online dissonance-based eating disorder intervention. Eat Behav 2018 Dec;31:88-98. [doi: 10.1016/j.eatbeh.2018.08.007] [Medline: 30199771]
11. Linardon J, Messer M, Lee S, Rosato J. Perspectives of e-health interventions for treating and preventing eating disorders: descriptive study of perceived advantages and barriers, help-seeking intentions, and preferred functionality. Eat Weight Disord 2021 May;26(4):1097-1109. [doi: 10.1007/s40519-020-01005-3] [Medline: 32959274]
12. Torous J, Bucci S, Bell IH, Kessing LV, Faurholt-Jepsen M, Whelan P, et al. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 2021 Oct;20(3):318-335 [FREE Full text] [doi: 10.1002/wps.20883] [Medline: 34505369]
13. Ruwaard J, Lange A, Broeksteeg J, Renteria-Agirre A, Schrieken B, Dolan CV, et al. Online cognitive-behavioural treatment of bulimic symptoms: a randomized controlled trial. Clin Psychol Psychother 2013;20(4):308-318. [doi: 10.1002/cpp.1767] [Medline: 22298417]
14. Linardon J, Shatte A, Rosato J, Fuller-Tyszkiewicz M. Efficacy of a transdiagnostic cognitive-behavioral intervention for eating disorder psychopathology delivered through a smartphone app: a randomized controlled trial. Psychol Med 2022 Jul;52(9):1679-1690. [doi: 10.1017/S0033291720003426] [Medline: 32972467]
15. Carrard I, Crépin C, Rouget P, Lam T, Golay A, van der Linden M. Randomised controlled trial of a guided self-help treatment on the internet for binge eating disorder. Behav Res Ther 2011 Aug;49(8):482-491. [doi: 10.1016/j.brat.2011.05.004] [Medline: 21641580]
16. de Zwaan M, Herpertz S, Zipfel S, Svaldi J, Friederich H, Schmidt F, et al. Effect of internet-based guided self-help vs individual face-to-face treatment on full or subsyndromal binge eating disorder in overweight or obese patients: the INTERBED randomized clinical trial. JAMA Psychiatry 2017 Oct 01;74(10):987-995 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.2150] [Medline: 28768334]
17. Ahmadiankalati M, Steins-Loeber S, Paslakis G. Review of randomized controlled trials using e-Health interventions for patients with eating disorders. Front Psychiatry 2020 Jun 12;11:568 [FREE Full text] [doi: 10.3389/fpsyt.2020.00568] [Medline: 32595546]
18. Yardley L, Morrison L, Bradbury K, Muller I. The person-based approach to intervention development: application to digital health-related behavior change interventions. J Med Internet Res 2015 Jan 30;17(1):e30 [FREE Full text] [doi: 10.2196/jmir.4055] [Medline: 25639757]
19. Davis R, Campbell R, Hildon Z, Hobbs L, Michie S. Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review. Health Psychol Rev 2015;9(3):323-344 [FREE Full text] [doi: 10.1080/17437199.2014.941722] [Medline: 25104107]
20. Ogden J. Celebrating variability and a call to limit systematisation: the example of the Behaviour Change Technique Taxonomy and the Behaviour Change Wheel. Health Psychol Rev 2016 Sep;10(3):245-250. [doi: 10.1080/17437199.2016.1190291] [Medline: 27189585]
21. Muntaner A, Vidal-Conti J, Palou P. Increasing physical activity through mobile device interventions: a systematic review. Health Informatics J 2016 Sep;22(3):451-469 [FREE Full text] [doi: 10.1177/1460458214567004] [Medline: 25649783]
22. Tombor I, Shahab L, Brown J, Crane D, Michie S, West R. Development of SmokeFree baby: a smoking cessation smartphone app for pregnant smokers. Transl Behav Med 2016 Dec;6(4):533-545 [FREE Full text] [doi: 10.1007/s13142-016-0438-0] [Medline: 27699682]
23. Cradock KA, ÓLaighin G, Finucane FM, Gainforth HL, Quinlan LR, Ginis KA. Behaviour change techniques targeting both diet and physical activity in type 2 diabetes: a systematic review and meta-analysis. Int J Behav Nutr Phys Act 2017 Feb 08;14(1):18 [FREE Full text] [doi: 10.1186/s12966-016-0436-0] [Medline: 28178985]
24. Samdal GB, Eide GE, Barth T, Williams G, Meland E. Effective behaviour change techniques for physical activity and healthy eating in overweight and obese adults; systematic review and meta-regression analyses. Int J Behav Nutr Phys Act 2017 Mar 28;14(1):42 [FREE Full text] [doi: 10.1186/s12966-017-0494-y] [Medline: 28351367]
25. van Rhoon L, Byrne M, Morrissey E, Murphy J, McSharry J. A systematic review of the behaviour change techniques and digital features in technology-driven type 2 diabetes prevention interventions. Digit Health 2020;6:2055207620914427 [FREE Full text] [doi: 10.1177/2055207620914427] [Medline: 32269830]
26. Hailey V, Rojas-Garcia A, Kassianos AP. A systematic review of behaviour change techniques used in interventions to increase physical activity among breast cancer survivors. Breast Cancer 2022 Mar 06;29(2):193-208 [FREE Full text] [doi: 10.1007/s12282-021-01323-z] [Medline: 34989962]
27. Carraça E, Encantado J, Battista F, Beaulieu K, Blundell J, Busetto L, et al. Effective behavior change techniques to promote physical activity in adults with overweight or obesity: a systematic review and meta-analysis. Obes Rev 2021 Jul;22 Suppl 4(Suppl 4):e13258 [FREE Full text] [doi: 10.1111/obr.13258] [Medline: 33949778]
28. Howlett N, Trivedi D, Troop NA, Chater AM. Are physical activity interventions for healthy inactive adults effective in promoting behavior change and maintenance, and which behavior change techniques are effective? a systematic review and meta-analysis. Transl Behav Med 2019 Jan 01;9(1):147-157 [FREE Full text] [doi: 10.1093/tbm/iby010] [Medline: 29506209]
29. Dugdale S, Ward J, Hernen J, Elison S, Davies G, Donkor D. Using the behavior change technique taxonomy v1 to conceptualize the clinical content of breaking free online: a computer-assisted therapy program for substance use disorders.

Subst Abuse Treat Prev Policy 2016 Jul 22;11(1):26 [FREE Full text] [doi: 10.1186/s13011-016-0069-y] [Medline: 27449786]
30. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. Ann Behav Med 2013 Aug;46(1):81-95 [FREE Full text] [doi: 10.1007/s12160-013-9486-6] [Medline: 23512568]
31. Howick J. OCEBM levels of evidence. The Oxford Centre for Evidence-Based Medicine (CEBM). 2011. URL: http://wwwcebmnet/indexaspx?o=5653 [accessed 2024-04-29]
32. Pennesi J, Wade TD. A systematic review of the existing models of disordered eating: do they inform the development of effective interventions? Clin Psychol Rev 2016 Feb;43:175-192. [doi: 10.1016/j.cpr.2015.12.004] [Medline: 26781985]
33. Barak A, Klein B, Proudfoot JG. Defining internet-supported therapeutic interventions. Ann Behav Med 2009 Aug;38(1):4-17. [doi: 10.1007/s12160-009-9130-7] [Medline: 19787305]
34. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J Clin Epidemiol 2009 Oct;62(10):e1-34 [FREE Full text] [doi: 10.1016/j.jclinepi.2009.06.006] [Medline: 19631507]
35. Linardon J, Shatte A, Messer M, Firth J, Fuller-Tyszkiewicz M. E-mental health interventions for the treatment and prevention of eating disorders: an updated systematic review and meta-analysis. J Consult Clin Psychol 2020 Nov;88(11):994-1007. [doi: 10.1037/ccp0000575] [Medline: 32852971]
36. Lefebvre C, Glanville J, Briscoe S, Featherstone R, Littlewood A, Metzendorf MI, et al. Searching for and selecting studies. In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions. London, UK: The Cochrane Collaboration; 2019.
37. Fairburn CG, Beglin S. Eating disorder examination questionnaire. In: Fairburn CG, editor. Cognitive Behavior Therapy and Eating Disorders. New York, NY: The Guilford Press; 2008:309-313.
38. Michie S. Is cognitive behaviour therapy effective for changing health behaviours? commentary on Hobbis and Sutton. J Health Psychol 2005 Jan 01;10(1):33-37. [doi: 10.1177/1359105305048553] [Medline: 15576498]
39. Wood CE, Hardeman W, Johnston M, Francis J, Abraham C, Michie S. Reporting behaviour change interventions: do the behaviour change technique taxonomy v1, and training in its use, improve the quality of intervention descriptions? Implement Sci 2016 Jun 07;11(1):84 [FREE Full text] [doi: 10.1186/s13012-016-0448-9] [Medline: 27268131]
40. Abraham C, Wood CE, Johnston M, Francis J, Hardeman W, Richardson M, et al. Reliability of identification of behavior change techniques in intervention descriptions. Ann Behav Med 2015 Dec;49(6):885-900. [doi: 10.1007/s12160-015-9727-y] [Medline: 26290002]
41. Michie S, Wood CE, Johnston M, Abraham C, Francis JJ, Hardeman W. Behaviour change techniques: the development and evaluation of a taxonomic method for reporting and describing behaviour change interventions (a suite of five studies involving consensus methods, randomised controlled trials and analysis of qualitative data). Health Technol Assess 2015 Nov;19(99):1-188 [FREE Full text] [doi: 10.3310/hta19990] [Medline: 26616119]
42. Marques MM, Carey RN, Norris E, Evans F, Finnerty AN, Hastings J, et al. Delivering behaviour change interventions: development of a mode of delivery ontology. Wellcome Open Res 2020;5:125 [FREE Full text] [doi: 10.12688/wellcomeopenres.15906.2] [Medline: 33824909]
43. Michie S, Prestwich A. Are interventions theory-based? development of a theory coding scheme. Health Psychol 2010 Jan;29(1):1-8 [FREE Full text] [doi: 10.1037/a0016939] [Medline: 20063930]
44. Hobbis IC, Sutton S. Are techniques used in cognitive behaviour therapy applicable to behaviour change interventions based on the theory of planned behaviour? J Health Psychol 2005 Jan 01;10(1):7-37. [doi: 10.1177/1359105305048549] [Medline: 15576494]
45. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synth Methods 2010 Apr;1(2):97-111. [doi: 10.1002/jrsm.12] [Medline: 26061376]
46. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 06;327(7414):557-560 [FREE Full text] [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
47. Berry R, Kassavou A, Sutton S. Does self-monitoring diet and physical activity behaviors using digital technology support adults with obesity or overweight to lose weight? a systematic literature review with meta-analysis. Obes Rev 2021 Oct;22(10):e13306. [doi: 10.1111/obr.13306] [Medline: 34192411]
48. Aardoom JJ, Dingemans AE, Spinhoven P, van Ginkel JR, de Rooij M, van Furth EF. Web-based fully automated self-help with different levels of therapist support for individuals with eating disorder symptoms: a randomized controlled trial. J Med Internet Res 2016 Jun 17;18(6):e159 [FREE Full text] [doi: 10.2196/jmir.5709] [Medline: 27317358]
49. Rohrbach PJ, Dingemans AE, Spinhoven P, van Ginkel JR, Fokkema M, Wilderjans TF, et al. Effectiveness of an online self-help program, expert-patient support, and their combination for eating disorders: Results from a randomized controlled trial. Int J Eat Disord 2022 Oct;55(10):1361-1373 [FREE Full text] [doi: 10.1002/eat.23785] [Medline: 35906929]
50. Fitzsimmons-Craft EE, Taylor CB, Graham AK, Sadeh-Sharvit S, Balantekin KN, Eichen DM, et al. Effectiveness of a digital cognitive behavior therapy-guided self-help intervention for eating disorders in college women: a cluster randomized

clinical trial. JAMA Netw Open 2020 Aug 03;3(8):e2015633 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.15633] [Medline: 32865576]
51. Linardon J, Shatte A, McClure Z, Fuller-Tyszkiewicz M. A broad focused digital intervention for recurrent binge eating: a randomized controlled non-inferiority trial. Psychol Med 2023 Jul 27;53(10):4580-4591 [FREE Full text] [doi: 10.1017/S0033291722001477] [Medline: 35621217]
52. Wyssen A, Meyer AH, Messerli-Bürgy N, Forrer F, Vanhulst P, Lalanne D, et al. BED-online: acceptance and efficacy of an internet-based treatment for binge-eating disorder: a randomized clinical trial including waitlist conditions. Eur Eat Disord Rev 2021 Nov;29(6):937-954 [FREE Full text] [doi: 10.1002/erv.2856] [Medline: 34418221]
53. Melisse B, Berg EV, Jonge MD, Blankers M, Furth EV, Dekker J, et al. Efficacy of web-based, guided self-help cognitive behavioral therapy-enhanced for binge eating disorder: randomized controlled trial. J Med Internet Res 2023 May 01;25:e40472 [FREE Full text] [doi: 10.2196/40472] [Medline: 37126386]
54. Jacobi C, Völker U, Trockel MT, Taylor CB. Effects of an internet-based intervention for subthreshold eating disorders: a randomized controlled trial. Behav Res Ther 2012 Feb;50(2):93-99. [doi: 10.1016/j.brat.2011.09.013] [Medline: 22137366]
55. Strandskov SW, Ghaderi A, Andersson H, Parmskog N, Hjort E, Wärn AS, et al. Effects of tailored and ACT-influenced internet-based CBT for eating disorders and the relation between knowledge acquisition and outcome: a randomized controlled trial. Behav Ther 2017 Sep;48(5):624-637. [doi: 10.1016/j.beth.2017.02.002] [Medline: 28711113]
56. Melioli T, Bauer S, Franko DL, Moessner M, Ozer F, Chabrol H, et al. Reducing eating disorder symptoms and risk factors using the internet: a meta-analytic review. Int J Eat Disord 2016 Jan;49(1):19-31. [doi: 10.1002/eat.22477] [Medline: 26607683]
57. Fairburn CG. Cognitive Behavior Therapy and Eating Disorders. New York, NY: The Guilford Press; 2008.
58. Fairburn CG, Cooper Z, Shafran R. Cognitive behaviour therapy for eating disorders: a "transdiagnostic" theory and treatment. Behav Res Ther 2003 May;41(5):509-528. [doi: 10.1016/s0005-7967(02)00088-8] [Medline: 12711261]
59. Higgins JP, Savović J, Page MJ, Elbers RG, Sterne JA. Assessing risk of bias in a randomized trial. In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. Cochrane Handbook for Systematic Reviews of Interventions. London, UK: The Cochrane Collaboration; 2019:205-228.
60. Linardon J, Messer M, Shatte A, Skvarc D, Rosato J, Rathgen A, et al. Targeting dietary restraint to reduce binge eating: a randomised controlled trial of a blended internet- and smartphone app-based intervention. Psychol Med 2023 Mar;53(4):1277-1287. [doi: 10.1017/S0033291721002786] [Medline: 34247660]
61. Högdahl L, Birgegård A, Norring C, de Man Lapidoth J, Franko MA, Björck C. Internet-based cognitive behavioral therapy for bulimic eating disorders in a clinical setting: results from a randomized trial with one-year follow-up. Internet Interv 2023 Mar;31:100598 [FREE Full text] [doi: 10.1016/j.invent.2022.100598] [Medline: 36588668]
62. Tregarthen J, Paik Kim J, Sadeh-Sharvit S, Neri E, Welch H, Lock J. Comparing a tailored self-help mobile app with a standard self-monitoring app for the treatment of eating disorder symptoms: randomized controlled trial. JMIR Ment Health 2019 Nov 21;6(11):e14972 [FREE Full text] [doi: 10.2196/14972] [Medline: 31750837]
63. Linardon J, Messer M, Shatte A, Greenwood CJ, Rosato J, Rathgen A, et al. Does the method of content delivery matter? randomized controlled comparison of an internet-based intervention for eating disorder symptoms with and without interactive functionality. Behav Ther 2022 May;53(3):508-520. [doi: 10.1016/j.beth.2021.12.001] [Medline: 35473653]
64. Wagner G, Penelo E, Wanner C, Gwinner P, Trofaier M, Imgart H, et al. Internet-delivered cognitive-behavioural therapy v. conventional guided self-help for bulimia nervosa: long-term evaluation of a randomised controlled trial. Br J Psychiatry 2013 Feb;202:135-141. [doi: 10.1192/bjp.bp.111.098582] [Medline: 23222037]
65. Högdahl L, Levallius J, Björck C, Norring C, Birgegård A. Personality predicts drop-out from therapist-guided internet-based cognitive behavioural therapy for eating disorders. Results from a randomized controlled trial. Internet Interv 2016 Sep;5:44-50 [FREE Full text] [doi: 10.1016/j.invent.2016.07.002] [Medline: 30135806]
66. Loucas CE, Fairburn CG, Whittington C, Pennant ME, Stockton S, Kendall T. E-therapy in the treatment and prevention of eating disorders: a systematic review and meta-analysis. Behav Res Ther 2014 Dec;63:122-131 [FREE Full text] [doi: 10.1016/j.brat.2014.09.011] [Medline: 25461787]
67. Antezana G, Venning A, Blake V, Smith D, Winsall M, Orlowski S, et al. An evaluation of behaviour change techniques in health and lifestyle mobile applications. Health Informatics J 2020 Mar;26(1):104-113 [FREE Full text] [doi: 10.1177/1460458218813726] [Medline: 30501364]
68. Roberts AL, Fisher A, Smith L, Heinrich M, Potts HW. Digital health behaviour change interventions targeting physical activity and diet in cancer survivors: a systematic review and meta-analysis. J Cancer Surviv 2017 Dec;11(6):704-719 [FREE Full text] [doi: 10.1007/s11764-017-0632-1] [Medline: 28779220]
69. Alkhaldi G, Hamilton FL, Lau R, Webster R, Michie S, Murray E. The effectiveness of prompts to promote engagement with digital interventions: a systematic review. J Med Internet Res 2016 Jan 08;18(1):e6 [FREE Full text] [doi: 10.2196/jmir.4790] [Medline: 26747176]
70. Collins LM, Murphy SA, Strecher V. The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): new methods for more potent eHealth interventions. Am J Prev Med 2007 May;32(5 Suppl):S112-S118 [FREE Full text] [doi: 10.1016/j.amepre.2007.01.022] [Medline: 17466815]

71. Aardoom JJ, Dingemans AE, Spinhoven P, van Furth EF. Treating eating disorders over the internet: a systematic review and future research directions. Int J Eat Disord 2013 Sep;46(6):539-552. [doi: 10.1002/eat.22135] [Medline: 23674367]
72. Sonneville KR, Lipson SK. Disparities in eating disorder diagnosis and treatment according to weight status, race/ethnicity, socioeconomic background, and sex among college students. Int J Eat Disord 2018 Jun 02;51(6):518-526 [FREE Full text] [doi: 10.1002/eat.22846] [Medline: 29500865]
73. Anastasiadou D, Folkvord F, Lupiañez-Villanueva F. A systematic review of mHealth interventions for the support of eating disorders. Eur Eat Disord Rev 2018 Sep;26(5):394-416. [doi: 10.1002/erv.2609] [Medline: 29927004]
74. Linardon J, Shatte A, Tepper H, Fuller-Tyszkiewicz M. A survey study of attitudes toward, and preferences for, e-therapy interventions for eating disorder psychopathology. Int J Eat Disord 2020 Jun 02;53(6):907-916. [doi: 10.1002/eat.23268] [Medline: 32239725]
75. Juarascio A, Srivastava P, Presseller E, Clark K, Manasse S, Forman E. A clinician-controlled just-in-time adaptive intervention system (CBT+) designed to promote acquisition and utilization of cognitive behavioral therapy skills in bulimia nervosa: development and preliminary evaluation study. JMIR Form Res 2021 May 31;5(5):e18261 [FREE Full text] [doi: 10.2196/18261] [Medline: 34057416]
76. Barakat S, Maguire S, Smith KE, Mason TB, Crosby RD, Touyz S. Evaluating the role of digital intervention design in treatment outcomes and adherence to eTherapy programs for eating disorders: a systematic review and meta-analysis. Int J Eat Disord 2019 Oct;52(10):1077-1094. [doi: 10.1002/eat.23131] [Medline: 31328815]
77. Fairburn CG, Bailey-Straebler S, Basden S, Doll HA, Jones R, Murphy R, et al. A transdiagnostic comparison of enhanced cognitive behaviour therapy (CBT-E) and interpersonal psychotherapy in the treatment of eating disorders. Behav Res Ther 2015 Jul;70:64-71 [FREE Full text] [doi: 10.1016/j.brat.2015.04.010] [Medline: 26000757]
78. Berg KC, Peterson CB, Frazier P, Crow SJ. Convergence of scores on the interview and questionnaire versions of the Eating Disorder Examination: a meta-analytic review. Psychol Assess 2011 Sep;23(3):714-724 [FREE Full text] [doi: 10.1037/a0023246] [Medline: 21517194]
79. Sadeh-Sharvit S, Kim JP, Darcy AM, Neri E, Vierhile M, Robinson A, et al. Subgrouping the users of a specialized app for eating disorders. Eat Disord 2018;26(4):361-372. [doi: 10.1080/10640266.2018.1440043] [Medline: 29452025]
80. Andersson G, Estling F, Jakobsson E, Cuijpers P, Carlbring P. Can the patient decide which modules to endorse? an open trial of tailored internet treatment of anxiety disorders. Cogn Behav Ther 2011;40(1):57-64. [doi: 10.1080/16506073.2010.529457] [Medline: 21337215]
81. Marques MM, Wright AJ, Corker E, Johnston M, West R, Hastings J, et al. The behaviour change technique ontology: transforming the behaviour change technique taxonomy v1. Wellcome Open Res 2023;8:308 [FREE Full text] [doi: 10.12688/wellcomeopenres.19363.1] [Medline: 37593567]
82. Spek V, Cuijpers P, Nyklícek I, Riper H, Keyzer J, Pop V. Internet-based cognitive behaviour therapy for symptoms of depression and anxiety: a meta-analysis. Psychol Med 2007 Mar;37(3):319-328. [doi: 10.1017/S0033291706008944] [Medline: 17112400]
83. Andersson G, Cuijpers P. Internet-based and other computerized psychological treatments for adult depression: a meta-analysis. Cogn Behav Ther 2009;38(4):196-205. [doi: 10.1080/16506070903318960] [Medline: 20183695]
84. Rohrbach PJ, Dingemans AE, Evers C, van Furth EF, Spinhoven P, Aardoom JJ, et al. Cost-effectiveness of internet interventions compared with treatment as usual for people with mental disorders: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2023 Jan 05;25:e38204 [FREE Full text] [doi: 10.2196/38204] [Medline: 36602854]
85. Martin DJ, Garske JP, Davis MK. Relation of the therapeutic alliance with outcome and other variables: a meta-analytic review. J Consult Clin Psychol 2000 Jun;68(3):438-450. [doi: 10.1037/0022-006X.68.3.438] [Medline: 10883561]
86. Knaevelsrud C, Maercker A. Internet-based treatment for PTSD reduces distress and facilitates the development of a strong therapeutic alliance: a randomized controlled clinical trial. BMC Psychiatry 2007 Apr 19;7:13 [FREE Full text] [doi: 10.1186/1471-244X-7-13] [Medline: 17442125]
87. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. Front Digit Health 2022;4:847991 [FREE Full text] [doi: 10.3389/fdgth.2022.847991] [Medline: 35480848]
88. O'Leary T, Torous J. Smartphone apps for eating disorders: an overview of the marketplace and research trends. Int J Eat Disord 2022 May 17;55(5):625-632. [doi: 10.1002/eat.23690] [Medline: 35175640]
89. Graham AK, Fitzsimmons-Craft EE. Accelerating the research-to-practice translation of eating disorder apps and other digital interventions: commentary on O'Leary and Torous (2022). Int J Eat Disord 2022 Nov;55(11):1635-1638 [FREE Full text] [doi: 10.1002/eat.23811] [Medline: 36111655]

Abbreviations

ACT: acceptance commitment therapy
BCT: behavior change technique
CBT: cognitive behavioral therapy

ED: eating disorder
EDE-Q: Eating Disorder Examination Questionnaire
F2F: face-to-face
IRR: interrater reliability
MD: mean difference
MeSH: Medical Subject Headings
mHealth: mobile health
MOA: mechanism of action
OBE: objective binge episode
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
TAU: treatment as usual
TCS: theory coding scheme
WL: waiting list

Edited by J Torous; submitted 20.02.24; peer-reviewed by E Neter, H Oh, R Hawkes; comments to author 09.05.24; revised version received 23.05.24; accepted 25.05.24; published 01.08.24.

Please cite as:
Thomas PC, Curtis K, Potts HWW, Bark P, Perowne R, Rookes T, Rowe S
Behavior Change Techniques Within Digital Interventions for the Treatment of Eating Disorders: Systematic Review and Meta-Analysis
JMIR Ment Health 2024;11:e57577
URL: https://mental.jmir.org/2024/1/e57577
doi:10.2196/57577
PMID:

Outcomes of Providing Children Aged 7-12 Years With Access to Evidence-Based Anxiety Treatment Via a Standalone Digital Intervention Using Immersive Gaming Technology: Real-World Evaluation

Brioney Gee[1], PhD; Bonnie Teague[1,2], PhD; Andrew Laphan[2], MSc; Tim Clarke[1,2], ClinPsyD; Georgianna Coote[3], MSc; Jessica Garner[3], MSc; Jon Wilson[1,2], MD

1

2

3

Corresponding Author:
Bonnie Teague, PhD

Abstract

Background: Anxiety disorders are among the most common mental health conditions in childhood, but most children with anxiety disorders do not access evidence-based interventions. The delivery of therapeutic interventions via digital technologies has been proposed to significantly increase timely access to evidence-based treatment. Lumi Nova (BfB Labs Limited) is a digital therapeutic intervention designed to deliver evidence-based anxiety treatment for those aged 7 - 12 years through a mobile app incorporating immersive gaming technology.

Objective: We aimed to evaluate the real-world impact of providing access to Lumi Nova through UK National Health Service–funded mental health services.

Methods: We analyzed precollected anonymized data routinely captured through the implementation of Lumi Nova from children aged 7 - 12 years, who lived in the United Kingdom and had the opportunity to use the intervention for at least 1 week over an 18-month period. Engagement indices included whether the game key was activated, number of unique sessions, time spent engaging, and number of “challenges” completed. Clinical outcomes were assessed using the Goal-Based Outcomes measure and Child Outcome Rating Scale. Demographic data were analyzed to assess the health equality implications of Lumi Nova.

Results: Of 1029 eligible families invited to use Lumi Nova, 644 (62.5%) activated their game key, of whom 374 (58.1%) completed at least one in-game graded exposure challenge. The median number of unique sessions was 6 (IQR 3 - 12) and the median time spent engaging with the intervention was 42 (IQR 15 - 79) minutes. For the subset of young people with paired outcomes, there were statistically significant small to medium improvements in goal-based outcome scores (n=224; t_{223}=5.78, P<.001; d=0.37, 95% CI 0.25 - 0.52) and Child Outcome Rating Scale scores (n=123; t_{122}=5.10, P<.001; d=0.46, 95% CI 0.27 - 0.65) between the first and last data points. Two in 5 young people’s scores reflected a change that would be considered reliable. Analysis of demographic characteristics tentatively suggested that children from ethnic minority backgrounds and those living in the most deprived neighbourhoods may be less likely to access Lumi Nova, but children from socioeconomically deprived areas were more likely to successfully complete a challenge once they accessed the intervention (P=.02). However, the level of missing data and small number of children in some demographic groups limited meaningful statistical comparisons.

Conclusions: This study provides initial evidence that Lumi Nova may be associated with improved outcomes for those aged 7 - 12 years seeking anxiety treatment in real-world settings. However, the lack of a control comparator group and information about concurrent treatments accessed by the young people, in addition to substantial attrition, limited the analysis that could be conducted and confidence in the conclusions drawn.

(JMIR Ment Health 2024;11:e52866) doi:10.2196/52866

KEYWORDS

anxiety; children; young people; exposure therapy; graded exposures; cognitive behavioural therapy; digital intervention; mobile app; gaming; real-world evaluation; gaming technology; real-world implementation

Introduction

Access to Evidence-Based Treatment for Children Experiencing Mental Health Problems

Evidence suggests that levels of mental ill-health among children and young people are rising, exacerbated in recent years by the COVID-19 pandemic [1]. In England, 16.7% of children aged 5-6 years were identified as having a probable mental health disorder in 2022, increasing from 12.1% in 2017 [2]. Providing timely access to evidence-based treatment to the growing number of children and young people seeking care from mental health services poses an urgent public policy challenge [3]. Given significant shortages in the mental health workforce [4], innovative approaches are required to meet this challenge.

The delivery of therapeutic interventions via digital technologies has been proposed as 1 innovation with the potential to significantly increase access to evidence-based treatment for young people [5]. Systematic reviews examining the effectiveness of internet-based or digital cognitive behavioral therapy (CBT) support the view that these digital health interventions are effective and feasible to treat depression and anxiety in children and young people, although the quality and depth of the research field has been questioned [6-8]. Meta-analytic evidence suggests that digital CBT programmes can produce similar outcomes to face-to-face therapy [9,10]. Further, given that many children are comfortable and familiar with using digital devices (eg smartphones and tablets), young people may find this mode of delivery particularly appealing [11].

However, this research evidence has not established if digital therapies based on CBT work for all young people equitably. For example, children from disadvantaged and marginalized groups are more likely to experience poor mental health than their peers [12,13], while also being disproportionately affected by obstacles to accessing health care [14]. There is potential for digital interventions to promote equity in mental health outcomes by reducing disparities in access to care between young people of different backgrounds [15]. However, it is also important to remain cognisant of the potential for new technologies to exacerbate existing health inequalities [16] and to ensure that any such impacts can be mitigated against.

Of particular focus to this paper, there is limited evidence for the use of digital health interventions for young people's anxiety. Anxiety disorders are among the most common mental health conditions in childhood. Prevalence rates from 2003 - 2020 were estimated to be around 5% in high-income countries before the COVID-19 pandemic [17], and appear to have increased following the pandemic [18]. Childhood anxiety disorders can have significant negative impact on educational attainment, peer relationships, and family life, and often persist into adulthood if left untreated [19]. However, at present, most children with anxiety disorders do not access evidence-based interventions. In a UK study, of the 65% of families of children with anxiety who sought support, only 38% accessed support and less than 3% accessed an evidence-based treatment [20]. While many mobile-based applications designed to help children with anxiety have been made available on consumer marketplaces, very few are underpinned by core therapeutic principles for the treatment of anxiety disorders or have been subject to empirical evaluation [21]. This paper will be examining one of these digital applications, Lumi Nova, designed to support young people's anxiety in real-life clinical settings.

Lumi Nova: Tales of Courage

Lumi Nova: Tales of Courage ("Lumi Nova") is a standalone digital therapeutic intervention designed to deliver evidence-based anxiety treatment for those aged 7 - 12 years [22]. It provides graded exposure therapy (a key active ingredient of CBT for anxiety [23]) and psychoeducation via an immersive, engaging app-based mobile game. Lumi Nova was launched by BfB Labs in September 2020 following a 3-year development process led by a human centred design approach. This process involved engagement with academics, clinicians, service managers, commissioners, games experts, education professionals, parents, and children. The aim of these development phases was to ensure the intervention has both a robust clinical basis and high-end functionality, incorporating the latest advances in immersive gaming technology to increase voluntary engagement. A preliminary evaluation found that Lumi Nova is safe and has the potential to benefit children experiencing mild to moderate anxiety [24]. Current guidance for the use of Lumi Nova in the home recommends that young people use the intervention for up to 30 - 40 minutes per day and are asked to select 3 goals that they would prefer to work on. These goals include working on personalized challenges selected by the young person, such as "be comfortable speaking in a group," "be able to sleep away from home overnight," or "feel comfortable going to school" [25]

The young person uses the intervention on their own, although they are supported by their parent or guardian to access the intervention on a device, who is notified when the young person accesses the intervention. In practice, the number of sessions and frequency of use is determined by the young person. Demographic, usage, and outcome data from 2 measures is collected on an embedded secure digital platform.

Young people presenting to the UK National Health Services (NHS) with anxiety are offered the opportunity to use Lumi Nova after initial screening in person or via telephone, or during later treatment appointments. Some families are supported to access the intervention by a practitioner, while others are given access via a promoted "self-sign-up" option (implemented from April 2022 onwards) which allows the child's parent or caregiver to access the intervention independently.

Young people are given access to the game either as a stand-alone treatment for mild to moderate anxiety, alongside other therapeutic support, or while on a waitlist for further support.

Lumi Nova is currently in use by providers of children and young people's mental health services across multiple regions of the United Kingdom, including both NHS and voluntary sector providers. The software is CE (Conformité Européene) marked and registered with the UK's Medicines and Healthcare Products Regulatory Agency as a medical device (class 1—low risk). Lumi Nova has recently been recommended by the National Institute for Health and Care Excellence in their Early Value Assessment of digital CBT for children and young people with mild to moderate symptoms of anxiety or low mood [26]. This assessment recommended that Lumi Nova be used as a first line treatment option for children and young people with mild to moderate anxiety difficulties but suggested that this usage should be accompanied by further evidence generation efforts.

Objectives

The current evaluation begins to address this National Institute for Health and Care Excellence recommendation to generate further evidence by using routinely captured engagement and outcome data to better understand how Lumi Nova has been used in real-world settings. Using this data, we were able to investigate (1) the rate and degree of engagement with Lumi Nova by young people who were provided with access to the intervention as part of their routine care over an 18-month period, (2) the clinical outcomes for young people provided with access to Lumi Nova over this period, and (3) whether engagement or outcomes for young people from disadvantaged or marginalized groups differ from those of their peers.

Methods

Design

We conducted a real-world evaluation of Lumi Nova using precollected quantitative data routinely captured by BfB Labs. User engagement and outcome data are collected by BfB Labs for the purposes of enabling families and professionals responsible for the care of young people using Lumi Nova to monitor the child's progress via the product's dashboard, as well as to enable evaluation and iterative improvement of their products and services. Relevant data were exported from the product's dashboard and anonymized by a member of BfB Labs team before being securely transferred to the research team for analysis.

This evaluation was part of a wider study investigating the utility of digital therapeutic interventions within NHS-funded children and young people's mental health services (the Investigating Digital Therapy study). This study aimed to provide insights into the potential for digital technologies to improve access to evidence-based treatment, to explore possible impacts on health inequalities, and identify factors that impact the commissioning and implementation of new digital technologies within the sector.

Participants

Young people whose precollected data were eligible for inclusion in the analysis were those who (1) had been referred by practitioners to use Lumi Nova during an 18-month period from July 2021 (when the current outcome measures were implemented) to December 2022, (2) had the opportunity to use the intervention for at least one week, (3) were aged 7 - 12 years, and (4) who were residing in the United Kingdom. No additional or new data was collected for analysis.

Measures

Demographic information collected from Lumi Nova users included the child's age, gender, ethnicity, disability status, and home postcode. Postcodes were used to classify the relative deprivation of individuals living in that area using the indices of multiple deprivation, which categorizes small areas of England, Wales, and Scotland from most (1) to least (10) deprived based on a combination of 37 indicators [27].

User data collected to provide a measure of engagement included: whether each family invited to access the game went on to activate their game key, the number of unique sessions completed, the amount of time spent engaging with the intervention (total length of sessions minus any idle time), and the number of "challenges" completed as part of the game (including both in-application challenges and real-world challenges completed outside of the application and verified as complete by the child's parent or caregiver).

Clinical outcome data was collected using the goal-based outcomes (GBOs [28]) measure and the Child Outcome Rating Scale (CORS [29]). GBOs are a method of quantifying progress toward the goals a young person sets for themselves at the beginning of a clinical intervention. Scoring range is 0 - 10, with higher scores indicating greater perceived progress toward the selected goals. Children were asked to complete this measure within the application after selecting their goals for the first time and then on a weekly basis (if the child logs into the application) before completing a challenge.

The CORS is a 4-item parent-reported measure developed to enable monitoring of a child's symptom distress, interpersonal well-being, social role functioning, and overall well-being on a regular basis during the process of therapeutic interventions. Scoring range is 0 - 40, with higher scores indicating better functioning. The child's parent or caregiver was first asked to complete the CORS during the process of accessing the application, before being provided with a game key. The child's parent or caregiver was subsequently prompted to repeat the measure weekly via a link sent to their mobile using SMS messaging; if the measure was not completed, 2 reminders were sent via SMS.

Young people were also asked to rate how anxious they felt before the challenge and, after completing the challenge, how anxious they felt during the challenge, and how anxious they would feel if they had to complete the challenge again on a 1 to 5 scale 5 being most anxious) via in-game ratings. Ratings of anxiety before and during the challenge were used as an indication of mismatch in anticipated and experienced anxiety, which has been posited as an important mechanism through which graded exposure therapy brings about reductions in anxiety symptoms [23].

Data Analysis

Descriptive statistics were calculated for all demographic characteristics, engagement indices, and outcome measures. After confirming all relevant assumptions were met, paired sample *t* tests were conducted to assess whether changes in outcome measures were statistically significant at the P=.05 level (GBOs were treated as the primary outcome and corrections for multiple comparisons used for other measures). Standardized effect sizes (Cohen *d*) and their 95% CIs were calculated to quantify the size of statistically significant differences. Only cases for which completed outcome measures were available at 2 or more time points were included in outcome assessments. Due to high levels of attrition and considerable variation in the time points of available follow-up data, comparisons were made between the first and last available data point for each participant, regardless of the spacing of these data points. Further, we focused on the first goal selected by the user to work toward only. Bivariate correlation coefficients were calculated to explore any associations between engagement indices and outcome change scores.

To assess any differences in access between demographic groups, the numbers of children invited to use the intervention and who activated their game key were disaggregated by gender, ethnicity, disability status, and neighbourhood deprivation. This information was compared with available data from the 2021 census in England and Wales to assess how closely the cohort of young people accessing Lumi Nova reflects the wider population of those aged 7 - 12 years.

To explore any difference in engagement or outcomes between groups, χ^2 tests were used to assess whether there are statistically significant between-group differences in activation or number of challenges completed, and mixed-effect ANOVAs were used to assess whether there is statistically significance between-group differences in mean change over time in GBO or CORS scores.

Patient and Public Involvement

Patient and public involvement has been central to BfB Labs' approach to designing, building, and testing digital interventions for young people. From the outset BfB Labs have worked alongside their target users to codesign interventions, using nonextractive iterative, user-led, agile design methodology. During this study, we worked alongside a panel of young advisors who advised on all aspects of this study, contributed to the interpretation of study findings, and supported dissemination efforts. Young advisors were paid for their time and offered opportunities to take part in training to build their research skills, as well as letters of recommendation or references to support their educational or professional development. The opportunity to be named as coauthors on this paper were discussed with the group, but by mutual agreement, young advisors wished to be acknowledged in this paper only.

Ethical Considerations

Ethical approval was obtained from the NHS Health Research Authority following a favourable ethical opinion from North West—Preston Research Ethics Committee (22/NW/0195). As the data were routinely collected and analysed retrospectively, study specific consent was not sought for reasons of practicality and limiting burden on users. However, parents or caregivers were informed that anonymized user data from the application would be used to evaluate the intervention and indicated their agreement to their child's data being used for this purpose during the sign-up process. Only approved staff members from BfB Labs had access to patient identifiable data, which was anonymized before analysis by the research team; the research team were not granted access to the anonymized data until confirmation of Health Research Authority approval for the evaluation was received.

Results

Participants

The initial dataset included data from 1202 families invited to use Lumi Nova during the period of interest. During the data cleaning process, we excluded 173 families from the evaluation who did not appear to meet the eligibility criteria (based on their responses to demographic questions or date of invitation or activation). Reasons for exclusion were: not having had the opportunity to use the intervention for at least one week (n=22), being outside of the target age range for the game (n=147), and not currently living in the United Kingdom (n=4). Demographic characteristics of the sample are summarized in Table 1.

Table . Sample characteristics.

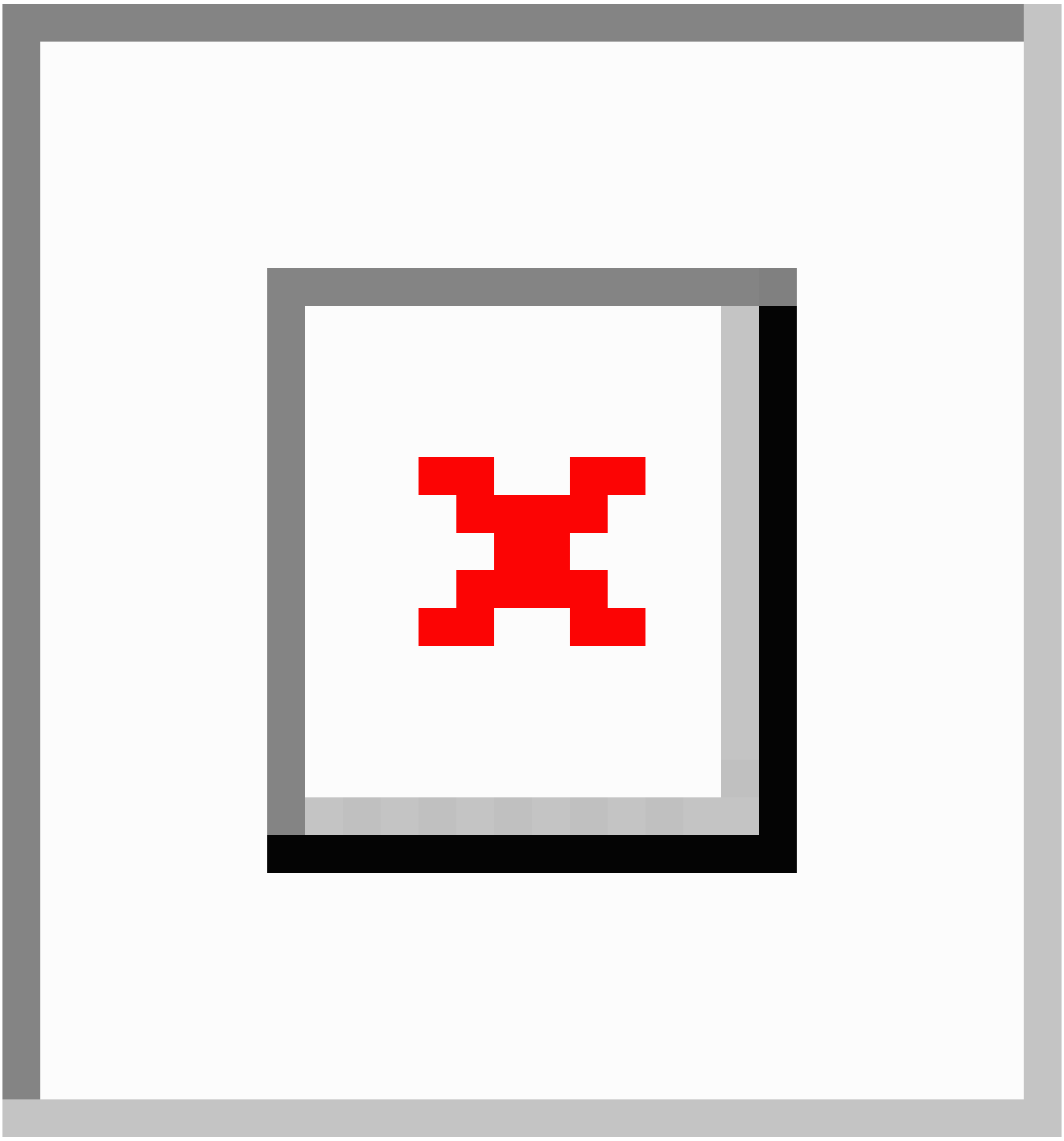
Characteristics		All cases (n=1029)	Activated cases (n=644)
Age (years), mean (SD, range)		9.71 (1.46, 7.00 - 12.98)	9.72 (1.44, 7.00 - 12.98)
Gender			
	Female, n (%)	492 (51.79)	312 (52)
	Male, n (%)	458 (48.21)	288 (48)
	Missing or not available, n	79	44
Ethnicity			
	Arab, n (%)	1 (0.1)	1 (0.2)
	Asian or Asian British, n (%)	12 (1.4)	9 (1.7)
	Black, African, Caribbean, or Black British, n (%)	17 (2)	7 (1.3)
	Chinese, n (%)	1 (0.1)	1 (0.2)
	Mixed or multiple ethnic group, n (%)	45 (5.3)	23 (4.4)
	Other ethnic group, n (%)	8 (0.9)	4 (0.8)
	White, n (%)	772 (90.19)	480 (91.42)
	Missing, n	173	119
Disability			
	True, n (%)	122 (11.86)	76 (11.8)
	False, n (%)	907 (88.14)	568 (88.2)
Referral type			
	Self–sign up, n (%)	618 (60.05)	381 (59.16)
	Clinician referral, n (%)	411 (39.95)	263 (40.84)
Operating system			
	Android, n (%)	—[a]	219 (34)
	iOS, n (%)	—	425 (66)

[a]N/A: not applicable.

Engagement

Of the 1029 families included in the final dataset, 644 (62.58%) parents or caregivers completed the baseline CORS to activate their Lumi Nova game key to enable their child to use the intervention. As the distributions of all engagement indices were positively skewed, medians and IQRs were calculated. The median number of unique sessions completed by those who activated their game key was 6 (IQR 3 - 12). The median total time spent engaging with the intervention was 42 (IQR 15 - 79) minutes and the median number of challenges completed was 1 (IQR 0 - 3); 58.1% (n=358) completed at least one challenge, 42% at least two (n=270), 28.2% at least three (n=181), and 17.7% (n=114) completed 4 or more. Those who completed at least one challenge also engaged in the psychoeducation element of the intervention as this is a prerequisite of commencing the first challenge. Histograms showing the distribution of unique sessions and challenges completed are in Figure 1.

Figure 1. Histograms showing distribution of engagement indices.



Outcomes

Paired GBO data was available for 224 cases and paired CORS data for 123 cases. There was a statistically significant increase in GBO scores between the first and last available data point (t_{223}=5.78, *P*<.001). The size of the effect was in the small to medium range (*d*=0.37, 95% CI 0.25 - 0.52), indicating that, on average, young people who engaged with the game made progress toward their selected goal. At the individual level, 39.3% (n=88) of cases saw a reliable improvement in GBO scores, 14.3% (n=32) a reliable deterioration, and 46.4% (n=104) had no reliable change.

There was also a statistically significant increase in CORS scores between the first and last available data point (t_{122}=5.10, *P*<.001). The size of the effect was in the small to medium range (*d*=0.46, 95% CI 0.27 - 0.65), indicating that parents or caregivers of children who engaged with the game reported improvements in functioning. At the individual level, 37.34% (n=46) of cases saw a reliable improvement in CORS score, 9.8% (n=12) a reliable deterioration, and 52.8% (n=65) had no reliable change.

Paired anxiety ratings pre and post at least one challenge were available for 363 cases. The average anxiety rating prechallenge was statistically significantly higher than the average postchallenge rating (t_{361}=8.35, *P*<.001). The size of the effect was again in the small to medium range (*d*=0.44, 95% CI 0.33 - 0.55). There was no association between the size of the average discrepancy between pre and post challenge anxiety ratings and change in GBO or CORS scores.

No associations were evident between the number of challenges completed or time spent engaging with the intervention and change score for either the GBO, CORS, or average anxiety ratings pre and post challenges.

Equality of Access, Engagement, and Outcomes

Of those young people whose parent or caregiver activated their game key, 51.79% (492/1029) identified as female and 48.21% (458/1029) as male. Females were significantly more likely to complete at least one challenge than males (n=200, 61.9% vs n=158, 53.5%, *P*=.02). There were no other significant differences between males and females in markers of engagement or outcomes.

More young people whose parent or caregiver activated their game key identified as White (480/644, 91.42%) than the percentage of those aged 7 - 12 years who identify as White in the general population of England and Wales (n =3,174,513, 73.5%). Conversely, children of Asian and Black ethnicities were underrepresented among Lumi Nova users compared to the population of those aged 7 - 12 years in England and Wales (n=12, Asian 1.7% vs n=486,124, 11.2% and Black n=17, 1.3% vs n=237,872, 5.5% respectively). It was not possible to conduct meaningful statistical analyses of any differences in engagement or outcomes between ethnic groups due to the small number of Lumi Nova users in the sample who identified as belonging to each group. Further, it should be noted that, due to commissioning arrangements, Lumi Nova was only available in certain areas which may not have reflected the demographics of the country as a whole, and often only to those already under the care of mental health services.

Compared to the general population of those aged 7 - 12 years, Lumi Nova users were more likely to live in the least deprived areas of the United Kingdom and less likely to live in the most deprived areas. However, young people living in the most deprived areas were significantly more likely than those living in more affluent areas to complete at least one challenge (n=87, 66.4% vs n=125, 53.8%; OR 1.68, 95% CI 1.08-2.64, *P*=.02). There were no significant differences in the outcomes of young people living in the most deprived areas compared to the most affluent areas.

Of those young people whose parent or caregiver activated their game key, 11.8% (n=121) reported a disability, a higher percentage with a disability than in the general population of those aged 7 - 12 years (n=190,287-445,469, 4.4% - 10.3%). However, young people with disabilities were less likely to successfully complete at least one challenge than nondisabled users (n=32, 42.7% vs n=326, 59.6%, OR 0.5, 95% CI 0.3-0.81, *P*=.005). The number of young people with a disability for whom paired outcome data was available was small (GBOs, n=20; CORS, n=11). However, within this small cohort of users, improvements in GBO and CORS scores were smaller than for nondisabled young people. The nature of reported disabilities (ie, physical, learning, or neurodevelopmental) was not recorded.

Discussion

Principal Findings

There is a pressing need to improve young people's access to evidence-based mental health treatments. To meet this need in the context of growing demand for mental health support and a limited pool of qualified practitioners, innovative approaches are required, including improved use of digital technologies. Lumi Nova, a digital therapeutic intervention which uses immersive gaming technology to deliver an evidence-informed anxiety intervention for those aged 7 - 12 years, is an example of an innovation with the potential to facilitate increased access to timely treatment. We conducted a real-world evaluation of Lumi Nova, to better understand user engagement and outcomes, and to explore whether these might differ across demographic groups.

We found that the majority (n=644, 62.58%) of families invited to access Lumi Nova activated their game key, of whom 58.1% (n=358) completed at least one challenge. However, only a relatively small proportion of users went on to complete multiple challenges, resulting in a median number of challenges completed of 1 (IQR 0 - 3). The median number of unique sessions was 6 (IQR 3 - 12), but these sessions were typically short relative to the usual duration of face-to-face therapy sessions, with the median total time spent engaging with the intervention being 42 (IQR 15 - 79) minutes.

For the subset of young people for whom paired outcome measures were available, there were statistically significant small to medium improvements in perceived progress toward goals and parent-reported functioning between the first and last data points. Two in 5 young people reported an improvement that would be considered reliable (ie, the magnitude of the change is such that it is unlikely to be accounted for by measurement error) for both the GBOs and CORS. Available paired anxiety ratings suggest that young people's anticipated anxiety prechallenge was, on average, higher than the anxiety they reported post challenge, providing some support that the mechanisms of graded exposure were operating as would be anticipated. However, there was no evidence of an association between either the extent of mismatch in anxiety expectation or the child's level of engagement and the outcomes observed.

Interpretation of these findings is complicated by the small proportion of the overall sample for whom paired data was available, the uncontrolled design and limited available comparison data from studies of interventions delivered in real-world treatment settings. However, analysis of routinely collected data from the Child Outcomes Research Consortium [30] suggests that the outcomes observed for young people accessing Lumi Nova are broadly in line with those observed in young people receiving routine treatment for child and adolescent mental health services.

Given the relatively brief duration of the intervention received, that we observed outcomes of a comparable magnitude to those seen in routine practice could be interpreted as preliminary evidence that Lumi Nova may offer a particularly efficient treatment option. However, we are unable to rule out the possibility that the improvements observed may simply reflect natural recovery or the effect of other support received alongside Lumi Nova. Further, as paired outcome data was only available for a subset of those who activated the intervention, we cannot be sure that these findings would generalize to all Lumi Nova users, let alone the wider population of children seeking treatment for anxiety.

Comparing the demographic characteristics of Lumi Nova users to the general population of those aged 7 - 12 years in England and Wales suggested that children from ethnic minority backgrounds and those living in the most deprived neighbourhoods may be less likely to access Lumi Nova than other young people. However, once users from more deprived areas did access Lumi Nova, they were more likely than affluent users to complete at least one challenge. Given that ethnic minority and economically disadvantaged young people are more likely to experience mental health difficulties than their peers [12,13], this is an important indicative finding that requires further investigation.

However, it should be borne in mind that, due to commissioning arrangements, Lumi Nova was only available in certain mental health services around the United Kingdom during the implementation period studied. These services were located in diverse areas of the United Kingdom including the East of England, London, Wales, Essex, Yorkshire, Bristol, Gloucestershire, and Kent, but only within certain organisations and specific NHS mental health teams. As such, the demographic comparative work relating to deprivation and ethnicity undertaken in this paper has used UK-wide data to reflect this diverse geography, in lieu of the absence of population data in these specific areas for those aged 7 - 12 years. Furthermore, since the option for families to access the intervention via self-sign-up was implemented toward the end of the period studied, and only in certain regions, in most cases, only those already in contact with mental health services would have been offered Lumi Nova, yet no quality mental health service access data on a specific age basis is publicly available which met the needs of this study analysis. It is possible that further mitigations are needed to overcome barriers to help-seeking among marginalized groups. For instance, expansion and further promotion of the self-sign-up option may help to overcome structural and attitudinal barriers to accessing traditional mental health services [11].

When examining differences in outcomes based on user characteristics, we found some evidence that females and those from less affluent areas may be more likely than other users to complete at least one challenge, but, overall, we identified few differences between the engagement and outcomes of Lumi Nova users across demographic groups. However, the small numbers of young people in the sample who identified as belonging to some minority groups and the high volume of missing data limited the statistical comparisons that could meaningfully be made. There were preliminary indications that young people with a disability may be less likely to successfully engage with the intervention than children without a disability. The small size of those subgroups involved, alongside the limited available information about the nature of disabilities reported, precludes any firm conclusions from being drawn from this finding. However, further research investigating the accessibility of digital interventions such as Lumi Nova to young people with disabilities is warranted.

Strengths and Limitations

One of the key strengths of this study was that it involved a real-world evaluation, reflecting pragmatic usage of the intervention as part of routine practice in NHS settings and families outside of experimental research contexts, conferring external validity to the findings [31]. This approach also allowed for analysis on a larger dataset of 1029 people, even given the extent of attrition, than would have been possible to collect in a controlled study carried out with similar resources and over a similar timeframe. Moreover, the scope of data collected and analysed in this study was more extensive than is typically made available for commercial digital interventions and services, incorporating demographic, engagement, or access and outcome data, not just data regarding initial access. Collectively, this allowed for an in depth exploration of not only whether Lumi Nova appears to be effective to treat young people's anxiety, but an insight into the platform's users, acceptance and real-life usage in the home, and early indicators about who it is more or less likely to benefit. These insights are important for practitioners and commissioners to consider if wishing to implement Lumi Nova into future services.

However, there were also several limitations to the approach taken in our study. Relying on retrospective routinely captured data completed by young people themselves without support by professionals meant the dataset had more missing data than would be anticipated in a prospective study design where missing data could be identified and addressed during the research process, posing some challenges during the data cleaning process. We were also limited by data relating to the population census statistics as a comparator, as no meaningful mental health service data is available for this specific age range. The UK-wide implementation of Lumi Nova, albeit across selected services only, meant that we were limited to comparisons of the user cohort with UK-level population data rather than using mental health service data or local small area census data. We were also limited by the choice of measures and other information available within the dataset. For instance, no standardized self-report measure of anxiety symptom severity was available and data on disabilities was captured as a binary variable, with no information available on the nature of the disability, which may have been relevant to understanding accessibility. Importantly, the real-life approach also meant that this study had no control comparator group and there was limited information on what (if any) treatment young people received in addition to Lumi Nova from the NHS or other care providers, compounding the limitations imposed by the uncontrolled study design when evaluating the cohort's clinical outcomes. This means that conclusions of effectiveness need to be treated with caution as improvements relating to natural recovery or the impact of other treatment interventions delivered in the same period as Lumi Nova usage cannot be ruled out.

Further, interpretation of the findings was made more challenging by the high level of missing data. This challenge was heightened by the impossibility of distinguishing between missing data due to a child disengaging from the intervention and missing data due to a child playing the game offline following the initial download and their data not being transferred to the central data hub as a result.

Implications and Recommendations

Providers of digital therapeutics should be encouraged to collect, monitor, and make available for analysis, data on how interventions are being used, by whom, and the outcomes observed. It should be noted that the quality and completeness of routine data collection by traditional child and adolescent mental health services is often poor [25]. As such, we must be cautious of holding digital interventions to standards of monitoring and evaluation that far exceed those of nondigital interventions. However, we should work toward thorough and systematic collection and analysis of routine data across all forms of child and adolescent mental health treatment.

Learning from this evaluation has already been used by BfB Labs to refine their processes to improve the quality and completeness of the data they are able to capture. For instance, extra data save points have been implemented to minimize the loss of data in the event the child does not continue to the next save point or connect to the internet for their data to be uploaded. This improved data will be used by BfB Labs to continue to evaluate and refine their product.

Further research is needed to explore how any barriers to access or engagement can be reduced among those with disabilities, young people from marginalized ethnic backgrounds, and those living in the most deprived areas to ensure all young people can benefit equally from interventions such as Lumi Nova. We are aware of an ongoing study (ISRCTN Registry, ISRCTN11131689) specifically aiming to understand the barriers to uptake and usage of Lumi Nova among children growing up in circumstances of economic disadvantage. This study aims to understand how the intervention can be implemented to maximize usage and engagement among those living in postcodes in the 2 most deprived deciles of England. Future research should also evaluate the benefits and costs of offering Lumi Nova and other digital therapeutic interventions via self-sign-up routes versus requiring families to be on-boarded by mental health clinicians.

On a broader level, this study also raises methodological questions regarding the best way to provide evidence for digital therapeutic interventions. Digital interventions must continually and rapidly evolve if they are to keep pace with technological advancements, compatibility issues, and user expectations. As such, there is a need to ensure that approaches to evaluation balance scientific rigour with the ability to provide evidence that is timely and has sufficient ecological validity to retain relevance to real-world settings. Researchers, commissioners, providers, guidance authors, and innovators must work together to agree on what constitutes a sufficiently robust approach to evaluation in this rapidly developing field.

Conclusions

Providing the growing number of children seeking mental health support with timely access to care poses a significant challenge. This study provides indicative evidence that use of Lumi Nova may be associated with improvements in functioning and goal-based outcomes among those aged 7 - 12 years seeking treatment for anxiety difficulties. However, the strength of this study’s findings is limited by a lack of a control comparator group and missing information about any concurrent treatments being received by young people using the intervention. As such, Lumi Nova has the potential to play an important role in improving access to mental health treatment for children experiencing anxiety. However, further research is needed to confirm these findings as natural recovery or the impact of other interventions cannot be ruled out. Future research should consider how engagement with the intervention can be maximized and sustained where beneficial for that child’s needs, particularly in children from less advantaged groups.

Acknowledgments

This study was funded by the National Institute for Health Research Invention for Innovation programme (NIHR202504). The views expressed are those of the authors and not necessarily those of the National Institute for Health Research or the Department of Health and Social Care. The authors wish to acknowledge the assistance of Manjul Rathee, CEO, Nabeena Mali, BFA, and Kelly Tham, BSc, from BfB Labs in providing and interpreting the dataset for analysis. We are grateful to Dr Sarah Robinson, ClinPsyD, and Sophie Knight, MPhil (Eastern Academic Health Science Network), for their contribution to the conception of this study and advisory input to the wider project. Thanks also to Aoife Dunne, MSc (research assistant psychologist), for her input to the wider project and contribution to co-ordination of patient and public involvement activities. We are also extremely grateful to the young people who gave up their time to contribute to this study as young advisors.

Conflicts of Interest

None declared.

References

1. Deng J, Zhou F, Hou W, et al. Prevalence of mental health symptoms in children and adolescents during the COVID-19 pandemic: a meta-analysis. Ann N Y Acad Sci 2023 Feb;1520(1):53-73. [doi: 10.1111/nyas.14947] [Medline: 36537131]
2. Mental health of children and young people in England 2022 - wave 3 follow up to the 2017 survey. NHS Digital. 2022. URL: https://digital.nhs.uk/data-and-information/publications/statistical/mental-health-of-children-and-young-people-in-england/2022-follow-up-to-the-2017-survey [accessed 2024-07-31]

3. Neufeld SAS. The burden of young people's mental health conditions in Europe: no cause for complacency. Lancet Reg Health Eur 2022 May;16:100364. [doi: 10.1016/j.lanepe.2022.100364] [Medline: 35345644]
4. British Medical Association. Measuring progress: commitments to support and expand the mental health workforce in england. 2020. URL: https://www.bma.org.uk/media/2405/bma-measuring-progress-of-commitments-for-mental-health-workforce-jan-2020.pdf [accessed 2024-09-28]
5. Schueller SM, Torous J. Scaling evidence-based treatments through digital mental health. Am Psychol 2020 Nov;75(8):1093-1104. [doi: 10.1037/amp0000654] [Medline: 33252947]
6. Eilert N, Enrique A, Wogan R, Mooney O, Timulak L, Richards D. The effectiveness of internet-delivered treatment for generalized anxiety disorder: an updated systematic review and meta-analysis. Depress Anxiety 2021 Feb;38(2):196-219. [doi: 10.1002/da.23115] [Medline: 33225589]
7. Hollis C, Falconer CJ, Martin JL, et al. Annual research review: digital health interventions for children and young people with mental health problems - a systematic and meta-review. J Child Psychol Psychiatry 2017 Apr;58(4):474-503. [doi: 10.1111/jcpp.12663] [Medline: 27943285]
8. Lehtimaki S, Martic J, Wahl B, Foster KT, Schwalbe N. Evidence on digital mental health interventions for adolescents and young people: systematic overview. JMIR Ment Health 2021 Apr 29;8(4):e25847. [doi: 10.2196/25847] [Medline: 33913817]
9. Ebert DD, Zarski AC, Christensen H, et al. Internet and computer-based cognitive behavioral therapy for anxiety and depression in youth: a meta-analysis of randomized controlled outcome trials. PLoS One 2015 Mar;10(3):e0119895. [doi: 10.1371/journal.pone.0119895] [Medline: 25786025]
10. Pennant ME, Loucas CE, Whittington C, et al. Computerised therapies for anxiety and depression in children and young people: a systematic review and meta-analysis. Behav Res Ther 2015 Apr;67:1-18. [doi: 10.1016/j.brat.2015.01.009] [Medline: 25727678]
11. Silk JS, Pramana G, Sequeira SL, et al. Using a smartphone app and clinician portal to enhance brief cognitive behavioral therapy for childhood anxiety disorders. Behav Ther 2020 Jan;51(1):69-84. [doi: 10.1016/j.beth.2019.05.002] [Medline: 32005341]
12. Abraham A, Walker-Harding L. The key social determinants of mental health: their effects among children globally and strategies to address them: a narrative review. Pediatr Med 2022 Feb;5(1):1-10. [doi: 10.21037/pm-20-78]
13. Reiss F, Meyrose AK, Otto C, Lampert T, Klasen F, Ravens-Sieberer U. Socioeconomic status, stressful life situations and mental health problems in children and adolescents: results of the German BELLA cohort-study. PLoS One 2019 Mar;14(3):e0213700. [doi: 10.1371/journal.pone.0213700] [Medline: 30865713]
14. Robards F, Kang M, Usherwood T, Sanci L. How marginalized young people access, engage with, and navigate health-care systems in the digital age: systematic review. J Adolesc Health 2018 Apr;62(4):365-381. [doi: 10.1016/j.jadohealth.2017.10.018]
15. Piers R, Williams JM, Sharpe H. Review: can digital mental health interventions bridge the 'digital divide' for socioeconomically and digitally marginalised youth? A systematic review. Child Adolesc Ment Health 2023 Feb;28(1):90-104. [doi: 10.1111/camh.12620] [Medline: 36510375]
16. Middle R, Welch L. Experiences of digital exclusion and the impact on health in people living with severe mental illness. Front Digit Health 2022 Nov;4:1-12. [doi: 10.3389/fdgth.2022.1004547]
17. Barican JL, Yung D, Schwartz C, Zheng Y, Georgiades K, Waddell C. Prevalence of childhood mental disorders in high-income countries: a systematic review and meta-analysis to inform policymaking. Evid Based Mental Health 2022 Feb;25(1):36-44. [doi: 10.1136/ebmental-2021-300277]
18. Creswell C, Shum A, Pearcey S, Skripkauskaite S, Patalay P, Waite P. Young people's mental health during the COVID-19 pandemic. Lancet Child Adolesc Health 2021 Aug;5(8):535-537. [doi: 10.1016/S2352-4642(21)00177-2]
19. Hill C, Waite P, Creswell C. Anxiety disorders in children and adolescents. Paediatr Child Health (Oxford) 2016 Dec;26(12):548-553. [doi: 10.20419/2020.29.505]
20. Reardon T, Harvey K, Creswell C. Seeking and accessing professional support for child anxiety in a community sample. Eur Child Adolesc Psychiatry 2020 May;29(5):649-664. [doi: 10.1007/s00787-019-01388-4] [Medline: 31410579]
21. Bry LJ, Chou T, Miguel E, Comer JS. Consumer smartphone apps marketed for child and adolescent anxiety: a systematic review and content analysis. Behav Ther 2018 Mar;49(2):249-261. [doi: 10.1016/j.beth.2017.07.008] [Medline: 29530263]
22. BfB Labs. Lumi Nova: tales of courage. URL: https://www.bfb-labs.com/luminova [accessed 2023-07-28]
23. Furnham A, Wilson E, Chapman A, Persaud R. Treatment hurts: lay theories of graded exposure in the treatment of four anxiety disorders. Eur J Psychother Couns 2013 Sep;15(3):253-273. [doi: 10.1080/13642537.2013.810657] [Medline: 26405431]
24. Lockwood J, Williams L, Martin JL, Rathee M, Hill C. User engagement and experience, and safety of a mobile app (Lumi Nova) delivering exposure-based cognitive behavioral therapy strategies to manage anxiety in children via Immersive gaming technology: preliminary evaluation study. JMIR Ment Health 2022 Jan;9(1):e29008. [doi: 10.2196/29008]
25. Guided self-help digital cognitive behavioural therapy for children and young people with mild to moderate symptoms of anxiety or low mood: early value assessment. National Institute for Health and Care Excellence. 2023. URL: https://www.nice.org.uk/guidance/hte3/resources/

guided-self-help-digital-cognitive-behavioural-therapy-for-children-and-young-people-with-mild-to-moderate-symptoms-of-anxiety-or-low-mood-early-value-assessment-pdf-1396168997317 [accessed 2024-07-31]
26. List of goals and challenges. Lumi Nova: Tales of Courage. 2021. URL: https://luminova.app/goals [accessed 2024-05-03]
27. English indices of deprivation 2019. UK.gov/Ministry of Housing, Communities & Local Government. 2019. URL: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019 [accessed 2023-08-17]
28. Law D, Jacob J. Goals and Goal Based Outcomes (GBOs), 3rd edition: CAMHS Press; 2015. URL: https://www.corc.uk.net/media/1219/goalsandgbos-thirdedition.pdf [accessed 2024-10-14]
29. Campbell A, Hemsley S. Outcome rating scale and session rating scale in psychological practice: clinical utility of ultra-brief measures. Clin Psychol 2009 Mar 1;13(1):1-9. [doi: 10.1080/13284200802676391]
30. Wolpert M, Jacob J, Napoleone E, Whale A, Calderon A, Edbrooke-Childs J. Child-and parent-reported outcomes and experience from child and young people's mental health services 2011–2015. Child Outcomes Research Consortium. 2016. URL: https://www.corc.uk.net/media/1544/0505207_corc-report_for-web.pdf [accessed 2024-09-28]
31. Real-world evaluation to facilitate adoption at scale. The AHSN Network. 2022. URL: https://www.ahsnnetwork.com/wp-content/uploads/2022/12/R58-Real-world-evaluation-to-facilitate-adoption-at-scale.pdf [accessed 2024-07-31]

Abbreviations

CBT: cognitive behavioral therapy
CE: Conformité Européene
CORS: Child Outcome Rating Scale
GBO: goal-based outcome
NHS: National Health Service

Edited by J Torous; submitted 18.09.23; peer-reviewed by M Janbakhsh, S Rennick-Egglestone; revised version received 28.05.24; accepted 06.06.24; published 22.10.24.

Please cite as:
Gee B, Teague B, Laphan A, Clarke T, Coote G, Garner J, Wilson J
Outcomes of Providing Children Aged 7-12 Years With Access to Evidence-Based Anxiety Treatment Via a Standalone Digital Intervention Using Immersive Gaming Technology: Real-World Evaluation
JMIR Ment Health 2024;11:e52866
URL: https://mental.jmir.org/2024/1/e52866
doi:10.2196/52866

Original Paper

Predictors of Use and Drop Out From a Web-Based Cognitive Behavioral Therapy Program and Health Community for Depression and Anxiety in Primary Care Patients: Secondary Analysis of a Randomized Controlled Trial

Armando J Rotondi[1,2,3], PhD; Bea Herbeck Belnap[2,4], Dr Biol Hum, DBH; Scott Rothenberger[4], DPhil; Robert Feldman[4], MA; Barbara Hanusa[1], DPhil; Bruce L Rollman[2,4], MPH, MD

[1]Mental Illness Research Education and Clinical Center, VA Pittsburgh Healthcare System, Veterans Administration, Pittsburgh, PA, United States
[2]Center for Behavioral Health, Media and Technology, University of Pittsburgh, Pittsburgh, PA, United States
[3]Center for Health Equity Research and Promotion, VA Pittsburgh Healthcare System, Veterans Administration, Pittsburgh, PA, United States
[4]Center for Research on Health Care, University of Pittsburgh School of Medicine, Pittsburgh, PA, United States

Corresponding Author:
Armando J Rotondi, PhD
Mental Illness Research Education and Clinical Center
VA Pittsburgh Healthcare System
Veterans Administration
Research Office Building (151R-U)
University Drive C
Pittsburgh, PA, 15240
United States
Phone: 1 412 360 2494
Fax: 1 412 360 2369
Email: armandorotondi1@gmail.com

Abstract

Background: A previously reported study examined the treatment of primary care patients with at least moderate severity depressive or anxiety symptoms via an evidence-based computerized cognitive behavioral therapy (CCBT) program (Beating the Blues) and an online health community (OHC) that included a moderated internet support group. The 2 treatment arms proved to be equally successful at 6-month follow-up.

Objective: Although highly promising, e-mental health treatment programs have encountered high rates of noninitiation, poor adherence, and discontinuation. Identifying ways to counter these tendencies is critical for their success. To further explore these issues, this study identified the primary care patient characteristics that increased the chances patients would not initiate the use of an intervention, (ie, not try it even once), initiate use, and go on to discontinue or continue to use an intervention.

Methods: The study had 3 arms: one received access to CCBT (n=301); another received CCBT plus OHC (n=302), which included a moderated internet support group; and the third received usual care (n=101). Participants in the 2 active intervention arms of the study were grouped together for analyses of CCBT use (n=603) because both arms had access to CCBT, and there were no differences in outcomes between the 2 arms. Analyses of OHC use were based on 302 participants who were randomized to that arm.

Results: Several baseline patient characteristics were associated with failure to initiate the use of CCBT, including having worse physical health (measured by the Short Form Health Survey Physical Components Score, P=.01), more interference from pain (by the Patient-Reported Outcomes Measurement Information System Pain Interference score, P=.048), less formal education (P=.02), and being African American or another US minority group (P=.006). Characteristics associated with failure to initiate use of the OHC were better mental health (by the Short Form Health Survey Mental Components Score, P=.04), lower use of the internet (P=.005), and less formal education (P=.001). Those who initiated the use of the CCBT program but went on to complete less of the program had less formal education (P=.01) and lower severity of anxiety symptoms (P=.03).

Conclusions: This study found that several patient characteristics predicted whether a patient was likely to not initiate use or discontinue the use of CCBT or OHC. These findings have clear implications for actionable areas that can be targeted during initial and ongoing engagement activities designed to increase patient buy-in, as well as increase subsequent use and the resulting success of eHealth programs.

Trial Registration: ClinicalTrials.gov NCT01482806; https://clinicaltrials.gov/study/NCT01482806

(JMIR Ment Health 2024;11:e52197) doi:10.2196/52197

KEYWORDS

e-mental health; user engagement; initiation; discontinue; depression; anxiety; cognitive behavioral therapy; computerized CBT; online health community; collaborative care; internet support group

Introduction

Background

Currently, there is a significant gap in mental health treatment. For example, almost 75% of US patients who screen positive for depressive symptoms do not participate in treatment [1]. One reason for this gap is the reliance on in-person delivery models [2]. In-person services have inherent barriers, including challenges scheduling sessions, transportation limitations, symptom exacerbations limiting the ability to travel, illness-reduced motivation, stigma, and difficulties affording out-of-pocket costs [2]. Availability of and timely access to in-person services are also limited, with >90% of psychologists and psychiatrists and 80% of professionals with a master's degree in social work practicing exclusively in metropolitan areas [3]. In addition, treatments may have limited effectiveness. Randomized clinical trials (RCTs) of pharmacotherapy and in-person psychotherapy for mental health disorders yield effect sizes that are typically small, with 0.30 standardized mean difference [4]. This contributes to the difficulty of finding effective treatments. A study on patient-reported helpfulness of mental health treatments found that only 26.1% were helped by the first treatment they tried [5]. Persisting to a second treatment resulted in a cumulative probability of feeling helped to 51.2%. After experiencing unhelpful treatment, patients would need to persist through up to 8 providers to reach a cumulative probability of 91%, obtaining a treatment that they found helpful [5].

These characteristics of in-person treatment contribute to high rates of noninitiation, poor adherence, and discontinuation. Even in RCTs, noninitiation and premature discontinuation average approximately 30% [6]. In practice settings, almost one-third of patients referred to psychotherapies do not initiate treatment, and those who do typically discontinue during the initial sessions (up to 44% after the first session and 82% by the fifth session) [7].

eHealth approaches offer promise in addressing many of these limitations by increasing convenience, reducing the level of effort and associated motivation involved in traveling to in-person treatments, overcoming limits on the availability of practitioners and treatments, reducing the obstacles involved in initiating and switching treatments, lowering relative costs, and reducing stigma. Despite their great potential and many successes [8-10], there are commonly encountered barriers, including the levels of user digital and health literacy required, lack of confidence [11], poor designs, lack of attention to cognitive design needs [12], limited technology access, and a lack of evidence-based methods to *engage* patients with eHealth programs [13-15]. Engagement has several different meanings in the literature [16]. The view taken in this paper is that a user is engaged when he or she (1) believes that using the intervention will result in positive changes that she or he values; (2) has the motivation, confidence, and ability to initiate use; and (3) can efficiently use the technologies and eHealth program. The 3 implications of this conceptualization are that engagement is a process, usability (actual and perceived) of the technologies and their information architecture influence engagement, and intervention methods should be designed to develop *sufficient* engagement of users initially and going forward. Evidence shows that many, eHealth programs too often fail to engage users sufficiently.

Consequently, noninitiation, discontinuation, and lack of adherence can be quite high [15,17]. In recognition of the significance of engagement in the success of eHealth, and how understudied this has been, Eysenbach [17] developed the "Law of Attrition." This law highlights inadequate buy-in by users as a fundamental methodological challenge that must be addressed if eHealth programs are to be successful in practice. In one of his examples, 99.5% of the participants discontinued MoodGym, an online, evidence-based program for depression. A recent meta-analysis of RCTs on the treatment of depressive symptoms found an estimated dropout rate of 47.8% (after adjusting for publication bias) [18]. A meta-analysis of the real-world uptake of interventions for depression, mood enhancement, and anxiety found that 12% to 79% of those who downloaded an intervention did not use the intervention once; of the remaining who did initiate use, 58% to 93% had poor adherence; and 71.4% to 99.5% discontinued before completing 40% of the intervention [19]. Eysenbach [17] argued that there is a need to develop scientific theories that provide a better understanding of the causes of such phenomena and can form the basis for creating best practices to counter them.

To gain insights into possible influences on engagement, this study used data from the Online Treatment for Mood and Anxiety Disorders in Primary Care (Online Treatment) Trial [20] to describe the characteristics of participants who did not initiate versus initiated the use of each of the 2 e-mental health interventions and discontinued versus continued using one or the other intervention. Participants were given access to a self-help computerized cognitive behavioral therapy (CCBT) program (Beating the Blues). It was provided in a collaborative care framework that included support from a "care manager." A subgroup of participants also had access to a password-protected online health community (OHC) platform, which included moderated asynchronous internet support groups (ISGs).

Beating the Blues (CCBT), a web-based program, has been shown to be effective in treating depressive and anxiety symptoms [21-23]. The program is self-administered and consists of a brief 10-minute introductory video followed by eight 50-minute-long interactive modules. The modules' topics are (1) problem definition and pleasurable events; (2) automatic thoughts; (3) thinking errors and distraction; (4 and 5) challenging unhelpful thinking; (6) core beliefs; (7) attributional style; and (8) review, action planning, and conclusion. Each module uses text, audio, audiovisual clips, and "homework" assignments designed to impart basic cognitive behavioral therapy techniques that target depression and anxiety. These 8 modules must be completed sequentially.

ISGs can improve users' illness knowledge, coping skills, emotional support, connectedness, self-efficacy, and mental health [24-27]. Studies on users of web-based peer support have found that more frequent use results in more mental health benefits [28]. To explore the potential of peer support to enhance the effectiveness of CCBT, which is fully self-guided and does not offer the option for interaction with peers or providers, we included OHC. This provided access to several ISGs that the participants could use. However, nontherapeutic interactions can occur in these groups, which may result in negative consequences for users [29]. Therefore, it is important to use methods to keep the interactions supportive. The ISGs in this study had a moderator to manage content, facilitate supportive interactions, and amplify potential benefits [24].

Objective

The objective of this study was to identify patient baseline characteristics that were associated with noninitiation, poor adherence, discontinuing, or continuing to use each of 2 eHealth treatments for anxiety and depression, CCBT and OHC. This study conducted secondary analyses of data collected from the parent randomized controlled trial, Online Treatment for Mood and Anxiety Disorders in Primary Care [20]. The parent study provided CCBT and OHC to primary care patients with elevated depressive or anxiety symptoms.

Methods

Ethics Approval

The institutional review board of the University of Pittsburgh approved all study procedures (reference number: 20030187), and all participants provided written informed consent.

Overview

This study is a secondary analysis of data from the parent randomized controlled trial, the Online Treatment for Mood and Anxiety Disorders in Primary Care (Online Treatment) trial that examined the impact of access to 2 e-mental health interventions or their primary care physician's (PCP's) usual care (UC) on clinical outcomes [20]. The institutional review board of the University of Pittsburgh approved the study protocol, and all participants provided written informed consent.

Briefly, the parent trial recruited patients from 26 primary care offices affiliated with the University of Pittsburgh Medical Center, which shared a common electronic medical record system. A PCP received an electronic medical record system reminder about the study at the time of the clinical encounter for all patients aged 18 to 75 years with a diagnosis of depression, anxiety, generalized anxiety, or panic disorder. If the patient agreed to a study referral, they were contacted by a study recruiter via telephone to confirm protocol eligibility.

Participants

Eligible patients needed to have at least a moderate level of depressive or anxiety symptoms with a score 10 or greater on the 9-item Patient Health Questionnaire (PHQ-9) [30] or the 7-item Generalized Anxiety Disorder Scale [31]; reliable access to the internet, email, and telephone; and no alcohol dependence (as determined by the Alcohol Use Disorders Identification Test) [32], active suicidality, or other serious mental illness. Research assessors then administered via telephone the baseline battery and collected information on patients' self-reported race, sex, and other sociodemographic characteristics. This study was registered to ClinicalTrials.gov under trial registration number NCT01482806.

Randomization Procedure

Following the baseline assessment, patients were randomized in a 3:3:1 ratio to (1) access to a self-guided CCBT program (CCBT-only [Beating the Blues]; n=301); (2) CCBT plus access to a password-protected OHC platform that included moderated ISGs (CCBT+OHC; n=302); or (3) UC of their PCP (n=101). The present report only used data from participants randomized to the CCBT-only and CCBT+OHC intervention groups (N=603).

Interventions

Computerized Cognitive Behavioral Therapy

The Beating the Blues web-based program (described earlier) served as the CCBT intervention [33]. CCBT was provided in a collaborative care framework via support from a care manager. The care managers encouraged participants to complete a module every 1 to 2 weeks, but the participants were free to proceed at their own pace.

Online Health Community

The OHC was password-protected and featured collections of links to external resources (eg, crisis hotlines, "find-a-therapist" sites, local US $4 generic pharmacy programs) and brief YouTube videos on stress management, sleep hygiene, meditation, exercise benefits, and nutrition. In addition, the participants could interact with one another on a variety of moderated ISGs. Each ISG was dedicated to a specific topic [34]. Unlike the CCBT, the OHC was designed so that participants had freedom of choice regarding the resources they used.

A study investigator logged into the ISGs daily to review new posts and monitor for the presence of suicidal thoughts or potentially inappropriate content. In addition, the OHC had a moderator who oversaw all communications by participants. The moderator was a care manager (described below) for this study. To promote participants' ongoing involvement with the OHC a variety of strategies were used: (1) weekly emails from the moderator that highlighted new discussions, new content, or self-management tips; (2) status indicators on participants' profiles and comments they posted to a discussion (eg, stars and "likes"); (3) automatically generated email notifications of new ISG activities, for example, when someone replied to a participant's ISG post or comment; (4) automated highlighting of recent comments on participants' home pages, which were personalized to their ISG profile and past comments; (5) invited participants as guest moderators; and (6) initiating contests (eg, scavenger hunts, respond to emails or posts). During regular care management contacts, the care manager directed the participants to pertinent content on the OHC. For safety, at least one study clinician logged into the ISGs daily to monitor user-generated posts and comments, and the participants were able to "flag" potentially inappropriate content for removal.

Collaborative Care Framework: Care Managers Promote Engagement, Ongoing Involvement, and Effectiveness

Both eHealth programs were delivered via a collaborative care strategy. The 2 care managers had a bachelor's or master's degree in the psychology field and had worked on conducting mental health research with human participants. After randomization, a care manager dedicated to each intervention arm contacted the patient for an introductory telephone call and provided guidance in the setup of access to the CCBT program and the OHC, if applicable. During the 6-month intervention, participants were encouraged to complete a CCBT module every 1 to 2 weeks and were provided reminders, if necessary, encouragement on their progress, and personalized feedback. Participants had the option to contact a care manager with questions or for assistance. Care managers, irrespective of eHealth program usage, monitored patients' progress, symptoms, use of the eHealth programs, and telephoned those who either were not doing well, as indicated by their scores on the PHQ-9, or had failed to log onto the CCBT or OHC regularly to inquire why. They entered all contacts and information gathered into an electronic registry that was used to track each patient's progress and guide care managers through their contacts. If a patient's symptoms did not improve or worsen, they contacted her or him via email or telephone (depending on the patient's preference) to discuss additional treatment options, depending on the patient's preference. At weekly case review meetings, with the assistance of the electronic registry, the care managers presented their patients to the study clinical team, which consisted of a PCP, psychiatrist, and psychologist-study coordinator (study team). In addition to providing patients with suggestions for general lifestyle adjustments, including social engagement, exercise, adequate sleep, and nutrition suggestions, the study team also recommended initiation or modification of antidepressant or anxiolytic pharmacotherapy based on patients' symptom response and treatment preferences and referral to a mental health specialist when a patient had either complex psychosocial issues or was not responding to treatment. Following the case review, the care managers discussed recommendations with the patient and then notified the patient's PCPs of progress and treatment suggestions. PCPs remained responsible for the treatment and were free to continue or modify the patients' treatments.

Assessments

Blinded telephone assessors administered the assessment battery at baseline, 3 months, 6 months (end of interventions), and 12 months to assess the durability of the interventions. It included the administration of the 12-Item Short Form Health Survey (SF-12) assessment of SF-12 Mental Component Summary (MCS; SF-12 MCS) and Physical Component Summary (PCS; SF-12 PCS) scores, composed of 6 items each from the SF-12 to measure these 2 aspects of health-related quality of life [35]; the PHQ-9, a fixed-length Patient-Reported Outcomes Measurement Information System (PROMIS) measure to assess the severity of depressive symptoms [30]; the Generalized Anxiety Disorder 7- item scale to assess anxiety severity [31]; the 8-item fixed-length PROMIS-PI to assess pain interference, that is, the degree to which pain interferes with an individual's physical, mental, and social daily activities [36]; and the Primary Care Evaluation of Mental Disorders to provide diagnoses of depressive and anxiety disorders [37]. In addition, an 11-item shortened version of the Pew Internet Use questionnaire was administered at baseline to assess participants' use of the internet; server logs were abstracted to measure the use of the CCBT and OHC programs; and care managers' electronic registry was used to assess the number of intervention emails and telephone contacts.

Statistical Analyses

As the purpose of this study was to identify subgroups of participants based on their CCBT or OHC use patterns, we limited the analyses to the CCBT alone and CCBT+OHC treatment arms (N=603). As documented in a previous study, both were more effective than PCPs' UC for depression and anxiety but were similarly effective to each other (ie, offering access to the OHC produced no added reduction in depressive or anxiety symptoms over CCBT alone) [20]. Consequently, for CCBT analyses, we combined CCBT users from both arms (N=603).

We first classified the participants according to whether they initiated the use of the CCBT program. Participants who logged in at least once were categorized as having initiated CCBT use, whereas participants who never logged in comprised the noninitiation group. We compared baseline sociodemographic, clinical, and functional status measures by initiation status using *t* tests for continuous data and chi-square tests for categorical data. Fisher exact test was used for categorical measures when the expected cell counts were less than 5. Participants were then stratified by whether they "continued" the use of CCBT (yes or no), as those who completed at least one CCBT module. Those who did not complete the first CCBT module were categorized as "discontinued" use. Baseline characteristics were compared between these 2 groups (ie, continued and discontinued use) using *t* tests and chi-squared tests (or Fisher exact test, as appropriate). On the basis of previous work with this data set that found differences in the effects of CCBT between African American participants and White participants [38], we also examined whether there might be differences in the use patterns.

Participants in the CCBT+OHC study arm were then classified as initiation or noninitiation of OHC. Those who logged in at least once were considered to have initiated OHC use, whereas those who never logged in were categorized as noninitiators of OHC use. The baseline measurements were compared between the 2 groups. Continuation of OHC use was evaluated by stratifying OHC participants into 3 mutually exclusive categories: those who logged in only once, those who logged in 2 to 3 times, and those who logged in more than 3 times during the 6-month intervention phase. The baseline characteristics were compared across these groups. Finally, continuation of OHC use was classified by the number of months during which participants logged in to the OHC: logged in during 1 month only, logged in during 2 or 3 months, and logged in over a period of more than 3 months during the intervention phase. The participant characteristics were compared between the groups. All statistical analyses were performed using R version 3.6.3 (R Foundation for Statistical Computing). A significance level of α of .05 was assumed, and no adjustments were made for multiplicity.

Results

Participant Description

Participants (N=603) in this study had a mean age of 42.8 (SD 14.2) years, were 79.6% (n=480) female, 82.8% (n=499) were White, 82.4% (n=497) had some college education, and 69.8% (n=421) were employed (Table 1). At baseline, they reported a mean SF-12 PCS of 50.9 (SD 12.3), which is similar to that of the general US population, 50 (SD 10) [35]. The mean SF-12 MCS was 31.5 (SD 8.9), which is considerably lower than the 50 (SD 10) for the general US population [35]. The participants' mean PROMIS depression and anxiety scores were 62.3 (SD 12.3), and 65.9 (SD 12.3), respectively. Both indicate a higher severity of symptoms than the general US population, which had a mean score of 50 [39]. The 2 intervention arms did not differ in their baseline sociodemographic or clinical characteristics by random treatment assignment (all $P \geq 0.4$, previously published) [20].

Table 1. Baseline characteristics for intervention initiators and noninitiators.

Measure and category	Total (n=603)	Computerized cognitive behavioral therapy			Online health community		
		Initiation (n=521)	Noninitiation (n=82)	*P* value	Initiation (n=228)	Noninitiation (n=74)	*P* value
Demographics							
Age (y), mean (SD)	42.8 (14.2)	42.8 (14.2)	43.1 (14.4)	.84	42.2 (14)	43.9 (15.5)	.38
Sex, n (%)				.94			.31
Female	480 (79.6)	415 (79.7)	65 (79.3)		182 (79.8)	63 (85.1)	
Male	123 (20.4)	106 (20.3)	17 (20.7)		46 (20.2)	11 (14.9)	
Race, n (%)				.006			.09
White	499 (82.8)	441 (84.6)	58 (70.7)		188 (82.5)	54 (73)	
African American	91 (15.1)	69 (13.2)	22 (26.8)		34 (14.9)	19 (25.7)	
Other	13 (2.2)	11 (2.1)	2 (2.4)		6 (2.6)	1 (1.4)	
Living situation, n (%)				.06			.60
Alone with child	82 (13.6)	64 (12.3)	18 (22)		23 (10.1)	11 (14.9)	
Alone no child	114 (18.9)	100 (19.2)	14 (17.1)		45 (19.7)	15 (20.3)	
Living together with child	106 (17.6)	89 (17.1)	17 (20.7)		45 (19.7)	11 (14.9)	
Living together with no child	300 (49.8)	267 (51.3)	33 (40.2)		115 (50.4)	37 (50)	
Working, n (%)				.85			.15
Employed	421 (69.8)	363 (69.7)	58 (70.7)		159 (69.7)	45 (60.8)	
Other	182 (30.2)	158 (30.3)	24 (29.3)		69 (30.3)	29 (39.2)	
Education, n (%)				.02			<.001
High school or less	106 (17.6)	84 (16.1)	22 (26.8)		25 (11)	23 (31.1)	
Attended college but did not receive a 4-year degree (also business or technical school)	216 (35.8)	184 (35.3)	32 (39)		86 (37.7)	24 (32.4)	
College degree or higher	281 (46.6)	253 (48.6)	28 (34.1)		117 (51.3)	27 (36.5)	
Clinical characteristics							
Short Form Health Survey Physical Components Score, mean (SD)	50.9 (12.3)	51.4 (12.1)	47.7 (12.8)	.01	51.3 (12.5)	50.3 (11.8)	.56
PROMIS[a] pain interference, mean (SD)	31.5 (8.9)	31.3 (8.8)	32.4 (10)	.30	31.1 (8.8)	33.6 (10.9)	.04
Short Form Health Survey Mental Components Score, mean (SD)	17.8 (9.3)	17.5 (9.3)	19.7 (9.4)	.048	17.2 (9.2)	18.8 (9.7)	.19
PROMIS depression, mean (SD)	62.3 (6.2)	62.2 (6.4)	62.6 (4.8)	.55	62.2 (6.3)	61.5 (6.1)	.41
PROMIS anxiety, mean (SD)	65.9 (6.1)	65.9 (6)	65.7 (6.4)	.82	66 (6.1)	65.2 (6.2)	.35
PROMIS sleep impairment, mean (SD)	24.1 (5.8)	24 (5.8)	25.1 (5.6)	.12	24.2 (6)	24.6 (5.5)	.68
Mobile and internet use, n (%)							
Mobile phone use				.82			.32
Yes	405 (67.2)	349 (67)	56 (68.3)		162 (71.1)	48 (64.9)	
No	198 (32.8)	172 (33)	26 (31.7)		66 (28.9)	26 (35.1)	
Nonwork internet use				.13			.005
Never or rare	46 (7.6)	36 (6.9)	10 (12.2)		12 (5.3)	11 (14.9)	
Occasional	90 (14.9)	75 (14.4)	15 (18.3)		30 (13.2)	15 (20.3)	
Consistent	467 (77.4)	410 (78.7)	57 (69.5)		186 (81.6)	48 (64.9)	

Measure and category	Total (n=603)	Computerized cognitive behavioral therapy			Online health community		
		Initiation (n=521)	Noninitiation (n=82)	*P* value	Initiation (n=228)	Noninitiation (n=74)	*P* value
Work internet use				.38			.19
Never or rare	282 (46.8)	239 (45.9)	43 (52.4)		102 (44.7)	42 (56.8)	
Occasional	29 (4.8)	24 (4.6)	5 (6.1)		10 (4.4)	3 (4.1)	
Consistent	292 (48.4)	258 (49.5)	34 (41.5)		116 (50.9)	29 (39.2)	

[a]PROMIS: Patient-Reported Outcomes Measurement Information System.

Noninitiation Versus Initiation of CCBT Use

When pooling CCBT users from the CCBT-only and CCBT+OHC arms, 603 participants had access to the CCBT program. During the 6-month intervention, 13.6% (82/603) did not initiate the use of the program (Table 1). Those who did not initiate use were more likely to have a lower (worse) SF-12 PCS (*P*=.01), higher (worse) PROMIS Pain Interference measure (*P*=.048), and less likely to have a 4-year college degree (*P*=.02). The rate of noninitiation was 24% (22/91) for African American participants, 11.6% (58/499) for White participants, and 15% (2/13) for participants of "other" races (*P*=.006).

Discontinue Versus Continue Use of CCBT

Of those who initiated use (n=521), 97% (504/521) completed the first module (Table 2), and 12.9% (65/504) discontinued on completing module 1. If a patient was going to complete the first module, she or he did so during the first 3 months (495/504, 98%). Only an additional 9 (1.8%) participants completed the first module in the subsequent 3 months (data not shown).

Table 2. Baseline characteristics for sample, intervention users, and nonusers.

Categories, measure, and category	Total (n=603)	CCBT[a] user[b] (n=504)	CCBT nonuser[c] (n=99)	*P* value	OHC[d] user (n=228)	OHC nonuser (n=74)	*P* value
Demographics							
Age (y), mean (SD)	42.8 (14.2)	43 (14.2)	42 (14.4)	.54	42.2 (14)	43.9 (15.5)	.38
Sex, n (%)				.62			.31
Female	480 (79.6)	403 (80)	77 (77.8)		182 (79.8)	63 (85.1)	
Male	123 (20.4)	101 (20)	22 (22.2)		46 (20.2)	11 (14.9)	
Race, n (%)				.001			.09
White	499 (82.8)	429 (85.1)	70 (70.7)		188 (82.5)	54 (73)	
African American	91 (15.1)	64 (12.7)	27 (27.3)		34 (14.9)	19 (25.7)	
Other	13 (2.2)	11 (2.2)	2 (2)		6 (2.6)	1 (1.4)	
Living situation, n (%)				.048			.60
Alone with child	82 (13.6)	60 (11.9)	22 (22.2)		23 (10.1)	11 (14.9)	
Alone with no child	114 (18.9)	98 (19.5)	16 (16.2)		45 (19.7)	15 (20.3)	
Living together with child	106 (17.6)	88 (17.5)	18 (18.2)		45 (19.7)	11 (14.9)	
Living together with no child	300 (49.8)	257 (51.1)	43 (43.4)		115 (50.4)	37 (50)	
Working, n (%)				.65			.15
Employed	421 (69.8)	350 (69.4)	71 (71.7)		159 (69.7)	45 (60.8)	
Other	182 (30.2)	154 (30.6)	28 (28.3)		69 (30.3)	29 (39.2)	
Education, n (%)				<.001			<.001
High school or less	106 (17.6)	76 (15.1)	30 (30.3)		25 (11)	23 (31.1)	
Attended college but did not receive a 4-year degree (also business or technical school)	216 (35.8)	177 (35.1)	39 (39.4)		86 (37.7)	24 (32.4)	
College degree or higher	281 (46.6)	251 (49.8)	30 (30.3)		117 (51.3)	27 (36.5)	
Clinical characteristics, mean (SD)							
Short Form Health Survey Physical Components Score	50.9 (12.3)	51.4 (12.1)	48.4 (12.7)	.03	51.3 (12.5)	50.3 (11.8)	.56
PROMIS[e] pain interference	17.8 (9.3)	17.5 (9.3)	19.5 (9.4)	.049	17.2 (9.2)	18.8 (9.7)	.19
Short Form Health Survey Mental Components Score	31.5 (8.9)	31.3 (8.7)	32.2 (9.9)	.36	31.1 (8.8)	33.6 (10.9)	.04
PROMIS depression	62.3 (6.2)	62.2 (6.5)	62.6 (5)	.56	62.2 (6.3)	61.5 (6.1)	.41
PROMIS anxiety	65.9 (6.1)	65.8 (6.1)	65.9 (6.2)	.92	66 (6.1)	65.2 (6.2)	.35
PROMIS sleep impairment	24.1 (5.8)	24 (5.8)	24.8 (5.6)	.22	24.2 (6)	24.6 (5.5)	.68
Mobile and internet use							
Mobile phone use, n (%)				.56			.32
Yes	405 (67.2)	336 (66.7)	69 (69.7)		162 (71.1)	48 (64.9)	
No	198 (32.8)	168 (33.3)	30 (30.3)		66 (28.9)	26 (35.1)	
Nonwork internet use, n (%)				.12			.005
Never or rare	46 (7.6)	35 (6.9)	11 (11.1)		12 (5.3)	11 (14.9)	
Occasional	90 (14.9)	71 (14.1)	19 (19.2)		30 (13.2)	15 (20.3)	
Consistent	467 (77.4)	398 (79)	69 (69.7)		186 (81.6)	48 (64.9)	

Categories, measure, and category	Total (n=603)	CCBT[a] user[b] (n=504)	CCBT nonuser[c] (n=99)	*P* value	OHC[d] user (n=228)	OHC nonuser (n=74)	*P* value
Work internet use, n (%)				.22			.19
Never or rare	282 (46.8)	231 (45.8)	51 (51.5)		102 (44.7)	42 (56.8)	
Occasional	29 (4.8)	22 (4.4)	7 (7.1)		10 (4.4)	3 (4.1)	
Consistent	292 (48.4)	251 (49.8)	41 (41.4)		116 (50.9)	29 (39.2)	

[a]CCBT: computerized cognitive behavioral therapy.

[b]Completed the first CCBT module.

[c]Did not complete the first CCBT module.

[d]OHC: online health community.

[e]PROMIS: Patient-Reported Outcomes Measurement Information System.

Before completing the first module, 16.4% (99/603) discontinued CCBT. These participants were more likely at baseline to be a single parent living with their child, have a lower SF-12 PCS, have a higher PROMIS pain interference with life score, and less likely to have a 4-year college degree. African American participants were more likely to discontinue use than White participants (Figure S1 in Multimedia Appendix 1). Just over half (261/504, 51.8%) discontinued CCBT before completing 7 or 8 modules. Those who completed more modules were less likely to be mobile phone users, more likely to have a 4-year college degree, and more likely to have a higher severity of anxiety symptoms at baseline. There were no differences in baseline characteristics between those who completed module 7, started but did not finish module 8, or completed module 8. For analysis purposes, this group was defined as the completers of the CCBT. Those who initiated use but discontinued before completing were 64% (41/64) African Americans and 50.6% (217/429) White (*P*=.04).

Noninitiation Versus Initiation of OHC Use

During the 6-month intervention, 24.5% (74/302) of the participants with OHC access did not initiate use (Table 1). This group, when compared with those who initiated use, was less likely to have a 4-year college degree (*P*<.001), use the internet outside of work (*P*=.005), and more likely to have a better SF-12 MCS (*P*=.04). The rate of noninitiation was higher for African American participants (19/53, 36%) than for White participants (54/242, 22.3%; *P*=.04). Patients who waited to initiate use until sometime after the first month of having access had a higher PROMIS pain interference with life score (*P*=.03), poorer SF-12 PCS (*P*=.04), and were less likely to use the internet outside of work (*P*=.003).

Discontinue Versus Continue Use of the OHC

Of the 75% (n=228) of patients who initiated the use of the OHC during the 6-month intervention, 19% (43/228) discontinued use after logging in only once, 25% (56/228) discontinued use after logging in 2 to 3 times, and 56.6% (129/228) logged in 4 or more times (Figure S2 in Multimedia Appendix 1). Those who logged in fewer times tended to be younger (*P*=.03) and African American (*P*=.005).

Discussion

Principal Findings

In this study, the vast majority of patients initiated the use of the eHealth programs. Over half of the participants discontinued using CCBT before completion, and 43% (99/228) logged into the OHC 3 or fewer times. These latter groups provide an opportunity to explore areas where improvements in engagement methods and eHealth interventions could be made.

Noninitiation of CCBT and OHC

Patients who did not initiate the use of the CCBT or OHC programs were more likely to have less formal education, and those who did not initiate the use of the OHC also tended to be less frequent internet users. These findings are consistent with the conclusion that those who have less experience with technology may be less savvy or confident with technology, and that there is more reluctance to initiate use of web-based programs. Reluctance due to one's level of digital or health literacy is addressable at the start by helping to develop greater confidence in the technologies involved in the interventions and the ability of the contents to be effective. In this study, all participants needed to have access to the necessary technologies for participation; however, the care managers also offered telephone assistance with technical difficulties with the e-mental health programs. Some may have felt stigmatized asking for help, particularly with their personal technologies; felt that phone guidance would not work; or felt that this was not an offer to assist with using the programs or their personal technologies but for technical issues outside of their control.

Poor physical well-being, as indicated by lower SF-12 PCS scores or higher pain interference with life, also reduced the likelihood of patients initiating use. This raises several possibilities. First, it may be that these patients had more difficulty using the technologies and websites due to physical limitations and associated pain or they may have had a reduced ability to concentrate on the contents. Another possibility is that they were less enthusiastic about the interventions because they were not directly focused on physical health issues; thus, they were perceived as not addressing their high-priority need. However, it has been well established that chronic pain, depression, anxiety, and somatic amplification co-occur [40]. Such issues would be important to address during early engagement activities designed to both develop confidence in the ability of the interventions to help and to assess and address reasons for patients' hesitancy to use the interventions.

Those who did not initiate the use of the interventions during the first month of the trial were unlikely to later. If the methods designed to develop engagement with the interventions did not sufficiently influence a patient during the initial weeks, they were unlikely to have a subsequent effect. This may help identify an area where certain patients could benefit from earlier or additional methods to increase their engagement.

Discontinuation of CCBT

Patients who did not complete the first CCBT module were likely to be less technology savvy, have poorer physical functioning, have higher pain interference with life, and have less severe anxiety symptoms at baseline. Mental health burden, comfort with technology, and physical functioning were recurring influences on patient initiation and discontinuation. Others have also found that an increased mental health burden is associated with greater use of e-mental health interventions for anxiety and depressive symptoms [41]. A counterintuitive finding was that those less likely to be mobile phone users were more likely to complete a higher number of CCBT modules. This could be because they had less access to web-based resources, which made the offered treatments a more unique and valued opportunity to receive mental health services.

Discontinuation of CCBT and OHC

Given that over half of the participants discontinued the CCBT program before completion, this may indicate the need for additional ongoing methods to maintain users' buy-in. It should be noted that CCBT did have a positive effect on clinical outcomes compared with UC [20], indicating that to receive benefit, at least partial benefit, it was not necessary to complete the program. Some patients may have felt that the program was no longer addressing their needs or that they improved sufficiently and discontinued [42]. This raises the issue of whether setting patients' treatment expectations, for example, for when full benefit has been gained, would influence discontinuation and improve outcomes for those who leave before completion. Examining patients' reasons for discontinuing would provide patient-centered insights on this issue and potentially how the interventions could be tailored as needs change during use.

African American participants who initiated the use were more likely than White participants to discontinue OHC use after only one login and more likely to discontinue CCBT use before completing the program. Taken together, these findings indicate that more could be done to facilitate the engagement of African Americans, and possibly other minority groups. This may also indicate that content adaptations could improve the intervention's fit for diverse groups [38]. Digital literacy and internet connectivity, which include skills, confidence, connectivity and its ongoing affordability, level of access, devices, training, and technical support, have been called the "super social determinants of health" because they influence all other social determinants of health [43]. The study examined only a few social determinants of health, for example, employment and education. It is likely that other social determinants of health influenced these findings, and this is a key area for future investigations.

The Importance of Strategies to Facilitate Engagement

Only a minority of patients did not initiate the use of e-mental eHealth interventions, and almost half of them completed CCBT. This argues for the effectiveness of the engagement offered by the ongoing collaborative care strategies used. This is supported by the findings from 2 earlier studies on CCBT. The original study that established the efficacy of CCBT provided in-person treatment at a research office [21]. One module was completed during each visit. Each user was provided with 1:1 supervision from a practice nurse. The nurse ensured that each participant interacted successfully with the computer and treatment program. A subsequent study examined the web-based home delivery of CCBT. Investigators found no effect of CCBT when compared with UC [44]. In this study, although participants were called weekly at their homes by a research staff member who encouraged continued program use, the amount of contact over the 4-month study period was an average of only 23 seconds per week.

These findings support the potential of providing eHealth programs via collaborative care and supported and "guided" models [45]. They also point to the potential of including additional and improved methods to facilitate engagement. Highly effective engagement practices, research findings, and theoretical models have identified several best practices to engage patients with interventions. These identify at least 5 characteristics of an intervention model that facilitate strong engagement [14,17,46-49]. One is compatibility. Engagement activities should establish how the intervention will be personally meaningful and able to meet the needs a user finds important. For example, it will facilitate better illness management or an improved ability to socialize or participate in work. In this study, it may have been helpful if a patient's PCP or care manager had introduced the programs and engaged in shared decision-making to identify specific benefits to the patient. In addition, participants may have considered the interventions more of a study than personalized treatment because they were not presented by their PCP. Another characteristic is relative advantage. This is the level of belief that the effort and resources involved will provide sufficient added value over alternatives. In total, 2 alternatives that patients commonly consider are not initiating the treatment and/or discontinuing. The study's CCBT program may have been too rigid for some patients, as they could not skip a module or tailor the program to their immediate needs. This feedback was provided to the care managers. Several existing and recent eHealth programs have developed more flexible approaches that allow users to move through the program and tailor the presentation according to their current needs and previous experience with mental health treatment [50-52]. However, many eHealth interventions with fixed presentations are highly effective [53]. Creating approaches that can accommodate diverse and even changing needs of patients may help support broader engagement. A third characteristic is the complexity of the intervention. This is the extent to which the intervention and required technologies are intuitive to use either before or after training. This can be especially important for users without sufficient digital or health literacy, and in-person training may be beneficial. Fourth characteristic is observability. This is the extent to which it will be easy for users to see improvements in issues that are important to them because of their use of the intervention. This may have manifested in 2 opposite ways. Some may have perceived initial improvement and thought that it was all that could be expected. Others may have felt hampered by the rigidity of the CCBT program because patients may have had to complete modules that did not seem relevant to them; thus, they did not see added benefit in continuing the program. Both these factors could have led to discontinuation. The fifth characteristic is supportive accountability. This includes support from a human coach or care manager who can provide guidance and develop accountability on the part of users. This can increase adherence. In this study, the personalized and continuous support provided via the collaborative care strategy, and knowing that care managers would call to check on progress and issues around completing homework, likely contributed to the high initiation and ongoing use rates. The ability of such "guidance" to improve initiation and adherence by some users has been documented [45,51]. This argues for the considerable advantage of guided collaborative care models such as was used in this study.

Conclusions

These analyses produced several findings: those with greater mental health needs had a greater predisposition to initiate and continue use, those with poorer physical well-being were less likely to initiate use, and comfort with technologies seemed to influence patients' likelihood of initiating and continuing use. These findings support the conclusion that methods are needed to build engagement with eHealth interventions that can be tailored to these as well as other specific needs of patients.

Although discontinuation has been explored to a somewhat greater extent in the eHealth literature, less attention has been devoted to noninitiation, and neither has been thoroughly investigated. The findings highlight a potentially important need for additional studies of both noninitiation and discontinuation and the relevance of these 2 phenomena to the types of personalized engagement and ongoing support methods that could benefit users and help address the gaps that lead to Eisenbach's [17] Law of Attrition.

Acknowledgments

This trial was supported by grant R01 MH093501 from the National Institute of Mental Health.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Number of computerized cognitive behavioral therapy program modules completed and number of logins to the online health community platform.

[DOCX File , 348 KB - mental_v11i1e52197_app1.docx]

References

1. Olfson M, Blanco C, Marcus SC. Treatment of adult depression in the United States. JAMA Intern Med 2016 Oct 01;176(10):1482-1491. [doi: 10.1001/jamainternmed.2016.5057] [Medline: 27571438]
2. Kazdin AE. Addressing the treatment gap: a key challenge for extending evidence-based psychosocial interventions. Behav Res Ther 2017 Jan;88:7-18. [doi: 10.1016/j.brat.2016.06.004] [Medline: 28110678]
3. Wilson W, Bangs A, Hatting T. Building a national plan for rural behavioral health: first steps—workforce development. National Rural Health Administration. 2007. URL: https://www.ruralhealth.us/getattachment/Advocate/Policy-Documents/FutureofRuralHealthFeb-2013.pdf.aspx [accessed 2022-08-04]
4. Leichsenring F, Steinert C, Rabung S, Ioannidis JP. The efficacy of psychotherapies and pharmacotherapies for mental disorders in adults: an umbrella review and meta-analytic evaluation of recent meta-analyses. World Psychiatry 2022 Feb;21(1):133-145 [FREE Full text] [doi: 10.1002/wps.20941] [Medline: 35015359]
5. Kessler RC, Kazdin AE, Aguilar-Gaxiola S, Al-Hamzawi A, Alonso J, Altwaijri YA, WHO World Mental Health Survey collaborators. Patterns and correlates of patient-reported helpfulness of treatment for common mental and substance use disorders in the WHO World Mental Health Surveys. World Psychiatry 2022 Jun;21(2):272-286 [FREE Full text] [doi: 10.1002/wps.20971] [Medline: 35524618]
6. Leichsenring F, Steinert C, Ioannidis JP. Toward a paradigm shift in treatment and research of mental disorders. Psychol Med 2019 Oct;49(13):2111-2117. [doi: 10.1017/S0033291719002265] [Medline: 31474241]
7. Simon GE, Ding V, Hubbard R, Fishman P, Ludman E, Morales L, et al. Early dropout from psychotherapy for depression with group- and network-model therapists. Adm Policy Ment Health 2012 Nov;39(6):440-447 [FREE Full text] [doi: 10.1007/s10488-011-0364-x] [Medline: 21710256]
8. Griffiths K, Christensen H. Review of randomised controlled trials of Internet interventions for mental disorders and related conditions. Clin Psychol 2007 Feb 03;10(1):16-29 [FREE Full text] [doi: 10.1080/13284200500378696]

9. Rogers MA, Lemmen K, Kramer R, Mann J, Chopra V. Internet-delivered health interventions that work: systematic review of meta-analyses and evaluation of website availability. J Med Internet Res 2017 Mar 24;19(3):e90 [FREE Full text] [doi: 10.2196/jmir.7111] [Medline: 28341617]
10. Mamukashvili-Delau M, Koburger N, Dietrich S, Rummel-Kluge C. Efficacy of computer- and/or internet-based cognitive-behavioral guided self-management for depression in adults: a systematic review and meta-analysis of randomized controlled trials. BMC Psychiatry 2022 Nov 24;22(1):730 [FREE Full text] [doi: 10.1186/s12888-022-04325-z] [Medline: 36424570]
11. Alkureishi MA, Choo ZY, Rahman A, Ho K, Benning-Shorb J, Lenti G, et al. Digitally disconnected: qualitative study of patient perspectives on the digital divide and potential solutions. JMIR Hum Factors 2021 Dec 15;8(4):e33364 [FREE Full text] [doi: 10.2196/33364] [Medline: 34705664]
12. Rotondi AJ, Grady J, Hanusa BH, Haas GL, Spring MR, Abebe KZ, et al. Key variables for developing effective e-health intervention designs for those with, and without, mental health disorders: navigational depth, number of hyperlinks per page, use of a popup menu, reading level, page length, participant abilities with the technology. J Med Internet Res 2021 Mar 24;23(3):e23137 [FREE Full text] [doi: 10.2196/23137] [Medline: 33759796]
13. Saleem M, Kühne L, De Santis KK, Christianson L, Brand T, Busse H. Understanding engagement strategies in digital interventions for mental health promotion: scoping review. JMIR Ment Health 2021 Dec 20;8(12):e30000 [FREE Full text] [doi: 10.2196/30000] [Medline: 34931995]
14. Chiauzzi E, Newell A. Mental health apps in psychiatric treatment: a patient perspective on real world technology usage. JMIR Ment Health 2019 Apr 22;6(4):e12292. [doi: 10.2196/12292] [Medline: 31008711]
15. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018 Aug;21(3):116-119. [doi: 10.1136/eb-2018-102891] [Medline: 29871870]
16. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
17. Eysenbach G. The law of attrition. J Med Internet Res 2005 Mar 31;7(1):e11 [FREE Full text] [doi: 10.2196/jmir.7.1.e11] [Medline: 15829473]
18. Torous J, Lipschitz J, Ng M, Firth J. Dropout rates in clinical trials of smartphone apps for depressive symptoms: a systematic review and meta-analysis. J Affect Disord 2020 Feb 15;263:413-419. [doi: 10.1016/j.jad.2019.11.167] [Medline: 31969272]
19. Fleming T, Bavin L, Lucassen M, Stasiak K, Hopkins S, Merry S. Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. J Med Internet Res 2018 Jun 06;20(6):e199 [FREE Full text] [doi: 10.2196/jmir.9275] [Medline: 29875089]
20. Rollman B, Herbeck Belnap B, Abebe KZ, Spring MB, Rotondi AJ, Rothenberger SD, et al. Effectiveness of online collaborative care for treating mood and anxiety disorders in primary care: a randomized clinical trial. JAMA Psychiatry 2018 Jan 01;75(1):56-64 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.3379] [Medline: 29117275]
21. Proudfoot J, Goldberg D, Mann A, Everitt B, Marks I, Gray JA. Computerized, interactive, multimedia cognitive-behavioural program for anxiety and depression in general practice. Psychol Med 2003 Feb 14;33(2):217-227. [doi: 10.1017/s0033291702007225] [Medline: 12622301]
22. Cavanagh K, Shapiro DA, Van Den Berg S, Swain S, Barkham M, Proudfoot J. The effectiveness of computerized cognitive behavioural therapy in routine care. Br J Clin Psychol 2006 Nov;45(Pt 4):499-514. [doi: 10.1348/014466505X84782] [Medline: 17076960]
23. Learmonth D, Rai S. Taking computerized CBT beyond primary care. Br J Clin Psychol 2008 Mar;47(Pt 1):111-118. [doi: 10.1348/014466507X248599] [Medline: 17939879]
24. Fortuna KL, Naslund JA, LaCroix JM, Bianco CL, Brooks JM, Zisman-Ilani Y, et al. Digital peer support mental health interventions for people with a lived experience of a serious mental illness: systematic review. JMIR Ment Health 2020 Apr 03;7(4):e16460 [FREE Full text] [doi: 10.2196/16460] [Medline: 32243256]
25. Skousen T, Safadi H, Young C, Karahanna E, Safadi S, Chebib F. Successful moderation in online patient communities: inductive case study. J Med Internet Res 2020 Mar 17;22(3):e15983 [FREE Full text] [doi: 10.2196/15983] [Medline: 32181743]
26. Naslund JA, Aschbrenner KA, Marsch LA, McHugo GJ, Bartels SJ. Facebook for supporting a lifestyle intervention for people with major depressive disorder, bipolar disorder, and schizophrenia: an exploratory study. Psychiatr Q 2018 Mar;89(1):81-94. [doi: 10.1007/s11126-017-9512-0] [Medline: 28470468]
27. Prescott J, Rathbone AL, Brown G. Online peer to peer support: qualitative analysis of UK and US open mental health Facebook groups. Digit Health 2020 Dec 10;6:2055207620979209 [FREE Full text] [doi: 10.1177/2055207620979209] [Medline: 33354335]
28. Merchant R, Goldin A, Manjanatha D, Harter C, Chandler J, Lipp A, et al. Opportunities to expand access to mental health services: a case for the role of online peer support communities. Psychiatr Q 2022 Jun 28;93(2):613-625 [FREE Full text] [doi: 10.1007/s11126-022-09974-7] [Medline: 35226251]

29. Easton K, Diggle J, Ruethi-Davis M, Holmes M, Byron-Parker D, Nuttall J, et al. Qualitative exploration of the potential for adverse events when using an online peer support network for mental health: cross-sectional survey. JMIR Ment Health 2017 Oct 30;4(4):e49 [FREE Full text] [doi: 10.2196/mental.8168] [Medline: 29084710]
30. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]
31. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
32. Gordon AJ, Maisto SA, McNeil M, Kraemer KL, Conigliaro RL, Kelley ME, et al. Three questions can detect hazardous drinkers. J Fam Pract 2001 Apr;50(4):313-320. [Medline: 11300982]
33. Proudfoot J, Ryden C, Everitt B, Shapiro DA, Goldberg D, Mann A, et al. Clinical efficacy of computerised cognitive-behavioural therapy for anxiety and depression in primary care: randomised controlled trial. Br J Psychiatry 2004 Jul 02;185(1):46-54. [doi: 10.1192/bjp.185.1.46] [Medline: 15231555]
34. Geramita EM, Herbeck Belnap BH, Abebe KZ, Rothenberger SD, Rotondi AJ, Rollman BL. The association between increased levels of patient engagement with an internet support group and improved mental health outcomes at 6-month follow-up: post-hoc analyses from a randomized controlled trial. J Med Internet Res 2018 Jul 17;20(7):e10402 [FREE Full text] [doi: 10.2196/10402] [Medline: 30021711]
35. Ware Jr J, Kosinski M, Keller SD. A 12-item short-form health survey: construction of scales and preliminary tests of reliability and validity. Med Care 1996 Mar;34(3):220-233. [doi: 10.1097/00005650-199603000-00003] [Medline: 8628042]
36. Amtmann D, Cook KF, Jensen MP, Chen WH, Choi S, Revicki D, et al. Development of a PROMIS item bank to measure pain interference. Pain 2010 Jul;150(1):173-182 [FREE Full text] [doi: 10.1016/j.pain.2010.04.025] [Medline: 20554116]
37. Spitzer RL, Williams JB, Kroenke K, Linzer M, deGruy 3rd FV, Hahn SR, et al. Utility of a new procedure for diagnosing mental disorders in primary care. The PRIME-MD 1000 study. JAMA 1994 Dec 14;272(22):1749-1756. [Medline: 7966923]
38. Jonassaint CR, Belnap BH, Huang Y, Karp JF, Abebe KZ, Rollman BL. Racial differences in the effectiveness of internet-delivered mental health care. J Gen Intern Med 2020 Feb 19;35(2):490-497 [FREE Full text] [doi: 10.1007/s11606-019-05542-1] [Medline: 31745855]
39. Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D, PROMIS Cooperative Group. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): depression, anxiety, and anger. Assessment 2011 Sep;18(3):263-283 [FREE Full text] [doi: 10.1177/1073191111411667] [Medline: 21697139]
40. Kim S, Lee J, Boone D. Protective and risk factors at the intersection of chronic pain, depression, anxiety, and somatic amplification: a latent profile approach. J Pain Res 2022 Apr 15;15:1107-1121 [FREE Full text] [doi: 10.2147/JPR.S340382] [Medline: 35450061]
41. Christensen H, Griffiths KM, Farrer L. Adherence in internet interventions for anxiety and depression. J Med Internet Res 2009 Apr 24;11(2):e13 [FREE Full text] [doi: 10.2196/jmir.1194] [Medline: 19403466]
42. Amagai S, Pila S, Kaat AJ, Nowinski CJ, Gershon RC. Challenges in participant engagement and retention using mobile health apps: literature review. J Med Internet Res 2022 Apr 26;24(4):e35120. [doi: 10.2196/35120] [Medline: 35471414]
43. Sieck CJ, Sheon A, Ancker JS, Castek J, Callahan B, Siefer A. Digital inclusion as a social determinant of health. NPJ Digit Med 2021 Mar 17;4(1):52 [FREE Full text] [doi: 10.1038/s41746-021-00413-8] [Medline: 33731887]
44. Gilbody S, Littlewood E, Hewitt C, Brierley G, Tharmanathan P, Araya R, REEACT Team. Computerised cognitive behaviour therapy (cCBT) as treatment for depression in primary care (REEACT trial): large scale pragmatic randomised controlled trial. BMJ 2015 Nov 11;351:h5627 [FREE Full text] [doi: 10.1136/bmj.h5627] [Medline: 26559241]
45. Zarski A, Lehr D, Berking M, Riper H, Cuijpers P, Ebert DD. Adherence to internet-based mobile-supported stress management: a pooled analysis of individual participant data from three randomized controlled trials. J Med Internet Res 2016 Jun 29;18(6):e146 [FREE Full text] [doi: 10.2196/jmir.4493] [Medline: 27357528]
46. Anderson CM, Reiss DJ, Hogarty GE. Schizophrenia and the Family: A Practitioner's Guide to Psychoeducation and Management. New York, NY: The Guilford Press; 1986.
47. Fogg BJ. A behavior model for persuasive design. In: Proceedings of the 4th International Conference on Persuasive Technology. 2009 Presented at: Persuasive '09; April 26-29, 2009; Claremont, CA p. 1-7 URL: https://dl.acm.org/doi/10.1145/1541948.1541999 [doi: 10.1145/1541948.1541999]
48. Rogers EM. Diffusion of Innovations, 4th Edition. New York, NY: The Free Press; 1995.
49. Mohr DC, Cuijpers P, Lehman K. Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions. J Med Internet Res 2011 Mar 10;13(1):e30 [FREE Full text] [doi: 10.2196/jmir.1602] [Medline: 21393123]
50. Mohr DC, Tomasino KN, Lattie EG, Palac HL, Kwasny MJ, Weingardt K, et al. IntelliCare: an eclectic, skills-based app suite for the treatment of depression and anxiety. J Med Internet Res 2017 Jan 05;19(1):e10 [FREE Full text] [doi: 10.2196/jmir.6645] [Medline: 28057609]
51. Brunette MF, Rotondi AJ, Ben-Zeev D, Gottlieb JD, Mueser KT, Robinson DG, et al. Coordinated technology-delivered treatment to prevent rehospitalization in schizophrenia: a novel model of care. Psychiatr Serv 2016 Apr 01;67(4):444-447. [doi: 10.1176/appi.ps.201500257] [Medline: 26725297]

52. Rotondi AJ, Anderson CM, Haas GL, Eack SM, Spring MB, Ganguli R, et al. Web-based psychoeducational intervention for persons with schizophrenia and their supporters: one-year outcomes. Psychiatr Serv 2010 Nov;61(11):1099-1105. [doi: 10.1176/ps.2010.61.11.1099] [Medline: 21041348]
53. Ritterband LM, Thorndike FP, Ingersoll KS, Lord HR, Gonder-Frederick L, Frederick C, et al. Effect of a web-based cognitive behavior therapy for insomnia intervention with 1-year follow-up: a randomized clinical trial. JAMA Psychiatry 2017 Jan 01;74(1):68-75. [doi: 10.1001/jamapsychiatry.2016.3249] [Medline: 27902836]

Abbreviations

CCBT: computerized cognitive behavioral therapy
ISG: internet support group
MCS: Mental Components Score
OHC: online health community
PCP: primary care physician
PCS: Physical Components Score
PHQ-9: 9-item Patient Health Questionnaire
PROMIS: Patient-Reported Outcomes Measurement Information System
RCT: randomized clinical trial
SF-12: Short Form Health Survey
UC: usual care

Edited by J Torous; submitted 26.08.23; peer-reviewed by C Benjet, S Pila; comments to author 03.10.23; revised version received 06.11.23; accepted 07.11.23; published 17.01.24.

Please cite as:
Rotondi AJ, Belnap BH, Rothenberger S, Feldman R, Hanusa B, Rollman BL
Predictors of Use and Drop Out From a Web-Based Cognitive Behavioral Therapy Program and Health Community for Depression and Anxiety in Primary Care Patients: Secondary Analysis of a Randomized Controlled Trial
JMIR Ment Health 2024;11:e52197
URL: https://mental.jmir.org/2024/1/e52197
doi:10.2196/52197
PMID:38231552

Original Paper

Testing the Efficacy of a Brief, Self-Guided Mindfulness Ecological Momentary Intervention on Emotion Regulation and Self-Compassion in Social Anxiety Disorder: Randomized Controlled Trial

Nur Hani Zainal[1,2], BS, MS, PhD; Hui Han Tan[1], BS, MS; Ryan Yee Shiun Hong[1], BS, MS, PhD; Michelle Gayle Newman[3], BS, MA, PhD

[1]Department of Psychology, National University of Singapore, Singapore, Singapore
[2]Department of Health Policy, Harvard Medical School, Boston, MA, United States
[3]Department of Psychology, The Pennsylvania State University, University Park, PA, United States

Corresponding Author:
Nur Hani Zainal, BS, MS, PhD
Department of Psychology
National University of Singapore
9 Arts Link
Singapore, 117572
Singapore
Phone: 65 917 767 7088
Fax: 65 6516 3749
Email: hanizainal@nus.edu.sg

Abstract

Background: Theories propose that brief, mobile, self-guided mindfulness ecological momentary interventions (MEMIs) could enhance emotion regulation (ER) and self-compassion. Such changes are posited to be mechanisms of change. However, rigorous tests of these theories have not been conducted.

Objective: In this assessor-blinded, parallel-group randomized controlled trial, we aimed to test these theories in social anxiety disorder (SAD).

Methods: Participants with SAD (defined as having a prerandomization cut-off score ≥20 on the Social Phobia Inventory self-report) were randomized to a 14-day fully self-guided MEMI (96/191, 50.3%) or self-monitoring app (95/191, 49.7%) arm. They completed web-based self-reports of 6 clinical outcome measures at prerandomization, 15-day postintervention (administered the day after the intervention ended), and 1-month follow-up time points. ER and self-compassion were assessed at preintervention and 7-day midintervention time points. Multilevel modeling determined the efficacy of MEMI on ER and self-compassion domains from pretrial to midintervention time points. Bootstrapped parallel multilevel mediation analysis examined the mediating role of pretrial to midintervention ER and self-compassion domains on the efficacy of MEMI on 6 clinical outcomes.

Results: Participants demonstrated strong compliance, with 78% (149/191) engaging in at least 80% of the MEMI and self-monitoring prompts. MEMI was more efficacious than the self-monitoring app in decreasing ER goal–directed behavior difficulties (between-group Cohen d=−0.24) and lack of emotional clarity (Cohen d=0.16) and increasing self-compassion social connectedness (Cohen d=0.19), nonidentification with emotions (Cohen d=0.16), and self-kindness (Cohen d=0.19) from pretrial to midintervention time points. The within-group effect sizes from pretrial to midintervention were larger in the MEMI arm than in the self-monitoring app arm (ER goal–directed behavior difficulties: Cohen d=−0.73 vs −0.29, lack of emotional clarity: Cohen d=−0.39 vs −0.21, self-compassion domains of social connectedness: Cohen d=0.45 vs 0.19, nonidentification with emotions: Cohen d=0.63 vs 0.48, and self-kindness: Cohen d=0.36 vs 0.10). Self-monitoring, but not MEMI, alleviated ER emotional awareness issues (between-group Cohen d=0.11 and within-group: Cohen d=−0.29 vs −0.13) and reduced self-compassion acknowledging shared human struggles (between-group Cohen d=0.26 and within-group: Cohen d=−0.23 vs 0.13). No ER and self-compassion domains were mediators of the effect of MEMI on SAD symptoms (P=.07-<.99), generalized anxiety symptoms (P=.16-.98), depression severity (P=.20-.94), repetitive negative thinking (P=.12-.96), and trait mindfulness (P=.18-.99) from pretrial to postintervention time points. Similar nonsignificant mediation effects emerged for all of these clinical outcomes from pretrial to 1-month follow-up time points (*P*=.11-.98).

Conclusions: Brief, fully self-guided, mobile MEMIs efficaciously increased specific self-compassion domains and decreased ER difficulties associated with goal pursuit and clarity of emotions from pretrial to midintervention time points. Higher-intensity MEMIs may be required to pinpoint the specific change mechanisms in ER and self-compassion domains of SAD.

Trial Registration: Open Science Framework (OSF) Registries; osf.io/m3kxz https://osf.io/m3kxz

(JMIR Ment Health 2024;11:e53712) doi:10.2196/53712

KEYWORDS

social anxiety disorder; mindfulness; ecological momentary intervention; randomized controlled trial; emotion regulation; self-compassion; mechanisms of change; mobile phone; momentary interventions; self-monitoring app; regulations; participant

Introduction

Background

Emotion regulation (ER) has been defined as a deliberate action taken to modulate emotional states [1]. Situations can trigger emotional reactions, including short- and long-term emotional responses [2]. Persistent suboptimal behaviors that limit opportunities and quality of life are inherent to various emotional disorders, such as avoidance of feared situations in social anxiety disorder (SAD) [3], brooding in depression [4], and impulsivity in attention-deficit/hyperactivity disorder [5]. Elevating self-compassion could facilitate better ER by fostering discomfort tolerance [6]. Self-compassion is described as accepting one's experience without judgment and showing oneself kindness [7-9]. Increased self-judgment might create anxiety and depression via emotional dysregulation. Such anxiety and depression could lead to avoidance as a way to manage negative affect [10]. Chronic avoidance of social situations and failure to optimally face challenges head-on can hinder one from building self-efficacy; decrease pleasure derived from recreational activities; and restrict work, school, and other life opportunities [11-13]. Therefore, developing efficacious interventions to improve ER and self-compassion is essential.

Mindfulness-based interventions (MBIs) might enhance ER and self-compassion simultaneously. The monitoring and acceptance theory proposed that MBIs could boost ER by increasing acceptance, curiosity, and equanimity, coinciding with better interpersonal functioning [14]. This theory suggests that enhancing various outcomes in MBIs relies on integrating attention monitoring and acceptance skills [15]. It proposes that acceptance functions as an ER skill that promotes nonreactivity, self-compassion, and receptiveness to current experiences. The omission of acceptance skills training could potentially negate the stress-buffering and social cognitive advantages of MBIs. Consistent with the monitoring and acceptance theory, meta-analyses have shown inverse relations among self-compassion and anxiety, depressive symptoms, and stress [16] and found that MBIs were superior to controls in improving mental health symptoms and self-compassion [17]. However, traditional in-person MBIs, such as mindfulness-based stress reduction [18] and mindfulness-based cognitive therapy [19], typically involve significant financial costs, time commitments, and potential travel requirements, often extending from 1 hour to 2.5 hours per week across 8 to 16 sessions with 6-hour retreat sessions [20]. Individuals with high levels of emotional dysregulation and self-compassion deficits tend to self-report shame, stigma, and other hindrances to treatment seeking [21,22]. Those experiencing these struggles may not seek in-person therapy or may not remain in treatment, fearing the negative emotions associated with the treatment process. Collectively, such research emphasizes the importance of developing scalable (eg, app based) and concise MBIs.

Concise and scalable MBIs, typically lasting up to 2 weeks [23], may enhance ER and self-compassion. For instance, enhanced ER skills and reduced stress levels were observed among smokers and nonsmokers after a 5-day MBI, compared to ER and stress observed before the intervention [24]. Similarly, a single-session MBI was associated with increased cognitive reappraisal and reduced emotional suppression (ER domains) among graduate students over 3 months [25]. However, the absence of a control arm in these studies precluded knowing whether the brief MBI would outperform an active control, an essential aspect for drawing causal inferences [26]. Another study, a 14-day MBI randomized controlled trial (RCT) [27], implied this possibility by showing that acceptance was critical in effectively reducing loneliness and enhancing social engagement in everyday experiences. Another RCT showed that a single-session MBI was linked to better ER in trauma-distressed people [28]. Relatedly, 5-session MBIs consistently outperformed controls, raising self-compassion and mindfulness among undergraduates [29,30]. Together, these studies suggested that brief mindfulness ecological momentary interventions (MEMIs) that repeatedly instructed mindfulness exercises and recorded symptom levels in real time could outperform an active control in enhancing ER and self-compassion over time in clinical samples.

Moreover, improved ER and self-regulation may be a theoretical change mechanism explaining why brief, fully self-guided mobile MEMIs conferred mental health benefits such as symptom reduction over time [31,32]. Such theories posit that brief, fully self-guided mobile MEMIs may enhance the capacity to observe internal reactions in emotionally charged situations, enabling individuals to recognize when they are caught in their emotions, to pause to regain composure before reacting, and to practice self-kindness. Indeed, 2 recent experiments of diverse mindfulness practices showed that state-level and trait-level self-compassion improvements were linked to increased self-guided mindfulness practices and quality of life across 14 days in nonclinical samples [33]. Relatedly, baseline higher trait observing and describing mindfulness facets predicted reduced anxiety and depression via enhanced ER across an internet-delivered 8-week MBI in healthy adults [34]. This finding might extend to brief, fully self-guided mobile MEMIs for clinical samples. The dearth of mediation analyses testing the mechanistic role of ER and self-compassion highlights the importance of conducting such studies, particularly among individuals with emotional disorders. Discovering treatment mediators (proxy change mechanisms) might aid with fine-tuning and refining existing brief, scalable MEMIs for clinical samples [35].

SAD presents as one potential clinical sample case in point. Fostering better ER and self-compassion as treatment targets via brief, fully self-guided, mobile MEMIs could be especially advantageous for individuals with SAD, as they tend to exhibit poor ER, such as excessive avoidance and difficulties applying cognitive modification techniques [36]. Teaching nonjudgment and nonreactivity skills in SAD via the MEMI might pave the way for fostering cognitive modification and other ER skills and promote active efforts to increase exposure to various interpersonal situations. Furthermore, self-compassion has shown inverse associations with overall SAD symptoms and specific cognitive processes, such as the fear of both negative and positive evaluation [37] and postevent ruminative processing [38]. Those with versus without SAD felt it was essential to control their emotions, did not believe in emotion malleability [39], and had trouble practicing acceptance of emotional responses [40]. Therefore, MEMIs instructing present-moment focus and acceptance of myriad emotional experiences could remedy these psychological rigidities by enhancing ER skills, thereby improving a wide array of responses to fear-inducing social situations. Construed clinically, these studies suggest that individuals with SAD have room for growth in terms of benefiting from brief, fully self-guided, mobile MEMIs to reduce various ER difficulties and enhance diverse self-compassion domains across time.

Such ER difficulties could pertain to *nonacceptance of emotions* (proclivity to exhibit nonaccepting responses to one's distress) and *emotional awareness issues* (the inclination to focus upon and recognize emotions within oneself). They might also be linked to a *lack of emotional clarity* (lucidity regarding present-moment emotions) and *goal-directed behavior difficulties* (maintaining focus and achieving tasks during episodes of negative emotional states). Furthermore, ER difficulties could encompass *impulse control issues* (inability to sustain behavioral self-regulation amid negative emotional states) and *strategy use problems* (lack of conviction for having the ability to effectively regulate emotions when feeling upset [41]). The facets of self-compassion might also pertain to *acknowledging shared human struggles* (acknowledgment of collective human experience, recognizing that all individuals encounter failures, make errors, and navigate imperfect life trajectories) and *social connectedness* (a sense of being linked socially with other humans). Cultivating self-compassion also embraces *mindfulness* (a state of cognizant immersion in one's immediate experience, characterized by clarity and equilibrium) and *nonidentification with emotions* (absence or lack of tendency to become ensnared in an exaggerated narrative about the adverse facets of one's self or life experiences). In addition, self-compassion includes *nonjudgment toward oneself* (absence or lack of tendency to be overly self-critical) and *self-kindness* (practice of approaching oneself with support and understanding) [42].

Objectives

On the basis of the theory and logic outlined, the aims of this study were 2-fold. First, we hypothesized that a 14-day MEMI would be superior to a self-monitoring control condition in reducing various domains of ER difficulties and enhancing self-compassion from pretrial to midintervention time points across 7 days (hypothesis 1). Specifically, we examined the 6 difficulties in ER and self-compassion domains mentioned above as pretrial to midintervention outcomes. Second, we hypothesized that the effect of the 14-day MEMI versus self-monitoring app on change in SAD-related outcomes from pretrial to postintervention time points (pre-post; hypothesis 2a) or prerandomization to 1-month follow-up (1MFU) time points (pre-1MFU; hypothesis 2b) would be mediated via a change in ER difficulties and self-compassion domains assessed from pretrial to midintervention time points. Specifically, the outcomes examined were SAD symptoms, generalized anxiety symptoms, depression severity, repetitive negative thinking, and trait mindfulness. Our study was an extension of a primary RCT, which showed that brief MEMI and self-monitoring app led to sustained changes in all of these clinical outcomes, with small-to-large effect sizes from pre-post and pre-1MFU time points. There were no differences between MEMI and self-monitoring on the main outcome changes [43].

Methods

Ethical Considerations

All study procedures were approved by the National University of Singapore (NUS) before participant recruitment and all participants provided informed consent (institutional review board #S-20-025). This RCT was preregistered on Open Science Framework [44]. All data were de-identified. Participants were reimbursed up to $30, 8 subject pool hour credits, or both, pro-rated based on their degree of participation.

Study Design

In this assessor-blinded RCT, we randomized individuals into 1 of 2 arms with a parallel design and 1:1 allocation ratio. Randomization was stratified according to age and sex. We used a mixed design of 2 groups (group: MEMI vs self-monitoring) by 3 (prerandomization, postintervention, and 1MFU postrandomization) time points to evaluate the efficacy of the 14-day MEMI compared to self-monitoring on ER difficulties and self-compassion domain outcomes. Random treatment assignment to the MEMI and self-monitoring arms was the between-participant factor, whereas time points served as the within-participant factor. The trial was advertised as a "digital mindfulness intervention study" by emailing NUS students via a listserve, posting advertisements across the campus and NUS-affiliated mental health clinics, and permitting recruitment from both the student body and the broader community.

Eligibility Criteria

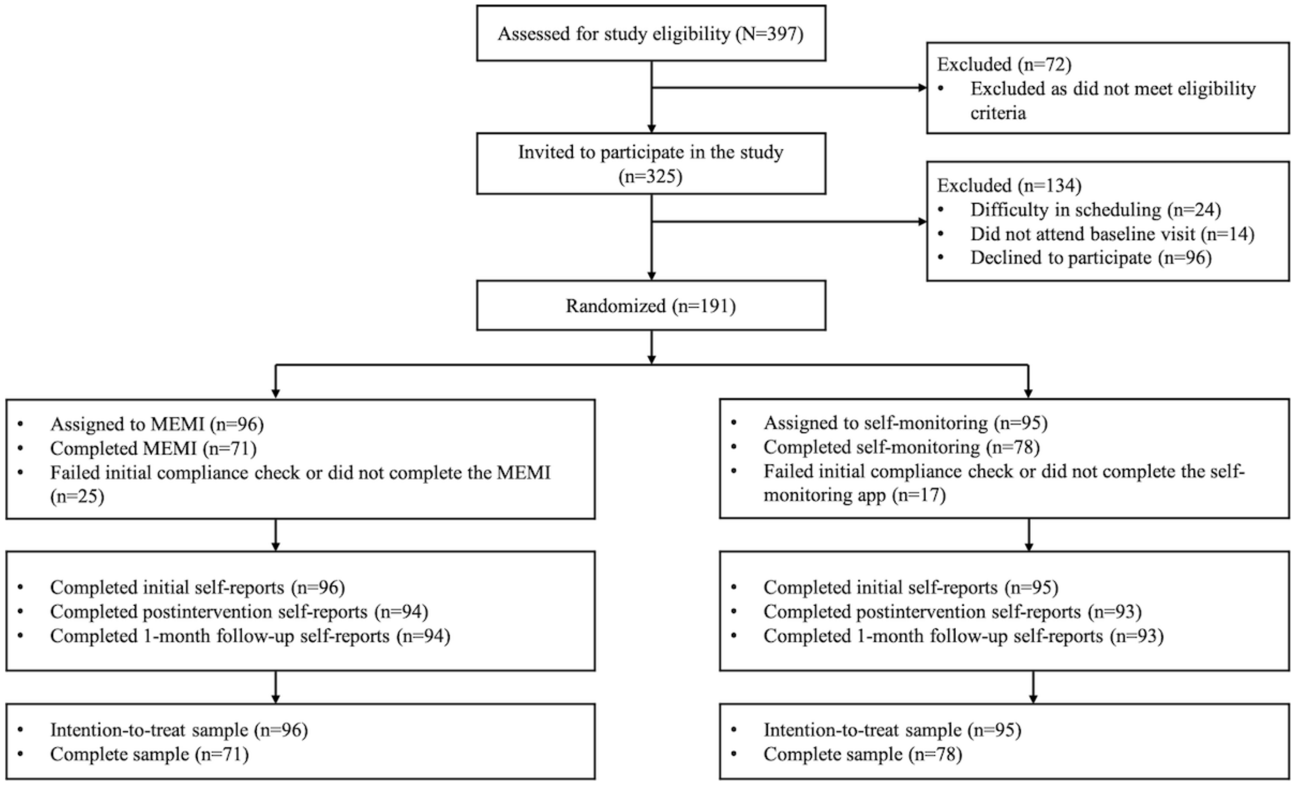
Details of the study methods can be found in Multimedia Appendix 1 and in an earlier report by Zainal et al [43]. Eligible participants were required to self-report SAD with a Social Phobia Inventory (SPIN) score ≥20 [45], be aged at least 18 years, own a smartphone, and seek help for mental health issues. We recruited treatment-seeking individuals from the psychology participant pool and the local community, excluding those with self-reported suicidal ideation, mania, or psychosis. Eligible participants were recruited to this web-based trial on campus (before the COVID-19 pandemic) and on the web (during the pandemic) between September 1, 2019, and May 31, 2021.

Several reasons prompted us to choose the SPIN measure to screen for participants with probable SAD. Unlike the Social Phobia Diagnostic Questionnaire (SPDQ), the SPIN was already integrated as part of a larger battery of screening assessments in a busy psychological clinic and an undergraduate psychology research participant pool at NUS. This clinical assessment battery was based on a series of *Diagnostic and Statistical Manual (DSM), Fifth Edition*–text revised web-based assessment measures made accessible and recommended by the American Psychiatric Association [46]. Although the SPDQ has superior psychometric properties to diagnose SAD, given its structural concordance with the *DSM*, the briefer SPIN was the more pragmatic and operationally efficient choice for this study. Moreover, both the SPIN and SPDQ baseline scores were highly correlated (r=0.89; P<.001). Furthermore, all eligible participants met criteria for probable SAD at baseline using the recommended SPDQ cut-off score of ≥7.38 that had optimal sensitivity (82%), specificity (85%), a positive predictive value (83%), and a negative predictive value (83%) with a clinical diagnosis of SAD with the Anxiety Disorder Interview Schedule-IV [47,48].

Participants

All participants provided voluntary informed consent. We randomized 191 participants into 2 groups: MEMI (96/191, 50.3%) and self-monitoring (95/191, 49.7%). Their average age was 21.84 (SD 3.37; range 18-53) years. Of the 191 participants, 41 (21.5%) identified as male, 149 (78%) as female, and 1 (1%) as other; 165 (86.4%) identified as Chinese and the remaining 26 (13.6%) identified as Indian, Malay, and other ethnicities; 167 (87.4%) were categorized as never married, whereas 24 (12.6%) were married, living with a partner, or in an intimate relationship but not living together; 145 (75.9%) had completed junior college as their highest level of education, whereas 46 (24.1%) held diplomas, university degrees, or graduate degrees; 139 (72.8%) were not employed, whereas 52 (27.2%) were engaged in part-time or full-time work; 178 (93.2%) were full-time students, whereas the remaining 13 (6.8%) were part-time students or nonstudents; 172 (90%) reported an annual income within the range of US $0 to US $7500, whereas 19 (9.9%) fell into higher income brackets; 11 (5.8%) had previously received a clinical diagnosis of anxiety or depressive disorder; and 10 (5.2%) were currently using psychotropic drugs (Table S1 in Multimedia Appendix 1). Eligible participants had a mean SPIN score of 35.65 (SD 13.18; range 20-67), with 31.9% (61/191) in the mild severity (score of 20-30), 31.9% (61/191) in the moderate severity (score of 31-40), 21.9% (42/191) in the severe (score of 41-50), and 14.1% (27/191) in the very severe (score of ≥51) categories [45,49,50]. Figure 1 presents the CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) [51] diagram (Multimedia Appendix 2).

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) diagram. MEMI: mindfulness ecological momentary intervention. SM: self-monitoring.



Procedures

The MEMI and self-monitoring arms were codeveloped by the lead author (NHZ) and senior author (MGN; both PhD-level psychologists). The study procedures were tested with various research assistants who helped provide feedback and troubleshoot technical issues and were refined from August 1, 2018, to May 31, 2019. Weekly meetings were held during this period to optimize the study procedures and reach a consensus regarding ambiguous procedural aspects. Therefore, the team did not encounter technical glitches with the Personal Analytics Companion app [52] used to deliver the MEMI and self-monitoring app during the subsequent data collection phase.

At baseline, eligible participants completed a counterbalanced series of web-based self-report measures. Subsequently, we randomly assigned participants to either the MEMI or self-monitoring arm using the Excel (Microsoft Corp) randomization function integrated into Qualtrics (Qualtrics International Inc). The mindfulness or self-monitoring video was provided toward the end of the baseline visit after the completion of all pretrial assessments. Then, participants installed the Personal Analytics Companion app [52], which contained either the MEMI or self-monitoring tool, on their smartphones, with the experimenter demonstrating its features. Assessors (also called experimenters herein) were blinded to group assignment. Participants were told that they would receive an intervention and experimenters could not know if they were randomly assigned to the MEMI or self-monitoring arm. The owner of the Personal Analytics Companion software had no role in the development of the MEMI and self-monitoring app. Participants were informed that they would receive prompts at 5 different times each day (around 9 AM, noon, 3 PM, 6 PM, and 9 PM) during the following 14-day period. These prompts could be edited within an approximately 2-hour window (ie, 8 AM-10 AM, 11 AM-1 PM, 2 PM-4 PM, 5 PM-7 PM, and 8 PM-10 PM) to fit each participant's schedule and did not substantially differ across participants. To ensure the validity of responses, participants were instructed to provide input on their current state of depression, anxiety, and mindfulness within 2 hours of receiving the MEMI or self-monitoring prompt. These prompts guided the participants in 1-minute mindfulness or self-monitoring activities based on their assigned group. The apps did not require the training of Bachelor of Arts—level coaches and were entirely self-guided. Following the 14-day treatment phase, participants received emails prompting them to complete the self-reported clinical outcome measures at postintervention and 1MFU time points.

RCT Arms of MEMI and Self-Monitoring

MEMI App

Participants assigned to the MEMI app viewed a standardized video presentation led by the principal investigator that described the evidence-based MBI protocol akin to mindfulness-based stress reduction [18,53]. This presentation familiarized MEMI participants with mindfulness, encouraging them to fully immerse themselves in their present moment and equipping them with skills encompassing open monitoring and attentiveness to present transient experiences. Proficiency in present-moment awareness may increase the richness of experiences, thereby amplifying emotional responsiveness [14]. Next, the video therapist skillfully demonstrated the paced, rhythmic diaphragmatic breathing technique and guided participants in its practice. Subsequently, the video therapist continued by imparting lessons on nonjudgmental acceptance (ie, allowing emotions to fluctuate and experiences to unfold without deliberately changing them), drawing from the mindfulness-based cognitive therapy principles [19]. These lessons fostered mindfulness-related attributes such as observation, nonreactivity, and nonjudgmental acceptance. Subsequently, each MEMI participant received a comprehensive explanation regarding the importance and benefits of consistent mindfulness practice.

The MEMI app prompted participants to practice mindfulness 5 times daily (at approximately 9 AM, noon, 3 PM, 6 PM, and 9 PM) for 14 days. During each prompt, MEMI participants received the following instructions:

> *Pay attention to your breathing. Breathe in a slow, steady, and rhythmic manner. Stay focused on the sensations of the air coming into your lungs and then letting it out. As you're breathing, observe your experience as it is. Let go of judgments that do not serve you. Focus on the here and now. Attend to the small moments right now (eg, reading a chapter, having a cool glass of water), as that is where enjoyment, peace, and serenity in life happen.*

Participants rated their state-level (ie, momentary) depression ("To what degree do you feel depressed right now?"), anxiety ("To what degree do you feel keyed up or on edge right now?"), and mindfulness ("To what extent are you experiencing the present moment fully?") levels on a 9-point Likert scale (1=*not at all* to 9=*extremely*) before and after receiving these instructions. Each MEMI prompt ended with encouragement to inculcate these skills in the long term as follows:

> *Remember that the cultivation of mindfulness is lifelong. The goal of therapy is to be your own therapist. Practice mindfulness between the prompts and after you have completed this study.*

These MEMI techniques were proposed to work for SAD in the following ways. First, *focused attention* to present-moment activities would break the habit of ruminating on the past unproductively, such as brooding over social events in self-critical ways often observed in SAD. Furthermore, this activity was designed to reduce pathological worry about the future, which is a common comorbid symptom of SAD. Second, *open monitoring* skills helped participants flexibly experience positive and negative emotions without resistance and cultivate stronger discomfort tolerance. Third, *mindful diaphragmatic breathing* might work therapeutically by inducing relaxation, improving heart rate variability, and lowering blood pressure [54]. Fourth, *acceptance* exercises might enhance the capacity to tolerate and manage emotional states, specifically addressing affect intolerance or sensitivity, an etiological and maintenance factor of SAD, and other psychopathology [55]. Unlike cognitive behavioral therapy (CBT) apps, the MEMI app does not work by continually instructing exposure therapy designed to encourage immersion in fear-inducing social situations to help persons with SAD cope with their anxiety and enhance self-efficacy in these contexts. No instructions were given to persuade participants to create a list of feared and avoided situations and to gradually and consistently approach and engage with these situations. Collectively, the MEMI app focused primarily on teaching mindfulness principles and skills.

The experimenter was available to address questions and then administered the 6-item Credibility and Expectancy Questionnaire [56]. After participants grasped the rationale and techniques of mindfulness, they set up the MEMI app on their smartphones. Furthermore, participants were provided with a copy of the MEMI rationale handout and were encouraged to engage with it consistently (Multimedia Appendix 1).

Self-Monitoring App

The video presentation created for self-monitoring participants commenced with the principal investigator (NHZ) explaining self-monitoring as an elevated awareness of one's thoughts and emotions, with particular attention to one's uncomfortable experiences. Following this, the video introduced the idea that monitoring thoughts and tracking associated distress could lead to healthier thinking. In essence, the self-monitoring video conveyed that self-monitoring (ie, focusing solely on distress) had the potential to alleviate anxiety. Self-monitoring was used in another set of studies that inspired the conceptual foundation of our self-monitoring app [57,58]. This adaptation sought to emulate the MEMI structure while intentionally excluding its presumed therapeutic elements, including acceptance, diaphragmatic breathing training, focusing on transient present moments, open monitoring, and a regimen of consistent mindfulness practices. Notably, it refrained from introducing the concept of mindfulness and abstained from guiding participants toward engaging with the present moment in ways that could influence their mood. In contrast to MEMI, the self-monitoring app encouraged participants to simply observe their thoughts and emotions and did not include emphasis on the need to accept these thoughts and feelings as they emerged. Furthermore, it did not include instructions for breathing retraining or the intention to elicit relaxation through abdominal breathing. Unlike MEMI, which promoted ongoing mindfulness practice, self-monitoring participants were not prompted to engage in self-monitoring between prompts and after the intervention. This self-monitoring approach was strategically designed to address potential credibility and expectancy effects, minimize the likelihood of regression to the mean, and mitigate the possibility of inflated effect sizes that could occur with a no-treatment or waitlist control group [59] (Multimedia Appendix 1).

Instead of receiving lengthier messages to practice mindfulness continually as with the MEMI group, self-monitoring participants received the following brief instruction 5 times daily (at about 9 AM, noon, 3 PM, 6 PM, and 9 PM) for 14 days: "Notice your thoughts and how distressing they may be." State-level depression, anxiety, and mindfulness were measured using the same 9-point Likert scale items before and after each self-monitoring prompt. Subsequently, similar to the MEMI group, experimenters administered the 6-item Credibility and Expectancy Questionnaire after confirming self-monitoring participants' comprehension of the rationale and self-monitoring technique. Participants received compensation through either course extra credit hours or monetary rewards.

The self-monitoring app was designed to work in the following manner. By suggesting that solely self-monitoring and focusing on distressing thought patterns might remedy anxiety, the app controlled for treatment expectancy and credibility effects across arms. Furthermore, by eliminating active therapeutic ingredients in the MEMI app (eg, focused attention on the present moment, open monitoring, diaphragmatic breathing, and nonjudgmental acceptance), the self-monitoring app functioned as a placebo comparator to maximize the odds of attaining between-arm efficacy on clinical outcomes.

Pre- and Midintervention Measures

Trait ER

The Difficulties in Emotion Regulation Scale (DERS) [41] is a 36-item web-based self-report tool that evaluates difficulties in regulating emotions. Participants responded on a 5-point scale (1=*almost never, 0%-10%* to 5=*almost always, 91%-100%*). The scale provides a total score (range 36-180), where higher scores reflect more ER difficulties. In addition, 6 ER difficulty facets were evaluated via the following DERS subscales: *nonacceptance of emotions* (acceptance), *emotional awareness issues* (awareness), *lack of emotional clarity* (clarity), *goal-directed behavior difficulties* (goals), *impulse control problems* (impulsivity), and *strategy use problems* (strategy and difficulty accessing ER skills) [41]. The DERS has shown strong internal consistency, excellent 2-month retest reliability [41], and good convergent and discriminant validity [60]. The internal consistency (Cronbach α values) was excellent at .95 at both prerandomization and 7-day midintervention time points.

Trait Self-Compassion

Participants rated on a 5-point Likert scale (1=*almost never* to 5=*almost always*) the extent to which they identified with each statement on the Self-Compassion Scale (SCS) [8]. The SCS involves six distinct aspects assessed via the following subscales: (1) *acknowledging shared human struggles* (common humanity), (2) *social connectedness* (feeling a sense of being linked with other humans), (3) *mindfulness* (nonjudgmental awareness), (4) *nonidentification with emotions* (the absence of intense focus on negative emotions), (5) *nonjudgment toward oneself* (self-soothing tendencies during times of distress), and (6) *self-kindness* (showing warmth toward one's imperfections) [8]. The SCS has shown strong internal consistency, predictive validity, convergent validity, discriminant validity, and retest reliability [61-63]. The Cronbach α values were good at 0.95 and 0.96 at prerandomization and 7-day midintervention time points, respectively.

SAD Symptoms

The 25-item SPDQ [48] assessed SAD fear and avoidance symptoms in different social situations, aligned with the *DSM Fourth Edition* criteria [64]. It has shown good retest reliability and strong internal consistency (Cronbach α=.96, .97, and .98 at prerandomization, postintervention, and 1MFU time points, respectively) [48]. Furthermore, it displayed robust convergent and discriminant validity, with high sensitivity (82%), specificity (85%) [48], and responsiveness to symptom reduction changes in clinical trials [13].

The 17-item SPIN [45] evaluated the SAD fear and avoidance symptoms in the past week, such as the fear of social embarrassment. Participants used a 5-point Likert scale (1=*not at all* to 5=*extremely*) to rate the relevance of each statement to their past week's experiences. Prior research has shown that the SPIN exhibited acceptable convergent validity with other established assessments of SAD [50,65]. The SPIN showed strong internal consistency in this study (Cronbach α=.93, .94, and .95 at prerandomization, postintervention, and 1MFU time points, respectively). Finally, a cut-off SPIN score ≥20 yielded excellent psychometric properties, with a sensitivity of 0.85, specificity of 0.86, positive predictive value of 0.85, negative predictive value of 0.85, and correct classification rate of 85% [50], when compared to the structured interview for the *DSM Fourth Edition* [66,67]. Excellent psychometric properties with a similar cut-off have been replicated in another Asian sample [68].

Generalized Anxiety Symptoms

The Generalized Anxiety Disorder Questionnaire–fourth edition (GADQ-IV) [69], comprising 14 items, evaluated the symptoms of generalized anxiety disorder (GAD) through a combination of dichotomous ("yes" or "no") and continuous responses. The continuous responses included a 9-point Likert scale for assessing the interference and distress caused by GAD symptoms. The GADQ-IV showed strong internal consistency (Cronbach α=0.93, 0.93, and 0.94 at prerandomization, postintervention, and 1MFU time points, respectively) and high retest reliability [69]. Furthermore, it exhibited strong convergent and discriminant validity, and aligned well with the structured diagnostic assessments of GAD [70,71].

Depression Severity

The Beck Depression Inventory–second edition (BDI-II) [72], which is a 21-item scale, assesses depression symptom severity. Participants rated the severity of each symptom based on their experiences in the past 2 weeks, using a scale of 0 to 3 to indicate increasing severity. The BDI-II showed excellent internal consistency (Cronbach α=.93, .95, .95 at prerandomization, postintervention, and 1MFU time points, respectively), high retest reliability, and strong convergent and discriminant validity [73,74].

Trait Repetitive Negative Thinking

The 45-item Perseverative Cognitions Questionnaire (PCQ; PCQ-45) [75] assessed persistent, repetitive negative thinking tendencies related to worrisome, obsessive, and ruminative thoughts. Participants rated the items on a 6-point Likert scale (0=*strongly disagree* to 5=*strongly agree*). The PCQ-45 included 6 domains: anticipating the worst, brooding over the past, future preparation, thoughts conflicting with the ideal self, uncontrollability, and seeking causes and meanings [75]. We calculated the total PCQ score by computing the average of all subscale scores. The scale showed high 2-week retest reliability and strong discriminant and convergent validity (Cronbach α=.96, .97, and .98 at prerandomization, postintervention, and 1MFU time points, respectively), and it demonstrated cross-cultural measurement invariance between the United States and Singapore [76].

Trait Mindfulness

The 39-item Five-Factor Mindfulness Questionnaire (FFMQ) [77] measured participants' inclination to practice mindfulness across 5 domains: awareness of the consequences of actions, description, nonjudgment, nonreactivity to inner experiences, and observation. Participants rated these aspects on a 5-point Likert scale (1=*never or very rarely true* to 5=*very often or always true*). The FFMQ total score showed strong convergent validity [78], differentiation from measures of unrelated factors (eg, psychological well-being) [77], and retest reliability [79]. The Cronbach α values were .90, .91, and .93 at prerandomization, postintervention, and 1MFU time points, respectively.

Statistical Analysis

Power Analysis and Missing Data Management

On the basis of a Monte Carlo power analysis [80,81], this study had 100% power to detect a substantial group × time interaction with a small effect size of Cohen *d*=0.2. We performed intent-to-treat [82] analyses with the enrolled 191 participants by incorporating data from participants who did not meet the 7-day compliance check (completing ≥80% of the app prompts in the 2-week intervention phase). Participants demonstrated strong adherence, with 78% (149/191) responding to at least 80% (56/70) of the MEMI and self-monitoring prompts. A total of 2.03% of the data from the pre-1MFU time points were missing. To address this, we applied multiple imputation using the recommended predictive mean matching algorithm by pooling data from 100 imputed data sets, each with 10 iterations [83].

Hierarchical Linear Modeling

To test hypothesis 1, we used hierarchical linear modeling [84], also known as multilevel modeling, to account for data nonindependence resulting from the nesting of repeated observations (level 1) within participants (level 2). For each hierarchical linear modeling, group (intervention), time, and group × time interaction were fixed-effect predictors of improvement in the prerandomization to midintervention ER and self-compassion domains, and the intercept (ie, time-coded as 0 for preintervention and 1 for midintervention time points) was the single random-effect predictor (permitting participants to vary in their average-outcome values). We used fitted models to determine the estimated mean scores at each time point.

Parallel Structural Equation Modeling Mediation Analyses

To test hypothesis 2, we conducted parallel mediation analyses using the *lavaan* structural equation modeling R package (R Foundation for Statistical Computing) [85] to assess whether pretrial to midintervention ER and self-compassion domains mediated group effects on the 6 clinical outcomes from pre-post and pre-1MFU time points. Similar to the primary RCT study, the clinical outcomes examined were the alleviation of pre-post and pre-1MFU SAD symptoms, generalized anxiety symptoms, depression severity, repetitive negative thinking, and trait mindfulness. Although there were no significant between-group effects from pre-post and pre-1MFU time points [43], we conducted mediation tests because significant indirect effects can exist without significant between-group total or direct effects [86]. Furthermore, our primary efficacy paper still showed significant within-group effects on these 5 clinical outcomes from pre-post and pre-1MFU time points [43]. Moreover, as recommended, our mediation tests aligned with theoretical considerations [87] and statistical recommendations to measure mediators at the midpoint of the intervention. In order to establish them as potential mechanisms of change, mediators should temporally precede the outcome measures [88]. The associations between groups, pretrial to midintervention mediators, and pre-post or pre-1MFU outcomes can be described using 4 regression coefficients (or paths) [89]: the group effect on the pretrial to midintervention mediator (path *a*), the effect of the pretrial to midintervention mediator on pre-post or pre-1MFU outcome (path *b*), the total effect (path *c*; the combination of the *a* and *b* paths), and the direct effect (path *c*'; the group effect on the outcome, irrespective of the mediator). The product of paths *a* and *b* signifies the mediated (or indirect) effect, that is, the focal estimate when examining mediation. We assessed the significance of the mediated effect via a nonparametric bootstrap method, generating 2000 random samples to determine point estimates, SEs, and bias-corrected bootstrap 95% CIs for each indirect effect. The mediated effect was considered significant if the CI bounds (both upper and lower) did not include 0. To safeguard against type I errors, we conducted post hoc comparisons for statistically significant effects using the Simes Bonferroni correction method [90]. Cohen *d* effect sizes and their 95% CIs were calculated to ease the interpretation of parameter estimates for both study hypotheses, such that Cohen *d* values of 0.2, 0.5, and 0.8 denote small, moderate, and large effects, respectively [91].

Results

Table S2 in Multimedia Appendix 1 displays descriptive statistics of the trait ER and self-compassion domain scores for participants in the MEMI and self-monitoring groups.

Between- and Within-Group Effects of Brief MEMI Versus Self-Monitoring on ER Domains

Significant between-group effects occurred from pretrial to midintervention time points on DERS domains, including lack of emotional clarity (Cohen *d*=0.16, 95% CI 0.02-0.31; *P*=.03) and difficulties engaging in goal-directed behavior (Cohen *d*=−0.24, 95% CI −0.39 to −0.10; *P*=.001; Tables 1 and 2). Regarding within-group effects, reductions in these difficulties were significantly stronger for the MEMI than self-monitoring groups from pretrial to midintervention time points (lack of emotional clarity: Cohen *d*=−0.39, 95% CI −0.54 to −0.25; *P*<.001 vs Cohen *d*=−0.21, 95% CI −0.35 to −0.06; *P*=.004 and goal-directed behavior difficulties: Cohen *d*=−0.73, 95% CI −0.88 to −0.58; *P*<.001 vs Cohen *d*=−0.29, 95% CI −0.44 to −0.15; *P*<.001). No significant between-group effects emerged from pretrial to midintervention time points on emotional awareness issues (Cohen *d*=0.11, 95% CI −0.03 to 0.26; *P*=.12), impulse control issues (Cohen *d*=0.11, 95% CI −0.03 to 0.26; *P*=.12), nonacceptance of emotions (Cohen *d*=−0.07, 95% CI −0.21 to 0.08; *P*=.36), and strategy use problems (Cohen *d*=−0.12, 95% CI −0.27 to 0.02; *P*=.10). Regarding within-group effects, both the MEMI and self-monitoring app significantly reduced impulse control issues (MEMI: Cohen *d*=−0.32, 95% CI −0.47 to −0.18; *P*=.006 and self-monitoring: Cohen *d*=−0.26, 95% CI −0.40 to −0.11; *P*<.001), nonacceptance of emotions (MEMI: Cohen *d*=−0.32, 95% CI −0.47 to −0.18; *P*<.001; self-monitoring: Cohen *d*=−0.24, 95% CI −0.38 to −0.09; *P*=.001), and strategy use problems (MEMI: Cohen *d*=−0.40, 95% CI −0.54 to −0.25; *P*<.001; self-monitoring: Cohen *d*=−0.21, 95% CI −0.35 to −0.06; *P*=.004) from pretrial to midintervention time points. A significant reduction in emotional awareness issues from pretrial to midintervention time points was observed in the self-monitoring group (Cohen *d*=−0.73, 95% CI −0.88 to −0.58; *P*<.001) but not in the MEMI group (Cohen *d*=−0.29, 95% CI −0.44 to 0.02; *P*=.07).

Table 1. Between-intervention hierarchical linear modeling of the mindfulness ecological momentary intervention versus self-monitoring app predicting emotion regulation and self-compassion domains.

Domains	β	*t* test (*df*)	*P* value	Cohen *d* (95% CI)
DERS[a] domains				
DERS emotional awareness issues				
Intercept	16.66	39.37 (953)	<.001	2.85 (2.64 to 3.05)
Group	−.58	−0.97 (189)	.34	−0.07 (−0.21 to 0.07)
Prerandomization to midintervention time points	−.91	−3.97 (953)	<.001	−0.29 (−0.43 to −0.14)
Group × pretrial to midintervention time points	.50	1.54 (953)	.12	0.11 (−0.03 to 0.26)
DERS lack of emotional clarity				
Intercept	13.02	38.69 (953)	<.001	2.80 (2.60 to 3.00)
Group	−0.19	−0.41 (189)	.68	−0.03 (−0.17 to 0.12)
Prerandomization to midintervention time points	−1.31	−7.33 (953)	<.001	−0.53 (−0.68 to −0.38)
Group × pretrial to midintervention time points	.56	2.21 (953)	.027	0.16 (0.02 to 0.31)
DERS goal-directed behavior difficulties				
Intercept	17.22	43.55 (953)	<.001	3.15 (2.93 to 3.37)
Group	.40	0.71 (189)	.48	0.05 (−0.09 to 0.20)
Prerandomization to midintervention time points	−1.02	−4.40 (953)	<.001	−0.32 (−0.46 to −0.17)
Group × pretrial to midintervention time points	−1.11	−3.39 (953)	.001	−0.24 (−0.39 to −0.10)
DERS impulse control issues				
Intercept	16.66	31.84 (953)	<.001	2.85 (2.64 to 3.05)
Group	−.58	0.12 (189)	.34	−0.07 (−0.21 to 0.07)
Prerandomization to midintervention time points	−.91	−3.30 (953)	<.001	−0.29 (−0.43 to −0.14)
Group × pretrial to midintervention time points	.50	0.29 (953)	.12	0.11 (−0.03 to 0.26)
DERS nonacceptance of emotions				
Intercept	15.74	32.08 (953)	<.001	2.32 (2.13 to 2.51)
Group	−.30	−0.43 (189)	.67	−0.03 (−0.18 to 0.11)
Prerandomization to midintervention time points	−1.01	−3.20 (953)	.001	−0.23 (−0.38 to −0.09)
Group × pretrial to midintervention time points	−.41	−0.92 (953)	.36	−0.07 (−0.21 to 0.08)
DERS strategy use problems				
Intercept	21.23	35.21 (953)	<.001	2.55 (2.35 to 2.74)
Group	−.36	−0.42 (189)	.68	−0.03 (−0.17 to 0.11)
Prerandomization to midintervention time points	−.93	−2.94 (953)	.003	−0.21 (−0.36 to −0.07)
Group × pretrial to midintervention time points	−.75	−1.67 (953)	.10	−0.12 (−0.27 to 0.02)
SCS[b] domains				
SCS acknowledging human struggles				
Intercept	12.20	34.92 (953)	<.001	2.53 (2.33 to 2.72)
Group	−.24	−0.49 (189)	.63	−0.04 (−0.18 to 0.11)
Prerandomization to midintervention time points	−.61	−3.48 (953)	.001	−0.25 (−0.40 to −0.11)
Group × pretrial to midintervention time points	.88	3.56 (953)	<.001	0.26 (0.11 to 0.40)
SCS social connectedness				
Intercept	10.25	31.05 (953)	<.001	2.25 (2.06 to 2.43)
Group	.15	0.31 (189)	.75	0.02 (−0.12 to 0.17)
Prerandomization to midintervention time points	.41	2.39 (953)	.02	0.17 (0.03 to 0.32)

Domains	β	*t* test (*df*)	*P* value	Cohen *d* (95% CI)
Group × pretrial to midintervention time points	.73	3.00 (953)	.003	0.22 (0.07 to 0.36)
SCS mindfulness				
Intercept	12.45	44.87 (953)	<.001	3.25 (3.03 to 3.47)
Group	−.21	−0.53 (189)	.60	−0.04 (−0.18 to 0.11)
Prerandomization to midintervention time points	.04	0.29 (953)	.78	0.02 (−0.12 to 0.17)
Group × pretrial to midintervention time points	.01	0.04 (953)	.97	0.00 (−0.14 to 0.15)
SCS nonidentification with emotions				
Intercept	10.05	34.19 (953)	<.001	2.47 (2.28 to 2.67)
Group	−0.25	−0.60 (189)	.55	−0.04 (−0.19 to 0.10)
Prerandomization to midintervention time points	1.02	6.18 (953)	<.001	0.45 (0.30 to 0.59)
Group × pretrial to midintervention time points	.50	2.15 (953)	.03	0.16 (0.01 to 0.30)
SCS nonjudgment toward oneself				
Intercept	12.74	33.37 (953)	<.001	2.41 (2.22 to 2.60)
Group	.18	0.33 (189)	.74	0.02 (−0.12 to 0.17)
Prerandomization to midintervention time points	1.29	6.66 (953)	<.001	0.48 (0.34 to 0.63)
Group × pretrial to midintervention time points	.46	1.66 (953)	.10	0.12 (−0.02 to 0.27)
SCS self-kindness				
Intercept	14.12	37.95 (953)	<.001	2.75 (2.54 to 2.95)
Group	.28	0.53 (189)	.60	0.04 (−0.11 to 0.18)
Prerandomization to midintervention time points	.24	1.33 (953)	.18	0.10 (−0.05 to 0.24)
Group × pretrial to midintervention time points	.67	2.61 (953)	.009	0.19 (0.04 to 0.33)

[a]DERS: Difficulties in Emotion Regulation Scale.

[b]SCS: Self-Compassion Scale.

Table 2. Within-intervention hierarchical linear modeling of MEMI[a] and self-monitoring app predicting emotion regulation and self-compassion domains.

Domains	β	t test (df)	P value	Cohen d (95% CI)
DERS[b] domains				
DERS emotional awareness issues				
Intercept (MEMI)	16.08	36.33 (479)	<.001	2.64 (2.45 to 2.84)
Time (MEMI)	−.41	–1.79 (479)	.07	−0.13 (−0.27 to 0.02)
Intercept (self-monitoring)	16.66	41.46 (474)	<.001	3.01 (2.80 to 3.23)
Time (self-monitoring)	−.91	–4.00 (474)	<.001	−0.29 (−0.43 to −0.14)
DERS lack of emotional clarity				
Intercept (MEMI)	20.87	35.74 (479)	<.001	2.60 (2.41 to 2.80)
Time (MEMI)	−1.68	−5.47 (479)	<.001	−0.39 (−0.54 to −0.25)
Intercept (self-monitoring)	21.23	34.31 (474)	<.001	2.48 (2.29 to 2.68)
Time (self-monitoring)	−.93	−2.87 (474)	.004	−0.21 (−0.35 to −0.06)
DERS goal-directed behavior difficulties				
Intercept (MEMI)	17.61	46.44 (479)	<.001	3.36 (3.14 to 3.59)
Time (MEMI)	−2.13	−10.11 (479)	<.001	−0.73 (−0.88 to −0.58)
Intercept (self-monitoring)	17.22	42.07 (474)	<.001	3.04 (2.83 to 3.26)
Time (self-monitoring)	−1.02	−4.07 (474)	<.001	−0.29 (−0.44 to −0.15)
DERS impulse control issues				
Intercept (MEMI)	14.94	32.73 (479)	<.001	2.28 (2.09 to 2.46)
Time (MEMI)	−.67	−2.73 (479)	.006	−0.32 (−0.47 to −0.18)
Intercept (self-monitoring)	14.87	31.26 (474)	<.001	2.26 (2.08 to 2.45)
Time (self-monitoring)	−.76	−3.56 (474)	<.001	−0.26 (−0.40 to −0.11)
DERS nonacceptance of emotions				
Intercept (MEMI)	15.44	31.45 (479)	<.001	2.28 (2.09 to 2.46)
Time (MEMI)	−1.42	−4.45 (479)	<.001	−0.32 (−0.47 to −0.18)
Intercept (self-monitoring)	15.74	32.26 (474)	<.001	2.33 (2.15 to 2.52)
Time (self-monitoring)	−1.01	−3.26 (474)	.001	−0.24 (−0.38 to −0.09)
DERS strategy use problems				
Intercept (MEMI)	20.87	35.74 (479)	<.001	2.59 (2.39 to 2.78)
Time (MEMI)	−1.68	−5.47 (479)	<.001	−0.40 (−0.54 to −0.25)
Intercept (self-monitoring)	21.23	34.31 (474)	<.001	2.48 (2.29 to 2.68)
Time (self-monitoring)	−.93	−2.87 (474)	.004	−0.21 (−0.35 to −0.06)
SCS[c] domains				
SCS acknowledging human struggles				
Intercept (MEMI)	11.96	33.83 (479)	<.001	2.45 (2.26 to 2.64)
Time (MEMI)	.27	1.74 (479)	.08	0.13 (−0.02 to 0.27)
Intercept (self-monitoring)	12.20	35.52 (474)	<.001	2.57 (2.37 to 2.77)
Time (self-monitoring)	−.61	−3.18 (474)	.002	−0.23 (−0.38 to −0.08)
SCS social connectedness				
Intercept (MEMI)	10.40	29.86 (479)	<.001	2.16 (1.98 to 2.34)
Time (MEMI)	1.14	6.20 (479)	<.001	0.45 (0.30 to 0.60)
Intercept (self-monitoring)	10.25	33.21 (474)	<.001	2.40 (2.21 to 2.59)

Domains	β	t test (df)	P value	Cohen d (95% CI)
Time (self-monitoring)	.41	2.59 (474)	.01	0.19 (0.04 to 0.33)
SCS mindfulness				
Intercept (MEMI)	12.25	42.63 (479)	<.001	3.08 (2.87 to 3.30)
Time (MEMI)	.05	0.35 (479)	.73	0.03 (−0.12 to 0.17)
Intercept (self-monitoring)	12.45	46.91 (474)	<.001	3.39 (3.17 to 3.62)
Time (self-monitoring)	.04	0.29 (474)	.77	0.02 (−0.12 to 0.17)
SCS nonidentification with emotions				
Intercept (MEMI)	9.80	32.35 (479)	<.001	2.34 (2.15 to 2.53)
Time (MEMI)	1.52	8.74 (479)	<.001	0.63 (0.48 to 0.78)
Intercept (self-monitoring)	10.05	35.53 (474)	<.001	2.57 (2.38 to 2.77)
Time (self-monitoring)	1.02	6.60 (474)	<.001	0.48 (0.33 to 0.62)
SCS self-judgment				
Intercept (MEMI)	12.92	34.08 (479)	<.001	2.47 (2.27 to 2.66)
Time (MEMI)	1.75	8.39 (479)	<.001	0.61 (0.46 to 0.76)
Intercept (self-monitoring)	12.74	33.28 (474)	<.001	2.41 (2.22 to 2.6)
Time (self-monitoring)	1.29	7.30 (474)	<.001	0.53 (0.38 to 0.68)
SCS self-kindness				
Intercept (MEMI)	14.40	35.81 (479)	<.001	2.59 (2.39 to 2.79)
Time (MEMI)	.91	4.91 (479)	<.001	0.36 (0.21 to 0.50)
Intercept (self-monitoring)	14.12	41.90 (474)	<.001	3.03 (2.82 to 3.24)
Time (self-monitoring)	.24	1.37 (474)	.17	0.10 (−0.05 to 0.24)

[a]MEMI: mindfulness ecological momentary intervention.

[b]DERS: Difficulties in Emotion Regulation Scale.

[c]SCS: Self-Compassion Scale.

Between- and Within-Group Effects of Brief MEMI Versus Self-Monitoring on Self-Compassion Domains

Significant between-group effects occurred from pretrial to midintervention time points on SCS domains, including acknowledging shared human struggles (Cohen *d*=0.26, 95% CI 0.11-0.40; *P*<.001), social connectedness (Cohen *d*=0.19, 95% CI 0.04-0.33; *P*=.01), nonidentification with emotions (Cohen *d*=0.16, 95% CI 0.01-0.30; *P*=.03), and self-kindness (Cohen *d*=0.19, 95% CI 0.04-0.33; *P*=.009; Tables S3 and S4 in Multimedia Appendix 1). Regarding within-group effects, increases in these self-compassion domains were significantly larger for the MEMI than self-monitoring groups from pretrial to midintervention time points (acknowledging shared human struggles: Cohen *d*=0.13, 95% CI −0.02 to 0.27; *P*=.08 vs Cohen *d*=−0.23, 95% CI −0.38 to −0.08; *P*=.002); social connectedness: Cohen *d*=0.45, 95% CI 0.30-0.60; *P*<.001 vs Cohen *d*=0.19, 95% CI 0.04-0.33; *P*=.01; identification with emotions: Cohen *d*=0.63, 95% CI 0.48-0.78; *P*<.001 vs Cohen *d*=0.48, 95% CI 0.33-0.62; *P*<.001; and self-kindness: Cohen *d*=0.36, 95% CI 0.21-0.50; *P*<.001 vs Cohen *d*=0.10, 95% CI −0.05 to 0.24; *P*=.17). No significant between-group effects emerged from pretrial to midintervention time points on mindfulness (Cohen *d*<0.01, 95% CI −0.14 to 0.15; *P*=.97) and nonjudgment toward oneself (Cohen *d*=0.12, 95% CI −0.02 to 0.27; *P*=.10) domains. Neither the MEMI (Cohen *d*=0.03, 95% CI −0.12 to 0.17; *P*=.73) nor self-monitoring app (Cohen *d*=0.02, 95% CI −0.12 to 0.17; *P*=.77) significantly changed mindfulness from pretrial to midintervention time points. However, both the MEMI (Cohen *d*=0.61, 95% CI 0.46-0.76; *P*<.001) and self-monitoring app (Cohen *d*=0.53, 95% CI 0.38-0.68; *P*<.001) significantly increased nonjudgment toward oneself from pretrial to midintervention time points.

Testing ER and Self-Compassion as Mediators of Outcomes From Pre-Post Time Points

Neither ER nor self-compassion domains significantly mediated the effect of the treatment (brief MEMI vs self-monitoring) on pre- and postintervention SAD symptoms indexed by SPDQ and SPIN scores, generalized anxiety symptoms (GADQ-IV score), depression severity (BDI-II score), trait repetitive negative thinking (PCQ score), and trait mindfulness (FFMQ score; Table S3 in Multimedia Appendix 1). Therefore, hypothesis 2a was not supported.

Testing ER and Self-Compassion as Mediators of Outcomes From Pre-1MFU Time Points

Neither ER nor self-compassion domains significantly mediated the effect of the treatment (brief MEMI vs self-monitoring) on pre-1MFU SAD symptoms indexed by SPDQ and SPIN scores, generalized anxiety symptoms (GADQ-IV score), depression severity (BDI-II score), trait repetitive negative thinking (PCQ score), and trait mindfulness (FFMQ score; Table S4 in Multimedia Appendix 1). Therefore, hypothesis 2b was not supported. In a sensitivity analysis of single-mediator models, we found that mediation effects were still not statistically significant when the overall scores of DERS or SCS were tested as a single mediator for all 6 clinical outcomes.

Discussion

Comparison With Prior Work

This study determined the efficacy of a 14-day, fully self-guided, mobile MEMI on ER difficulties and self-compassion domains from pretrial to midintervention time points and whether such changes mediated 6 clinical outcomes from pre-post and pre-1MFU time points. Extending a prior report of primary outcomes [43], this RCT determined which domains of ER difficulties and self-compassion a brief MEMI could impact from pretrial to midintervention time points. Encouragingly, the results showed that, similar to applied relaxation and self-compassion training [92], brief, fully self-guided, mobile MEMIs efficaciously increased specific self-compassion domains. The MEMI enhanced acknowledging shared human struggles (common humanity), social connectedness (sense of being linked with other people), nonidentification with emotions (not overassociating with one’s feelings), and self-kindness domains from pretrial to midintervention time points. Therefore, the MEMI was superior to self-monitoring in increasing the ER domains linked with goal-directed behavior pursuit and emotional clarity.

Why was the brief MEMI more efficacious than self-monitoring in enhancing social connectedness, nonidentification with emotions, self-kindness, acknowledging shared human struggles, reducing goal-directed behavior difficulties, and lack of emotional clarity from pretrial to midintervention time points? Consistent with the mindfulness-to-meaning theory [93], the MEMI might have helped cultivate nonjudgmental awareness of one’s and others’ experiences, steering clear of both rumination and the suppression of distressing emotions typical of SAD. By teaching the regulation of physiological anxiety and the practice of self-kindness during moments of confusion or insecurity, the brief MEMI might offer greater benefits than self-monitoring in managing processes related to SAD. This approach might encourage constructive engagement with experiences instead of getting stuck in the vicious cycle of avoidance and missing out on positive interactions that could disconfirm feared social outcomes and offer feelings of self-efficacy across diverse interpersonal contexts [94,95]. Furthermore, the emphasis on acceptance and constructive awareness by the MEMI instead of self-monitoring might have contributed to these findings. Prior dismantling studies of MBIs that evidenced stronger comparative efficacy of acceptance and monitoring over monitoring alone on clinically relevant outcomes have alluded to this possibility [96].

Moreover, it is noteworthy that both brief MEMI and self-monitoring reduced ER difficulties linked to impulse control issues, nonacceptance of emotional responses, strategy use problems, and nonjudgment toward oneself from pretrial to midintervention time points. These null between-group effects could be viewed as a comparison between awareness alone and awareness combined with acceptance, where acceptance did not contribute anything additional to generate differential comparative efficacy on these outcomes. Perhaps learning via both the MEMI and self-monitoring to calmly acknowledge current thoughts, emotions, and bodily sensations without judgment, regardless of intensity or discomfort, fostered the ability to endure distress and engage with experiences rather than avoid them [97]. Furthermore, these processes might have motivated actions and curbed the inclination to engage in self-sabotaging behaviors, including self-criticism and impulsive actions [98,99]. In addition, these alterations may have been due to placebo and related factors rather than the influence of “active ingredients” [100]. Without the MEMI showing superior efficacy over self-monitoring on these outcomes, we cannot determine whether the improvements exceeded the effects of the passage of time, expectancy effects, and related factors. Future RCTs could empirically evaluate the validity of these conjectures.

Notably, self-monitoring, but not the MEMI, led to an increased emotional awareness. This surprising finding could be attributed to repeated instructions by self-monitoring prompts to attend to thoughts and feelings with explicit instructions to notice how distressing they might be. Another explanation might be that self-monitoring could have enhanced emotional self-awareness and related mental health outcomes, such as psychological well-being [101].

Another counterintuitive finding was the absence of between- and within-group effectiveness of the MEMI on the mindfulness domain of the SCS from pretrial to midintervention time points. A potential account for these nonsignificant outcomes was that more frequent, intense, and longer-duration mindfulness practices were necessary. Consistent with this inference, a 2-week MEMI positively affected trait mindfulness from pre-1MFU but not from pre- to postintervention time points [102]. Alternatively, the mindfulness measure used in this study might lack sensitivity to detect a change in mindfulness from pretrial to midintervention time points.

In addition, contrary to our expectations, no examined pretrial to midintervention ER or self-compassion domains were mediators of the effect of MEMI on the pre-post and pre-1MFU SAD symptoms, generalized anxiety symptoms, depression severity, repetitive negative thinking, and trait mindfulness. Other studies on brief MEMIs similarly found that ER and self-compassion did not consistently mediate treatment effects on mental health outcomes in various populations, such as distressed college students [103], health care workers [104], and other nonclinical samples [105]. These nonsignificant mediation effects might suggest that ER and self-compassion domains were not the change mechanisms of brief, fully self-guided, mobile MEMIs, as previously suggested [31,32]. On the basis of prior mechanism-focused trials that delivered higher-intensity MBIs [106-110], longer-duration and more rigorous MEMIs may be necessary to observe ER and self-compassion domains as mediators of the effect of brief MEMI on various clinical outcomes in SAD. Higher-intensity

MBIs may nurture stronger consolidation of mindfulness and related ER skills over long durations and across various contexts by modifying habitual responses to social and associated stressors in SAD [111]. Furthermore, it is possible that the ER and self-compassion measures used in this study were not sensitive enough to detect any existing true mediation effects. Alternatively, based on emerging evidence, other constructs, such as increased acceptance-attention, nonreactivity [112], and cognitive reappraisal [113], should be examined as mediators or proxy change mechanisms of brief MEMI on diverse SAD outcomes. Brief MEMIs could potentially enhance alternative mechanisms that boost mental well-being, coping skills, and related factors in individuals with SAD. These testable ideas await experimental evaluation.

Limitations and Strengths

This study had some limitations. First, all putative mediators and clinical outcome measures relied on web-based self-reports. Future RCTs should incorporate diverse measures of related constructs and evaluate crucial yet understudied factors believed to drive change, such as attentional bias toward threats [114,115], executive functioning [102,116,117], and physiological anxiety indices [118,119]. Second, we did not evaluate the sustainability of mindfulness skill use after the 2-week intervention duration. Third, we measured difficulties in ER and self-compassion domains at prerandomization and midintervention time points but not at posttreatment and 1MFU time points. This approach enabled us to evaluate the initial mechanism conditions and changes during the active treatment phase and test whether such conditions led to pre-1MFU changes in clinical outcomes [120]. Future investigations should delve into whether persistent mindfulness engagement, even without recurring MEMI instructions, can predict efficacy and change mechanisms at follow-up evaluations [121]. Fourth, the level of intensity in the brief MEMI may have been insufficient to produce substantial improvement in clinical outcomes via putative mediators. Finally, we used a clinical cut-off score on a self-report measure to determine a SAD diagnosis, instead of relying on a diagnostic interview.

Despite these limitations, this study had notable strengths, including a robust RCT design featuring an active comparator and high adherence rates, an uncommon achievement in the digital mental health intervention field, which often faces challenges such as high dropout and low use [122]. Furthermore, participants exhibited good compliance, with 78% (149/191) engaging in at least 80% of the MEMI and self-monitoring prompts. In addition, these measures have a well-established history of use and psychometric validation in previous trials [13,102,123,124]. Another strength of our study was the recruitment of a sample from Singapore, a Southeast Asian country. This method strengthened the potential for cross-cultural applicability, addressing a vital lacuna within the field of clinical psychology, which has traditionally concentrated on Western settings [76,125].

Conclusions

If replicated, our study has several clinical implications. Our observed advantages of a brief MEMI versus self-monitoring in decreasing isolation (sense of social disconnection) and overidentification (excessive attention to negative emotions) and increasing self-kindness (tenderness toward flaws in oneself), goal-directed behaviors, and emotional clarity in people with SAD indicated that established first-line treatments, such as CBT, could benefit from integrating self-compassion and mindfulness strategies. The finding that brief MEMI was not significantly different than self-monitoring in decreasing impulse control difficulties, nonacceptance of emotional responses, limited access to ER strategies, and self-judgment (excessive self-criticism) implies that the benefits of self-monitoring alone could be highlighted in the treatment to promote self-efficacy before moving on to graded exposure therapy and other CBT components for SAD. Higher-intensity brief, fully self-guided mobile MEMIs are likely necessary to identify which ER and self-compassion domains are change mechanisms [126]. In addition to identifying change mechanisms and establishing efficacy, it will be essential to investigate which individuals with SAD will derive the greatest benefit from the brief MEMI.

Acknowledgments

NHZ, HHT, RYH, and MGN take full responsibility for the data, accuracy of analyses and interpretation, and conduct of the research. This study was conducted in compliance with the American Psychological Association and Declaration of Helsinki ethical standards in treating human participants and was approved by the institutional review board. Informed consent was obtained from participants as per the National University of Singapore Institutional Review Board requirements. This randomized controlled trial received partial funding from the National Institute of Mental Health (R01 MH115128), the Pennsylvania State University Office of Research and Graduate Studies Dissertation Award, the Penn State Susan Welch-Nagle Family Graduate Fellowship, the National University of Singapore Development Grant, and the Association for Behavioral and Cognitive Therapies Leonard Krasner Student Dissertation Award.

Authors' Contributions

All authors have (1) made substantial contributions to the analysis and interpretation of the study and its findings, (2) drafted and revised the paper for intellectual content, and (3) given their final approval of the version to be submitted. The manuscript has been read and approved by all authors. NHZ was involved in the conceptualization, methodology (design and creation of models), funding acquisition, formal analysis, coding with the R software (R Foundation for Statistical Computing) for data analysis, visualization, and writing (original draft, reviewing, and editing) of this work. HHT assisted heavily with the data collection and various logistical aspects of the study. RYH and MGN assisted with the study conceptualization, design, data curation, validation, supervision, and writing (original draft, reviewing, and editing).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Study methodology details.
[DOCX File , 2071 KB - mental_v11i1e53712_app1.docx]

Multimedia Appendix 2
CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) diagram.
[PDF File (Adobe PDF File), 1348 KB - mental_v11i1e53712_app2.pdf]

References

1. Gross JJ. Emotion regulation: affective, cognitive, and social consequences. Psychophysiology 2002 May;39(3):281-291. [doi: 10.1017/s0048577201393198] [Medline: 12212647]
2. Lazarus RS. Progress on a cognitive-motivational-relational theory of emotion. Am Psychol 1991;46(8):819-834. [doi: 10.1037/0003-066x.46.8.819]
3. Jazaieri H, Morrison AS, Goldin PR, Gross JJ. The role of emotion and emotion regulation in social anxiety disorder. Curr Psychiatry Rep 2015 Jan;17(1):531. [doi: 10.1007/s11920-014-0531-3] [Medline: 25413637]
4. Aldao A, Nolen-Hoeksema S, Schweizer S. Emotion-regulation strategies across psychopathology: a meta-analytic review. Clin Psychol Rev 2010 Mar;30(2):217-237. [doi: 10.1016/j.cpr.2009.11.004] [Medline: 20015584]
5. Christiansen H, Hirsch O, Albrecht B, Chavanon ML. Attention-deficit/hyperactivity disorder (ADHD) and emotion regulation over the life span. Curr Psychiatry Rep 2019 Mar 02;21(3):17. [doi: 10.1007/s11920-019-1003-6] [Medline: 30826879]
6. Diedrich A, Grant M, Hofmann SG, Hiller W, Berking M. Self-compassion as an emotion regulation strategy in major depressive disorder. Behav Res Ther 2014 Jul;58:43-51. [doi: 10.1016/j.brat.2014.05.006] [Medline: 24929927]
7. Gilbert P. The origins and nature of compassion focused therapy. Br J Clin Psychol 2014 Mar 21;53(1):6-41. [doi: 10.1111/bjc.12043] [Medline: 24588760]
8. Neff KD. The development and validation of a scale to measure self-compassion. Self Identity 2003 Jul;2(3):223-250. [doi: 10.1080/15298860309027]
9. Neff K. Self-compassion: an alternative conceptualization of a healthy attitude toward oneself. Self Identity 2003 Apr;2(2):85-101. [doi: 10.1080/15298860309032]
10. Inwood E, Ferrari M. Mechanisms of change in the relationship between self-compassion, emotion regulation, and mental health: a systematic review. Appl Psychol Health Well Being 2018 Jul;10(2):215-235. [doi: 10.1111/aphw.12127] [Medline: 29673093]
11. Zainal NH, Newman MG. Relation between cognitive and behavioral strategies and future change in common mental health problems across 18 years. J Abnorm Psychol 2019 May;128(4):295-304 [FREE Full text] [doi: 10.1037/abn0000428] [Medline: 31045412]
12. Hong RY, Zainal NH, Ong XL. Longitudinal associations between academic competence-building and depression symptoms in early adolescence. Dev Psychopathol 2022 Aug 12;35(4):2061-2072. [doi: 10.1017/s0954579422000694]
13. Zainal NH, Chan WW, Saxena AP, Taylor CB, Newman MG. Pilot randomized trial of self-guided virtual reality exposure therapy for social anxiety disorder. Behav Res Ther 2021 Dec;147:103984 [FREE Full text] [doi: 10.1016/j.brat.2021.103984] [Medline: 34740099]
14. Lindsay EK, Creswell JD. Mindfulness, acceptance, and emotion regulation: perspectives from Monitor and Acceptance Theory (MAT). Curr Opin Psychol 2019 Aug;28:120-125 [FREE Full text] [doi: 10.1016/j.copsyc.2018.12.004] [Medline: 30639835]
15. Lindsay EK, Creswell JD. Mechanisms of mindfulness training: Monitor and Acceptance Theory (MAT). Clin Psychol Rev 2017 Feb;51:48-59 [FREE Full text] [doi: 10.1016/j.cpr.2016.10.011] [Medline: 27835764]
16. MacBeth A, Gumley A. Exploring compassion: a meta-analysis of the association between self-compassion and psychopathology. Clin Psychol Rev 2012 Aug;32(6):545-552. [doi: 10.1016/j.cpr.2012.06.003] [Medline: 22796446]
17. Wasson RS, Barratt C, O'Brien WH. Effects of mindfulness-based interventions on self-compassion in health care professionals: a meta-analysis. Mindfulness (N Y) 2020;11(8):1914-1934 [FREE Full text] [doi: 10.1007/s12671-020-01342-5] [Medline: 32421083]

18. Kabat-Zinn J, Massion AO, Kristeller J, Peterson LG, Fletcher KE, Pbert L, et al. Effectiveness of a meditation-based stress reduction program in the treatment of anxiety disorders. Am J Psychiatry 1992 Jul;149(7):936-943. [doi: 10.1176/ajp.149.7.936] [Medline: 1609875]
19. Segal Z, Williams JM, Williams M, Teasdale J. Mindfulness-Based Cognitive Therapy for Depression. New York, NY: Guilford Publications; 2012.
20. Creswell JD. Mindfulness interventions. Annu Rev Psychol 2017 Jan 03;68:491-516. [doi: 10.1146/annurev-psych-042716-051139] [Medline: 27687118]
21. Harvey LJ, White FA. Emotion self-stigma as a unique predictor of help-seeking intentions: a comparative analysis of early adolescents and young adults. Psychol Psychother 2023 Sep;96(3):762-777. [doi: 10.1111/papt.12467] [Medline: 37070732]
22. Dschaak ZA, Spiker DA, Berney EC, Miller ME, Hammer JH. Collegian help seeking: the role of self-compassion and self-coldness. J Ment Health 2021 Jun;30(3):284-291. [doi: 10.1080/09638237.2019.1677873] [Medline: 31661997]
23. Schumer MC, Lindsay EK, Creswell JD. Brief mindfulness training for negative affectivity: a systematic review and meta-analysis. J Consult Clin Psychol 2018 Jul;86(7):569-583 [FREE Full text] [doi: 10.1037/ccp0000324] [Medline: 29939051]
24. Tang YY, Tang R, Posner MI. Mindfulness meditation improves emotion regulation and reduces drug abuse. Drug Alcohol Depend 2016 Jun 01;163 Suppl 1:S13-S18 [FREE Full text] [doi: 10.1016/j.drugalcdep.2015.11.041] [Medline: 27306725]
25. Bernstein EE, LeBlanc NJ, Bentley KH, Barreira PJ, McNally RJ. A single-session workshop to enhance emotional awareness and emotion regulation for graduate students: a pilot study. Cognit Behav Pract 2021 Aug;28(3):393-409. [doi: 10.1016/j.cbpra.2020.09.008]
26. Kazdin AE. Drawing causal inferences from randomized controlled trials in psychotherapy research. Psychother Res 2023 Nov;33(8):991-1003. [doi: 10.1080/10503307.2022.2130112] [Medline: 36226476]
27. Lindsay EK, Young S, Brown KW, Smyth JM, Creswell JD. Mindfulness training reduces loneliness and increases social contact in a randomized controlled trial. Proc Natl Acad Sci U S A 2019 Feb 26;116(9):3488-3493 [FREE Full text] [doi: 10.1073/pnas.1813588116] [Medline: 30808743]
28. McLean CL, Ruork AK, Ramaiya MK, Fruzzetti AE. Feasibility and initial impact of single-session internet-delivered acceptance vs change skills for emotions for stress- and trauma-related problems: a randomized controlled trial. Behav Cogn Psychother 2023 Sep;51(5):443-458. [doi: 10.1017/S1352465823000206] [Medline: 37288653]
29. Martínez-Rubio D, Navarrete J, Montero-Marin J. Feasibility, effectiveness, and mechanisms of a brief mindfulness- and compassion-based program to reduce stress in university students: a pilot randomized controlled trial. Int J Environ Res Public Health 2021 Dec 23;19(1):154 [FREE Full text] [doi: 10.3390/ijerph19010154] [Medline: 35010414]
30. Bergen-Cico D, Possemato K, Cheon S. Examining the efficacy of a brief mindfulness-based stress reduction (Brief MBSR) program on psychological health. J Am Coll Health 2013;61(6):348-360. [doi: 10.1080/07448481.2013.813853] [Medline: 23930749]
31. Gratz KL, Tull MT. Emotion regulation as a mechanism of change in acceptance-and mindfulness-based treatments. In: Assessing Mindfulness and Acceptance: Illuminating the Processes of Change. Oakland, CA: New Harbinger Publications; 2010.
32. Baer RA. Self-compassion as a mechanism of change in mindfulness- and acceptance-based treatments. In: Baer RA, editor. Assessing Mindfulness and Acceptance Processes in Clients: Illuminating the Theory and Practice of Change. Oakland, CA: New Harbinger Publications; 2010:135-153.
33. Maloney S, Surawy C, Martin M, Montero-Marin J, Kuyken W. The state- and trait-level effects and candidate mechanisms of four mindfulness-based cognitive therapy (MBCT) practices: two exploratory studies. Mindfulness (N Y) 2023 Aug 24;14(9):2155-2171 [FREE Full text] [doi: 10.1007/s12671-023-02193-6] [Medline: 37795338]
34. Ma Y, She Z, Siu AF, Zeng X, Liu X. Effectiveness of online mindfulness-based interventions on psychological distress and the mediating role of emotion regulation. Front Psychol 2018;9:2090 [FREE Full text] [doi: 10.3389/fpsyg.2018.02090] [Medline: 30429816]
35. Kraemer HC, Kiernan M, Essex M, Kupfer DJ. How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. Health Psychol 2008 Mar;27(2S):S101-S108 [FREE Full text] [doi: 10.1037/0278-6133.27.2(Suppl.).S101] [Medline: 18377151]
36. Werner KH, Goldin PR, Ball TM, Heimberg RG, Gross JJ. Assessing emotion regulation in social anxiety disorder: the emotion regulation interview. J Psychopathol Behav Assess 2011 Apr 15;33(3):346-354. [doi: 10.1007/s10862-011-9225-x]
37. Werner KH, Jazaieri H, Goldin PR, Ziv M, Heimberg RG, Gross JJ. Self-compassion and social anxiety disorder. Anxiety Stress Coping 2012;25(5):543-558 [FREE Full text] [doi: 10.1080/10615806.2011.608842] [Medline: 21895450]
38. Blackie RA, Kocovski NL. Forgive and let go: effect of self-compassion on post-event processing in social anxiety. Mindfulness 2017 Sep 15;9(2):654-663. [doi: 10.1007/s12671-017-0808-9]
39. Goodman FR, Kashdan TB, İmamoğlu A. Valuing emotional control in social anxiety disorder: a multimethod study of emotion beliefs and emotion regulation. Emotion 2021 Jun;21(4):842-855. [doi: 10.1037/emo0000750] [Medline: 32191093]
40. Dixon ML, Moodie CA, Goldin PR, Farb N, Heimberg RG, Gross JJ. Emotion regulation in social anxiety disorder: reappraisal and acceptance of negative self-beliefs. Biol Psychiatry Cogn Neurosci Neuroimaging 2020 Jan;5(1):119-129. [doi: 10.1016/j.bpsc.2019.07.009] [Medline: 31547972]

41. Gratz KL, Roemer L. Multidimensional assessment of emotion regulation and dysregulation: development, factor structure, and initial validation of the difficulties in emotion regulation scale. J Psychopathol Behav Assess 2004 Mar;26(1):41-54. [doi: 10.1023/B:JOBA.0000007455.08539.94]
42. Neff KD, Tóth-Király I. Self-Compassion Scale (SCS). In: Medvedev ON, Krägeloh CU, Siegert RJ, Singh NN, editors. Handbook of Assessment in Mindfulness Research. Cham, Switzerland: Springer International Publishing; 2022.
43. Zainal NH, Tan HH, Hong RY, Newman MG. Is a brief mindfulness ecological momentary intervention more efficacious than a self-monitoring app for social anxiety disorder? a randomized controlled trial. J Anxiety Disord (forthcoming) 2024 Mar 24:102858. [doi: 10.1016/j.janxdis.2024.102858]
44. Zainal NH, Tan HH, Hong RY, Newman MG. Is a brief mindfulness app efficacious in enhancing emotion regulation and self-compassion in social anxiety disorder? a randomized controlled trial and mediation analysis. OSF Registries. 2023. URL: https://osf.io/m3kxz [accessed 2023-09-03]
45. Connor KM, Davidson JR, Churchill LE, Sherwood A, Foa E, Weisler RH. Psychometric properties of the Social Phobia Inventory (SPIN). New self-rating scale. Br J Psychiatry 2000 Apr 02;176(4):379-386. [doi: 10.1192/bjp.176.4.379] [Medline: 10827888]
46. DSM-5-TR online assessment measures. American Psychiatric Association. URL: https://www.psychiatry.org/psychiatrists/practice/dsm/educational-resources/assessment-measures [accessed 2024-03-29]
47. Di Nardo PA, Brown TA, Barlow DH. Anxiety Disorder Interview Schedule for DSM-IV: Lifetime version (ADIS-IV-L). New York, NY: Oxford University Press; 1994.
48. Newman MG, Kachin KE, Zuellig AR, Constantino MJ, Cashman-McGrath L. The social phobia diagnostic questionnaire: preliminary validation of a new self-report diagnostic measure of social phobia. Psychol Med 2003 May;33(4):623-635. [doi: 10.1017/s0033291703007669] [Medline: 12785464]
49. Antony MM, Coons MJ, McCabe RE, Ashbaugh A, Swinson RP. Psychometric properties of the social phobia inventory: further evaluation. Behav Res Ther 2006 Aug;44(8):1177-1185. [doi: 10.1016/j.brat.2005.08.013] [Medline: 16257387]
50. Osório FL, Crippa JA, Loureiro SR. Evaluation of the psychometric properties of the Social Phobia Inventory in university students. Compr Psychiatry 2010;51(6):630-640. [doi: 10.1016/j.comppsych.2010.03.004] [Medline: 20965310]
51. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
52. PACO is an opensource, mobile, behavorial research platform. GitHub. URL: https://github.com/google/paco [accessed 2023-09-09]
53. Kabat-Zinn J. Mindfulness-based stress reduction (MBSR). Construct Hum Sci 2003;8(2):73-107 [FREE Full text]
54. Heckenberg RA, Eddy P, Kent S, Wright BJ. Do workplace-based mindfulness meditation programs improve physiological indices of stress? a systematic review and meta-analysis. J Psychosom Res 2018 Nov;114:62-71. [doi: 10.1016/j.jpsychores.2018.09.010] [Medline: 30314581]
55. Kraemer KM, Luberto CM, Hall DL, Ngo LH, Yeh GY. A systematic review and meta-analysis of mindfulness- and acceptance-based interventions for affect intolerance/sensitivity. Behav Res Ther 2020 Dec;135:103746 [FREE Full text] [doi: 10.1016/j.brat.2020.103746] [Medline: 33011486]
56. Devilly GJ, Borkovec TD. Psychometric properties of the credibility/expectancy questionnaire. J Behav Ther Exp Psychiatry 2000 Jun;31(2):73-86. [doi: 10.1016/s0005-7916(00)00012-4] [Medline: 11132119]
57. LaFreniere LS, Newman MG. A brief ecological momentary intervention for generalized anxiety disorder: a randomized controlled trial of the worry outcome journal. Depress Anxiety 2016 Sep;33(9):829-839. [doi: 10.1002/da.22507] [Medline: 27062682]
58. LaFreniere LS, Newman MG. Exposing worry's deceit: percentage of untrue worries in generalized anxiety disorder treatment. Behav Ther 2020 May;51(3):413-423 [FREE Full text] [doi: 10.1016/j.beth.2019.07.003] [Medline: 32402257]
59. Cunningham JA, Kypri K, McCambridge J. Exploratory randomized controlled trial evaluating the impact of a waiting list control design. BMC Med Res Methodol 2013 Dec 06;13:150 [FREE Full text] [doi: 10.1186/1471-2288-13-150] [Medline: 24314204]
60. Fowler JC, Charak R, Elhai JD, Allen JG, Frueh BC, Oldham JM. Construct validity and factor structure of the difficulties in emotion regulation scale among adults with severe mental illness. J Psychiatr Res 2014 Nov;58:175-180. [doi: 10.1016/j.jpsychires.2014.07.029] [Medline: 25171941]
61. Neff KD, McGehee P. Self-compassion and psychological resilience among adolescents and young adults. Self Identity 2010;9(3):225-240. [doi: 10.1080/15298860902979307]
62. Neff KD, Rude SS, Kirkpatrick KL. An examination of self-compassion in relation to positive psychological functioning and personality traits. J Res Pers 2007 Aug;41(4):908-916. [doi: 10.1016/j.jrp.2006.08.002]
63. Neff KD. The self-compassion scale is a valid and theoretically coherent measure of self-compassion. Mindfulness 2015 Dec 18;7:264-274. [doi: 10.1007/s12671-015-0479-3]
64. Diagnostic and Statistical Manual of Mental Disorders DSM-IV. Washington, DC: American Psychiatric Association; 1994.

65. Fogliati VJ, Terides MD, Gandy M, Staples LG, Johnston L, Karin E, et al. Psychometric properties of the Mini-Social Phobia Inventory (Mini-SPIN) in a large online treatment-seeking sample. Cogn Behav Ther 2016 Apr;45(3):236-257. [doi: 10.1080/16506073.2016.1158206] [Medline: 27046641]
66. First MB. User's Guide for the Structured Clinical Interview for DSM-IV Axis I Disorders SCID-I Clinician Version. Washington, DC: American Psychiatric Press; 1997.
67. Crippa JA, de Lima Osório F, Del-Ben CM, Filho AS, da Silva Freitas MC, Loureiro SR. Comparability between telephone and face-to-face structured clinical interview for DSM-IV in assessing social anxiety disorder. Perspect Psychiatr Care 2008 Oct;44(4):241-247. [doi: 10.1111/j.1744-6163.2008.00183.x] [Medline: 18826462]
68. Nagata T, Nakajima T, Teo AR, Yamada H, Yoshimura C. Psychometric properties of the Japanese version of the social phobia inventory. Psychiatry Clin Neurosci 2013 Apr;67(3):160-166 [FREE Full text] [doi: 10.1111/pcn.12037] [Medline: 23581867]
69. Newman MG, Zuellig AR, Kachin KE, Constantino MJ, Przeworski A, Erickson T, et al. Preliminary reliability and validity of the generalized anxiety disorder questionnaire-IV: a revised self-report diagnostic measure of generalized anxiety disorder. Behav Ther 2002;33(2):215-233. [doi: 10.1016/S0005-7894(02)80026-0]
70. Moore MT, Anderson NL, Barnes JM, Haigh EA, Fresco DM. Using the GAD-Q-IV to identify generalized anxiety disorder in psychiatric treatment seeking and primary care medical samples. J Anxiety Disord 2014 Jan;28(1):25-30. [doi: 10.1016/j.janxdis.2013.10.009] [Medline: 24334213]
71. Zainal NH, Newman MG. Worry amplifies theory-of-mind reasoning for negatively valenced social stimuli in generalized anxiety disorder. J Affect Disord 2018 Feb;227:824-833 [FREE Full text] [doi: 10.1016/j.jad.2017.11.084] [Medline: 29254067]
72. Beck AT, Steer RA, Ball R, Ranieri W. Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. J Pers Assess 1996 Dec;67(3):588-597. [doi: 10.1207/s15327752jpa6703_13] [Medline: 8991972]
73. García-Batista ZE, Guerra-Peña K, Cano-Vindel A, Herrera-Martínez SX, Medrano LA. Validity and reliability of the Beck Depression Inventory (BDI-II) in general and hospital population of Dominican Republic. PLoS One 2018 Jun 29;13(6):e0199750 [FREE Full text] [doi: 10.1371/journal.pone.0199750] [Medline: 29958268]
74. Huang SL, Hsieh CL, Wu RM, Lu WS. Test-retest reliability and minimal detectable change of the Beck Depression Inventory and the Taiwan Geriatric Depression Scale in patients with Parkinson's disease. PLoS One 2017 Sep 25;12(9):e0184823 [FREE Full text] [doi: 10.1371/journal.pone.0184823] [Medline: 28945776]
75. Szkodny LE, Newman MG. Delineating characteristics of maladaptive repetitive thought: development and preliminary validation of the perseverative cognitions questionnaire. Assessment 2019 Sep;26(6):1084-1104 [FREE Full text] [doi: 10.1177/1073191117698753] [Medline: 28355881]
76. Zainal NH, Newman MG, Hong RY. Cross-cultural and gender invariance of transdiagnostic processes in the United States and Singapore. Assessment 2021 Mar;28(2):485-502 [FREE Full text] [doi: 10.1177/1073191119869832] [Medline: 31538795]
77. Baer RA, Smith GT, Lykins E, Button D, Krietemeyer J, Sauer S, et al. Construct validity of the five facet mindfulness questionnaire in meditating and nonmeditating samples. Assessment 2008 Sep;15(3):329-342. [doi: 10.1177/1073191107313003] [Medline: 18310597]
78. Goldberg SB, Wielgosz J, Dahl C, Schuyler B, MacCoon DS, Rosenkranz M, et al. Does the five facet mindfulness questionnaire measure what we think it does? construct validity evidence from an active controlled randomized clinical trial. Psychol Assess 2016 Aug;28(8):1009-1014 [FREE Full text] [doi: 10.1037/pas0000233] [Medline: 26460893]
79. Veehof MM, Ten Klooster PM, Taal E, Westerhof GJ, Bohlmeijer ET. Psychometric properties of the Dutch Five Facet Mindfulness Questionnaire (FFMQ) in patients with fibromyalgia. Clin Rheumatol 2011 Aug;30(8):1045-1054 [FREE Full text] [doi: 10.1007/s10067-011-1690-9] [Medline: 21347605]
80. powerlmm: power analysis for longitudinal multilevel models. The Comprehensive R Archive Network. URL: https://CRAN.R-project.org/package=powerlmm [accessed 2023-09-09]
81. Arend MG, Schäfer T. Statistical power in two-level models: a tutorial based on Monte Carlo simulation. Psychol Methods 2019 Feb;24(1):1-19. [doi: 10.1037/met0000195] [Medline: 30265048]
82. McCoy CE. Understanding the intention-to-treat principle in randomized controlled trials. West J Emerg Med 2017 Oct;18(6):1075-1078 [FREE Full text] [doi: 10.5811/westjem.2017.8.35985] [Medline: 29085540]
83. Lee T, Shi D. A comparison of full information maximum likelihood and multiple imputation in structural equation modeling with missing data. Psychol Methods 2021 Aug;26(4):466-485. [doi: 10.1037/met0000381] [Medline: 33507765]
84. Grimm KJ, Ram N, Estabrook R. Growth Modeling: Structural Equation and Multilevel Modeling Approaches. New York City, NY: Guilford Press; 2016.
85. Rosseel Y. Lavaan: an R package for structural equation modeling. J Stat Softw 2012;48(2):1-36. [doi: 10.18637/jss.v048.i02]
86. Zhao X, Lynch JGJ, Chen Q. Reconsidering Baron and Kenny: myths and truths about mediation analysis. J Consum Res 2010 Aug;37(2):197-206. [doi: 10.1086/651257]
87. Stanton AL, Luecken LJ, MacKinnon DP, Thompson EH. Mechanisms in psychosocial interventions for adults living with cancer: opportunity for integration of theory, research, and practice. J Consult Clin Psychol 2013 Apr;81(2):318-335. [doi: 10.1037/a0028833] [Medline: 22663900]

88. Maxwell SE, Cole DA. Bias in cross-sectional analyses of longitudinal mediation. Psychol Methods 2007 Mar;12(1):23-44. [doi: 10.1037/1082-989X.12.1.23] [Medline: 17402810]
89. Lachowicz MJ, Preacher KJ, Kelley K. A novel measure of effect size for mediation analysis. Psychol Methods 2018 Jun;23(2):244-261. [doi: 10.1037/met0000165] [Medline: 29172614]
90. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. Biometrika 1986 Dec;73(3):751-754. [doi: 10.1093/biomet/73.3.751]
91. Feingold A. Confidence interval estimation for standardized effect sizes in multilevel and latent growth modeling. J Consult Clin Psychol 2015 Feb;83(1):157-168 [FREE Full text] [doi: 10.1037/a0037721] [Medline: 25181028]
92. Teale Sapach MJ, Carleton RN. Self-compassion training for individuals with social anxiety disorder: a preliminary randomized controlled trial. Cogn Behav Ther 2023 Jan;52(1):18-37. [doi: 10.1080/16506073.2022.2130820] [Medline: 36254613]
93. Garland EL, Hanley AW, Goldin PR, Gross JJ. Testing the mindfulness-to-meaning theory: evidence for mindful positive emotion regulation from a reanalysis of longitudinal data. PLoS One 2017 Dec 6;12(12):e0187727 [FREE Full text] [doi: 10.1371/journal.pone.0187727] [Medline: 29211754]
94. Guendelman S, Medeiros S, Rampes H. Mindfulness and emotion regulation: insights from neurobiological, psychological, and clinical studies. Front Psychol 2017 Mar 06;8:220 [FREE Full text] [doi: 10.3389/fpsyg.2017.00220] [Medline: 28321194]
95. Curreri AJ, Farchione TJ, Sauer-Zavala S, Barlow DH. Mindful emotion awareness facilitates engagement with exposure therapy: an idiographic exploration using single case experimental design. Behav Modif 2022 Jan;46(1):36-62. [doi: 10.1177/0145445520947662] [Medline: 32752883]
96. Chin B, Lindsay EK, Greco CM, Brown KW, Smyth JM, Wright AG, et al. Psychological mechanisms driving stress resilience in mindfulness training: a randomized controlled trial. Health Psychol 2019 Aug;38(8):759-768 [FREE Full text] [doi: 10.1037/hea0000763] [Medline: 31120272]
97. Herscu O, Somer E, Federman A, Soffer-Dudek N. Mindfulness meditation and self-monitoring reduced maladaptive daydreaming symptoms: a randomized controlled trial of a brief self-guided web-based program. J Consult Clin Psychol 2023 May;91(5):285-300. [doi: 10.1037/ccp0000790] [Medline: 36634023]
98. Roemer L, Williston SK, Rollins LG. Mindfulness and emotion regulation. Curr Opin Psychol 2015 Jun;3:52-57. [doi: 10.1016/j.copsyc.2015.02.006]
99. Ruffault A, Czernichow S, Hagger MS, Ferrand M, Erichot N, Carette C, et al. The effects of mindfulness training on weight-loss and health-related behaviours in adults with overweight and obesity: a systematic review and meta-analysis. Obes Res Clin Pract 2017;11(5 Suppl 1):90-111. [doi: 10.1016/j.orcp.2016.09.002] [Medline: 27658995]
100. Kazdin AE. Research Design in Clinical Psychology. London, UK: Pearson Education; 2016.
101. Bakker D, Rickard N. Engagement in mobile phone app for self-monitoring of emotional wellbeing predicts changes in mental health: MoodPrism. J Affect Disord 2018 Feb;227:432-442. [doi: 10.1016/j.jad.2017.11.016] [Medline: 29154165]
102. Zainal NH, Newman MG. A randomized controlled trial of a 14-day mindfulness ecological momentary intervention (MEMI) for generalized anxiety disorder. Eur Psychiatry 2023 Jan 16;66(1):e12 [FREE Full text] [doi: 10.1192/j.eurpsy.2023.2] [Medline: 36645098]
103. Sun S, Lin D, Goldberg S, Shen Z, Chen P, Qiao S, et al. A mindfulness-based mobile health (mHealth) intervention among psychologically distressed university students in quarantine during the COVID-19 pandemic: a randomized controlled trial. J Couns Psychol 2022 Mar;69(2):157-171 [FREE Full text] [doi: 10.1037/cou0000568] [Medline: 34264696]
104. Keng SL, Chin JW, Mammadova M, Teo I. Effects of mobile app-based mindfulness practice on healthcare workers: a randomized active controlled trial. Mindfulness (N Y) 2022 Sep 16;13(11):2691-2704 [FREE Full text] [doi: 10.1007/s12671-022-01975-8] [Medline: 36160038]
105. Golden HL, Vosper J, Kingston J, Ellett L. The impact of mindfulness-based programmes on self-compassion in nonclinical populations: a systematic review and meta-analysis. Mindfulness 2020 Sep 25;12(1):29-52. [doi: 10.1007/s12671-020-01501-8]
106. Sanilevici M, Reuveni O, Lev-Ari S, Golland Y, Levit-Binnun N. Mindfulness-based stress reduction increases mental wellbeing and emotion regulation during the first wave of the COVID-19 pandemic: a synchronous online intervention study. Front Psychol 2021;12:720965 [FREE Full text] [doi: 10.3389/fpsyg.2021.720965] [Medline: 34858260]
107. Buxton AE, Hofheinz C, Remmers C, Michalak J. When depressed-be mindful of and kind to yourself: self-compassion as mediator of change in a mindful depression treatment. Psychother Res 2024 Feb;34(2):182-194. [doi: 10.1080/10503307.2023.2179439] [Medline: 36803684]
108. tefan CA, Căpraru C, Szilágyi M. Investigating effects and mechanisms of a mindfulness-based stress reduction intervention in a sample of college students at risk for social anxiety. Mindfulness 2018 Feb 14;9(5):1509-1521. [doi: 10.1007/s12671-018-0899-y]
109. Sevel LS, Finn MT, Smith RM, Ryden AM, McKernan LC. Self-compassion in mindfulness-based stress reduction: an examination of prediction and mediation of intervention effects. Stress Health 2020 Feb;36(1):88-96 [FREE Full text] [doi: 10.1002/smi.2917] [Medline: 31874122]

110. Keng SL, Smoski MJ, Robins CJ, Ekblad AG, Brantley JG. Mechanisms of change in mindfulness-based stress reduction: self-compassion and mindfulness as mediators of intervention outcomes. J Cogn Psychother 2012 Jan 01;26(3):270-280. [doi: 10.1891/0889-8391.26.3.270]
111. Goldin PR, Morrison AS, Jazaieri H, Heimberg RG, Gross JJ. Trajectories of social anxiety, cognitive reappraisal, and mindfulness during an RCT of CBGT versus MBSR for social anxiety disorder. Behav Res Ther 2017 Oct;97:1-13 [FREE Full text] [doi: 10.1016/j.brat.2017.06.001] [Medline: 28654771]
112. Gavrilova L, Zawadzki MJ. Examining how headspace impacts mindfulness mechanisms over an 8-week app-based mindfulness intervention. Mindfulness 2023 Sep 22;14(9):2236-2249. [doi: 10.1007/s12671-023-02214-4]
113. Schulte-Frankenfeld PM, Trautwein FM. App-based mindfulness meditation reduces perceived stress and improves self-regulation in working university students: a randomised controlled trial. Appl Psychol Health Well Being 2022 Nov;14(4):1151-1171 [FREE Full text] [doi: 10.1111/aphw.12328] [Medline: 34962055]
114. Hsu KJ, Shumake J, Caffey K, Risom S, Labrada J, Smits JA, et al. Efficacy of attention bias modification training for depressed adults: a randomized clinical trial. Psychol Med 2021 Mar 26;52(16):3865-3873. [doi: 10.1017/s0033291721000702]
115. Zainal NH, Hellberg SN, Kabel KE, Hotchkin CM, Baker AW. Cognitive behavioral therapy (CBT) plus attention bias modification (ABM) reduces anxiety sensitivity and depressive symptoms in panic disorder: a pilot randomized trial. Scand J Psychol 2023 Aug;64(4):390-400. [doi: 10.1111/sjop.12902] [Medline: 36707979]
116. Zainal NH, Newman MG. Mindfulness enhances cognitive functioning: a meta-analysis of 111 randomized controlled trials. Health Psychol Rev 2023 Aug 30:1-27 (forthcoming). [doi: 10.1080/17437199.2023.2248222] [Medline: 37578065]
117. Zainal NH, Newman MG. Executive functioning constructs in anxiety, obsessive-compulsive, post-traumatic stress, and related disorders. Curr Psychiatry Rep 2022 Dec;24(12):871-880 [FREE Full text] [doi: 10.1007/s11920-022-01390-9] [Medline: 36401677]
118. Papenfuss I, Lommen MJ, Huisman M, Ostafin BD. Aversive response to uncertainty as a mediator for the effect of a mindfulness intervention on symptoms of anxiety. Int J Psychophysiol 2022 Sep;179:30-42 [FREE Full text] [doi: 10.1016/j.ijpsycho.2022.06.014] [Medline: 35753561]
119. Matos M, Duarte C, Duarte J, Pinto-Gouveia J, Petrocchi N, Gilbert P. Cultivating the compassionate self: an exploration of the mechanisms of change in compassionate mind training. Mindfulness 2021 Oct 21;13(1):66-79. [doi: 10.1007/s12671-021-01717-2]
120. Bauer DJ, Preacher KJ, Gil KM. Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: new procedures and recommendations. Psychol Methods 2006 Jun;11(2):142-163. [doi: 10.1037/1082-989X.11.2.142] [Medline: 16784335]
121. Kazdin AE. Understanding how and why psychotherapy leads to change. Psychother Res 2009 Jul 22;19(4-5):418-428. [doi: 10.1080/10503300802448899] [Medline: 19034715]
122. Torous J, Lipschitz J, Ng M, Firth J. Dropout rates in clinical trials of smartphone apps for depressive symptoms: a systematic review and meta-analysis. J Affect Disord 2020 Feb 15;263:413-419. [doi: 10.1016/j.jad.2019.11.167] [Medline: 31969272]
123. Bieber M, Görgülü E, Schmidt D, Zabel K, Etyemez S, Friedrichs B, et al. Effects of body-oriented yoga: a RCT study for patients with major depressive disorder. Eur Arch Psychiatry Clin Neurosci 2021 Oct;271(7):1217-1229 [FREE Full text] [doi: 10.1007/s00406-021-01277-5] [Medline: 34245349]
124. Samantaray NN, Nath B, Behera N, Mishra A, Singh P, Sudhir P. Brief cognitive behavior group therapy for social anxiety among medical students: a randomized placebo-controlled trial. Asian J Psychiatr 2021 Jan;55:102526. [doi: 10.1016/j.ajp.2020.102526] [Medline: 33360708]
125. Van Doren N, Zainal NH, Newman MG. Cross-cultural and gender invariance of emotion regulation in the United States and India. J Affect Disord 2021 Dec 01;295:1360-1370 [FREE Full text] [doi: 10.1016/j.jad.2021.04.089] [Medline: 34706449]
126. Koszycki D, Thake J, Mavounza C, Daoust JP, Taljaard M, Bradwejn J. Preliminary investigation of a mindfulness-based intervention for social anxiety disorder that integrates compassion meditation and mindful exposure. J Altern Complement Med 2016 May;22(5):363-374. [doi: 10.1089/acm.2015.0108] [Medline: 27070853]

Abbreviations

1MFU: 1-month follow-up
BDI-II: Beck Depression Inventory–second edition
CBT: cognitive behavioral therapy
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
DERS: Difficulties in Emotion Regulation Scale
DSM: Diagnostic and Statistical Manual
ER: emotion regulation
FFMQ: Five-Factor Mindfulness Questionnaire

GAD: generalized anxiety disorder
GADQ-IV: Generalized Anxiety Disorder Questionnaire–fourth edition
MBI: mindfulness-based intervention
MEMI: mindfulness ecological momentary intervention
NUS: National University of Singapore
OSF: Open Science Framework
PCQ: Perseverative Cognitions Questionnaire
RCT: randomized controlled trial
SAD: social anxiety disorder
SCS: Self-Compassion Scale
SPDQ: Social Phobia Diagnostic Questionnaire
SPIN: Social Phobia Inventory

Edited by J Torous; submitted 16.10.23; peer-reviewed by L LaFreniere, N Van Doren; comments to author 27.11.23; revised version received 27.12.23; accepted 12.01.24; published 19.04.24.

Please cite as:
Zainal NH, Tan HH, Hong RYS, Newman MG
Testing the Efficacy of a Brief, Self-Guided Mindfulness Ecological Momentary Intervention on Emotion Regulation and Self-Compassion in Social Anxiety Disorder: Randomized Controlled Trial
JMIR Ment Health 2024;11:e53712
URL: https://mental.jmir.org/2024/1/e53712
doi:10.2196/53712
PMID:38640015

Original Paper

Examining the Effects of a Brief, Fully Self-Guided Mindfulness Ecological Momentary Intervention on Empathy and Theory-of-Mind for Generalized Anxiety Disorder: Randomized Controlled Trial

Nur Hani Zainal[1,2], BS, MS, PhD; Michelle G Newman[3], BS, MA, PhD

[1]Department of Health Care Policy, Harvard Medical School, Boston, MA, United States

[2]Department of Psychology, National University of Singapore, Singapore, Singapore

[3]Department of Psychology, The Pennsylvania State University, University Park, PA, United States

Corresponding Author:
Nur Hani Zainal, BS, MS, PhD
Department of Psychology
National University of Singapore
9 Arts Link
Singapore, 117572
Singapore
Phone: 65 917 767 7088
Fax: 65 6516 3749
Email: hanizainal@nus.edu.sg

Abstract

Background: The utility of brief mindfulness ecological momentary interventions (EMIs) to improve empathy and theory-of-mind has been underinvestigated, particularly in generalized anxiety disorder (GAD).

Objective: In this randomized controlled trial, we aimed to examine the efficacy of a 14-day, fully self-guided, mindfulness EMI on the empathy and theory-of-mind domains for GAD.

Methods: Adults (aged ≥18 y) diagnosed with GAD were randomized to a mindfulness EMI (68/110, 61.8%) or self-monitoring app (42/110, 38.2%) arm. They completed the Interpersonal Reactivity Index self-report empathy measure and theory-of-mind test (Bell-Lysaker Emotion Recognition Task) at prerandomization, postintervention, and 1-month follow-up (1MFU) time points. Hierarchical linear modeling was conducted with the intent-to-treat principle to determine prerandomization to postintervention (pre-post intervention) and prerandomization to 1MFU (pre-1MFU) changes, comparing the mindfulness EMI to self-monitoring.

Results: Observed effects were generally stronger from pre-1MFU than from pre-post intervention time points. From pre-post intervention time points, the mindfulness EMI was more efficacious than the self-monitoring app on *fantasy* (the ability to imagine being in others' shoes; between-intervention effect size: Cohen d=0.26, P=.007; within-intervention effect size: Cohen d=0.22, P=.02 for the mindfulness EMI and Cohen d=−0.16, P=.10 for the self-monitoring app). From pre-1MFU time points, the mindfulness EMI, but not the self-monitoring app, improved *theory-of-mind* (a window into others' thoughts and intentions through abstract, propositional knowledge about their mental states, encompassing the ability to decipher social cues) and the *fantasy*, *personal distress* (stress when witnessing others' negative experiences), and *perspective-taking* (understanding others' perspective) empathy domains. The effect sizes were small to moderate (Cohen d=0.15-0.36; P<.001 to P=.01) for significant between-intervention effects from pre-1MFU time points. Furthermore, the within-intervention effect sizes for these significant outcomes were stronger for the mindfulness EMI (Cohen d=0.30-0.43; P<.001 to P=.03) than the self-monitoring app (Cohen d=−0.12 to 0.21; P=.001 to P>.99) from pre-1MFU time points. No between-intervention and within-intervention effects on *empathic concern* (feeling affection, compassion, and care when observing others in distress, primarily attending to their emotional well-being) were observed from pre-post intervention and pre-1MFU time points.

Conclusions: The brief mindfulness EMI improved specific domains of empathy (eg, fantasy, personal distress, and perspective-taking) and theory-of-mind with small to moderate effect sizes in persons with GAD. Higher-intensity, self-guided or coach-facilitated, multicomponent mindfulness EMIs targeting the optimization of social relationships are likely necessary to improve the empathic concern domain in this population.

Trial Registration: ClinicalTrials.gov NCT04846777; https://clinicaltrials.gov/study/NCT04846777

(JMIR Ment Health 2024;11:e54412) doi:10.2196/54412

KEYWORDS

empathy; theory-of-mind; mindfulness; ecological momentary intervention; generalized anxiety disorder; randomized controlled trial; mobile phone

Introduction

Background

Empathy encompasses the capacity to comprehend and resonate with the emotional experiences of others, facilitating caregiving, knowledge sharing, and collaborative goal attainment [1,2]. Relatedly, *theory-of-mind* (ToM) offers insight into peoples' thoughts and intentions via abstract, propositional knowledge concerning others' mental states and encompasses the capacity to interpret social cues from others [3,4]. Both empathy and ToM play vital roles in comprehending the cognitive and emotional dynamics of others in social contexts [3]. Specifically, empathy and ToM have been linked to enhanced emotional well-being [5], stronger social connections [6], and improved social health [7]. Therefore, developing efficacious interventions to improve various empathy domains and ToM is essential.

In particular, it is possible that mindfulness-based interventions (MBIs) could be efficacious in enhancing empathy and ToM skills [8]. Broadly, numerous theorists asserted that practicing mindfulness diligently should inherently give rise to interpersonal growth, including kindness, compassion, and empathy [9]. Plausibly, the receptive, nonjudgmental attitude fostered by mindfulness exercises could be pivotal in nurturing empathy toward others [10,11]. In addition, practicing mindfulness offers a conducive environment for efforts to comprehend another individual while acknowledging the inherent incompleteness of such understanding [12]. Collectively, mindfulness practices via MBIs can enhance individuals' attunement to others' circumstances, deepen their emotional resonance, increase their willingness to engage with negative emotions, and foster compassion toward others' experiences as an extension of self-compassion [13].

Supporting these theories, a narrative systematic review provided compelling evidence for the positive impact of MBIs on enhancing empathy in children and adolescents [14]. Similarly, in an open trial, an 8-week mindfulness-based cognitive therapy (MBCT) [15] among university students improved all empathy domains with moderate to large effects at pre-post intervention time points (Cohen d=0.48-1.19) [16]; however, the lack of a randomly assigned control intervention precluded the ability to make cause-effect inferences. Lending credence to this idea, meta-analytic data from experiments and randomized controlled trials (RCTs) indicated that MBIs were superior to control interventions in enhancing empathy in the healthy general population [17]. A particular case was how premedical and medical students' participation in a mindfulness-based stress reduction (MBSR) [18] program significantly elevated their overall empathy levels compared to those in the control intervention among premedical and medical students [19]. Another study showed that a 3-month perspective-taking–focused MBI led to greater ToM performance compared to an emotion-focused MBI [20]. Similarly, a recent pilot trial suggested the promise of MBIs in enhancing ToM in individuals with psychotic disorders [21]. Overall, the literature alludes to the high likelihood of efficacy of MBIs in enhancing empathy and ToM domains for various populations.

However, most of these MBIs, such as the popular 8-week MBCT and MBSR programs, necessitated in-person weekly individual or group intervention sessions lasting 60 to 150 minutes with day-long, 6-hour meditation retreats [22]. Thus, evaluating the impact of brief MBIs is crucial in light of the growing popularity of concise MBIs, such as brief smartphone-delivered mindfulness ecological momentary interventions (EMIs) and web-based audio streams [23]. Scant yet positive evidence for the possibility of brief mindfulness EMIs' efficacy on empathy-related outcomes exists across 2 studies. First, a 5-minute mindfulness induction enhanced both ToM and empathic concern among meditation-naive adults more than among control adults [24]. Second, an 11-minute MBI versus relaxation led to enhanced out-intervention altruism while adjusting for levels of in-intervention empathy in the US general population [25]. Despite that, to the best of our awareness, there is no evidence yet that these effects endured beyond these mindfulness inductions, emphasizing the need to assess the efficacy of brief mindfulness EMIs on longer-term outcomes.

In addition, a significant limitation in the current literature on the efficacy of MBIs on social-cognitive outcomes is the focus on healthy samples despite ample evidence demonstrating empathy and ToM problems in populations with psychiatric disorders, such as depression [26], eating disorders [27], and anxiety disorders [28]. In recent years, there has been growing recognition of dysregulation in social processes, particularly cognitive and emotional empathy, as significant transdiagnostic contributors to internalizing disorders such as major depressive disorder and generalized anxiety disorder (GAD) in terms of their etiology, diagnostic relevance, and maintenance [29,30]. A meta-analysis showed that increased anxiety symptoms (including excessive worry in GAD) correlated with reduced *perspective-taking* and heightened the ability to be deeply immersed in imaginative fictional worlds (or *fantasy*) [31]. In addition, higher anxiety symptoms were associated with less pronounced *empathic concern* for others' emotional well-being and others' *personal distress*, the inclination to absorb or experience stress in reaction to others' feelings. These relationships imply that suboptimal empathy and ToM processes could be a common factor across various forms of anxiety disorders. Nonetheless, basic science [31,32] and clinical research [33] attention on this topic has been more heavily weighted toward social anxiety disorder, obsessive-compulsive disorder [34], and posttraumatic stress disorder [35] than other anxiety disorders, such as GAD. This lacuna in the literature highlights the importance of investigating this topic in other anxiety disorders to inform novel treatment optimization efforts for GAD, a needed avenue that lacks exploration [36].

On that note, GAD, for which chronic excessive and uncontrollable worry features as a core symptom, is an understudied yet essential case in point. More specifically, individuals with GAD tended to have above-average ToM scores, but only for negative social cues when instructed to worry instead of relax [37]. In another study, those with GAD scored lower on a ToM assessment than non-GAD control intervention participants [38]. Furthermore, people with clinical depression and anxiety experienced increased worry symptoms on days when they pursued self-image goals above the sample average but encountered reduced worry symptoms when they pursued empathic, compassionate goals [39]. Broadly, persistent interpersonal challenges, including problems associated with empathy and social cognition, among a sizable subset of clients with GAD were linked to reduced progress both immediately after cognitive behavioral therapy (CBT) and during follow-up [40,41]. Altogether, these findings suggest that empathy and associated social-cognitive issues may be key maintenance factors in GAD. For these reasons, intensive psychotherapies integrating various theoretical modalities have been developed to enhance empathy in GAD [42], but their lengthiness and rigor preclude scalability. Collectively, suboptimal empathy, ToM, and interpersonal issues [42], coupled with a considerable reluctance to seek face-to-face mental health treatment in GAD [43], highlight the importance of determining whether brief mindfulness EMIs might be efficacious in targeting these social-cognitive outcomes for GAD.

This study was a secondary analysis of a published RCT, which showed that a brief mindfulness EMI reduced repetitive negative thinking and GAD severity and enhanced trait mindfulness and executive functioning among persons diagnosed with GAD [44]. Similar to this study, in prior research MBIs produced modest to moderate impacts on anxiety and depression symptoms [45], such as pathological worry [44,46-48]. In addition, a meta-analysis documented the substantial impact of MBIs on executive functioning in both nonclinical and clinical populations [49], such as persons with GAD [44]. Therefore, it is possible that a brief mindfulness EMI could help enhance ToM and empathy in those with GAD.

Objective

We aimed to examine how a brief mindfulness EMI, compared to a self-monitoring app, might improve empathy and ToM in GAD. On the basis of the theories and evidence outlined, we hypothesized that a brief mindfulness EMI would be more efficacious than a self-monitoring app in improving 4 established empathy domains [50] and ToM [51] from pre-post time points. The empathy domains included *empathic concern* (feeling affection, compassion, and genuine care when observing others' distress, with attentiveness centered on their emotional well-being); *fantasy* (the capacity for immersive engagement in fictional scenarios through imagination); *personal distress* (the anxiety and stress individuals experience when observing the adversity of others); and *perspective-taking* (understanding others' vantage point) [52]. Moreover, we expected such improvements to occur in the longer term from prerandomization to 1-month follow-up (1MFU) time points.

Methods

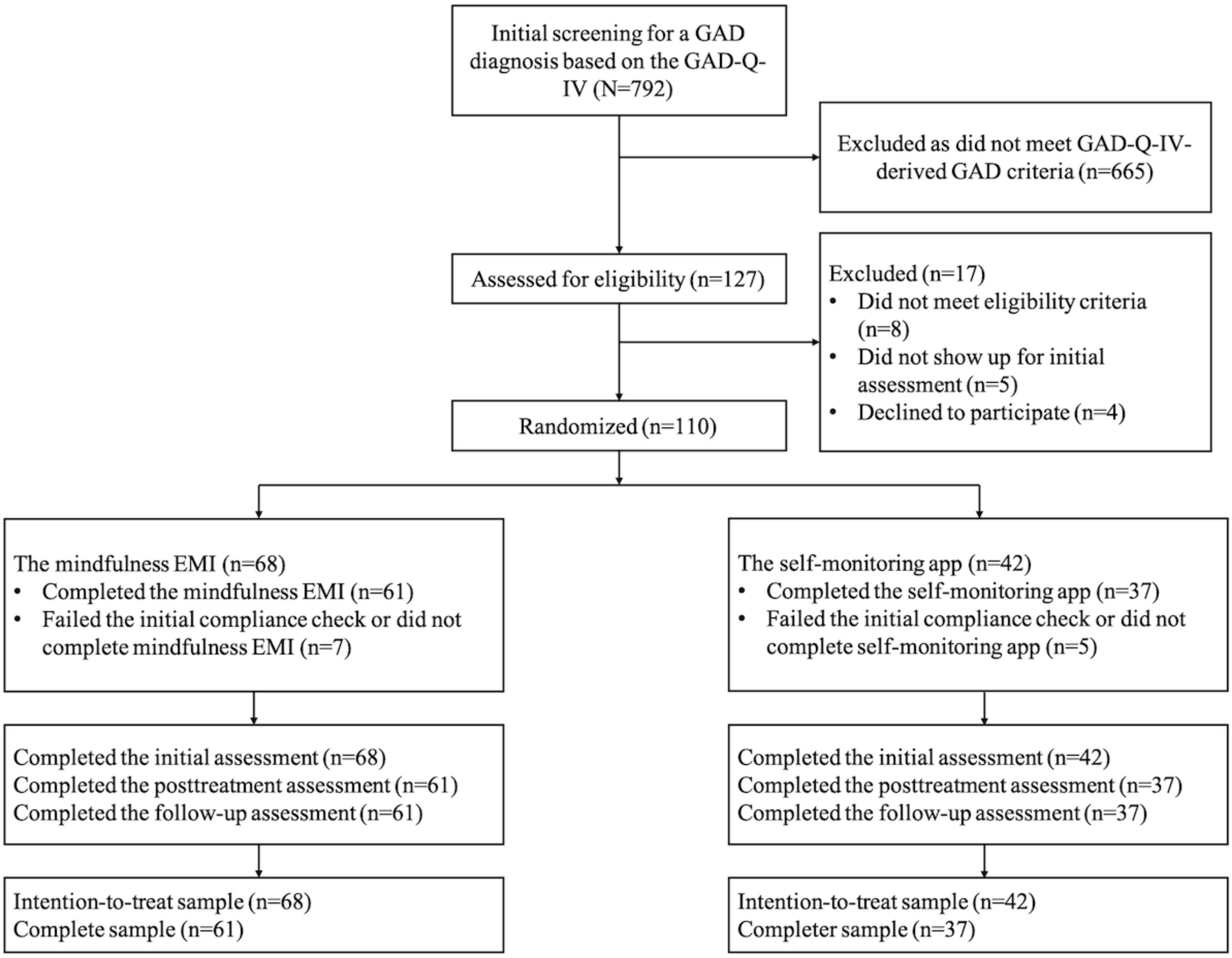
Ethical Considerations

This study was approved by the Pennsylvania State University Institutional Review Board (approval STUDY00010664). Participants were compensated up to US $30, subject pool credit hours, or a mixture of both. We preregistered the RCT on ClinicalTrials.gov (NCT04846777) and the hypotheses of this study on the Open Science Framework [53]. This study was conducted in compliance with the American Psychological Association and Declaration of Helsinki ethical standards in treating human participants. Informed consent was obtained from participants as per the Penn State Institutional Review Board. All data were de-identified.

Study Design

We used a 2 (intervention: mindfulness EMI; control: self-monitoring app) × 3 (time: prerandomization, postintervention, and 1MFU) mixed methods design to assess the efficacy of the mindfulness EMI versus self-monitoring app on each distinct empathy and ToM outcome. Intervention was the between-subject factor, whereas time was the within-subject factor. We recruited 110 participants, with 61.8% (68/110) in the mindfulness EMI arm and 38.2% (42/110) in the self-monitoring app arm. Multimedia Appendix 1 [54-66] provides a comprehensive overview of the study's methodology, including power analysis and reimbursement. Figure 1 displays the CONSORT (Consolidated Standards of Reporting Trials) diagram [67,68], illustrating participant enrollment and progression (refer to Multimedia Appendix 2 for the CONSORT-EHEALTH [Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth] checklist).

Figure 1. CONSORT (Consolidated Standards of Reporting Trials) flowchart of participant recruitment and progress. GAD: generalized anxiety disorder; GAD-Q-IV: Generalized Anxiety Disorder Questionnaire, Fourth Version; EMI: ecological momentary intervention.



Eligibility Criteria

Included participants were required to meet diagnostic criteria for GAD according to the *Diagnostic and Statistical Manual, Fifth Edition* (*DSM-5*) [54] and the GAD Questionnaire, Fourth Version (GAD-Q-IV) [55]. Initially, prospective participants underwent screening with the GAD-Q-IV. Subsequently, we invited those who scored at or above the clinical threshold for a brief clinical interview. The Anxiety and Related Disorders Interview Schedule (ADIS) for *DSM-5* [54,56] was used to establish their psychiatric diagnoses. Furthermore, participants were required to be aged at least 18 years, possess either an iPhone or Android phone, and provide informed consent. Exclusion criteria encompassed the presence of suicidal thoughts, manic episodes, disorders of psychosis, or substance use disorders.

Participants

We recruited help-seeking participants diagnosed with GAD, who were not currently under the care of a mental health professional, from the local community and psychology subject pool. Table 1 displays the sociodemographic characteristics of the participants. Furthermore, the prevalence of comorbid diagnoses (eg, current or recurrent major depressive disorder, panic disorder, social anxiety disorder, obsessive compulsive disorder, posttraumatic stress disorder, alcohol use disorder, substance use disorder, anorexia nervosa, and binge-eating disorder) at baseline did not significantly differ between arms (all P>.05).

Table 1. Sociodemographic data of study participants in the mindfulness EMI[a] and self-monitoring app arms (N=110).

Sociodemographic characteristics	Mindfulness EMI arm (n=68)	Self-monitoring app arm (n=42)	*P* value
Continuous variables, mean (SD)			
Age (y)	20.53 (3.91)	21.24 (7.24)	.51
14-item GAD-Q-IV[b] score	9.52 (2.10)	9.94 (1.96)	.30
Treatment expectations			
Credibility	6.00 (1.39)	5.72 (1.58)	.34
Expectancy	43.46 (17.33)	44.29 (18.13)	.31
Categorical variables, n (%)			
Gender orientation			.85
Women	10 (14.7)	5 (11.9)	
Men	57 (83.8)	37 (88.1)	
Declined to disclose	1 (1.5)	N/A[c]	
Race			.99
African American	5 (7.4)	1 (2.4)	
Asian or Asian American	11 (16.2)	4 (9.5)	
Declined to disclose	1 (1.5)	0 (0)	
Hispanic	3 (4.4)	5 (11.9)	
Other race	4 (5.9)	2 (4.8)	
White	44 (64.7)	27 (64.3)	
Comorbid diagnoses			
Current major depressive episode	32 (47.1)	24 (57.1)	.30
Recurrent major depressive episode	25 (36.8)	20 (47.6)	.26
Current panic disorder	16 (23.5)	5 (11.9)	.13
Current social anxiety disorder	15 (22.1)	14 (33.3)	.19
Current OCD[d]	4 (5.9)	4 (9.5)	.48
Current PTSD[e]	9 (13.2)	4 (9.5)	.56
Current alcohol use disorder	7 (10.3)	1 (2.4)	.12
Current substance use disorder	3 (4.4)	1 (2.4)	.58
Current anorexia nervosa	0 (0)	0 (0)	N/A
Current binge-eating disorder	1 (1.5)	0 (0)	.39

[a]EMI: ecological momentary intervention.

[b]GAD-Q-IV: Generalized Anxiety Disorder Questionnaire, Fourth Edition.

[c]N/A: not applicable.

[d]OCD: obsessive compulsive disorder.

[e]PTSD: posttraumatic stress disorder.

Prerandomization Diagnostic Interview and Screening Assessment

Mental Disorder Diagnoses

The ADIS-5 [56] was used as a semistructured interview based on the *DSM-5* criteria [69]. Every ADIS-5 interview, whether face-to-face or via Zoom (Zoom Video Communications), was video-recorded and meticulously conducted by highly trained undergraduate and Bachelor of Arts–level assessors. All assessors had watched standardized training videos and completed quality assurance tests to maximize implementation fidelity to the study protocol. The study protocol was implemented remotely via Zoom during the COVID-19 pandemic. A subset comprising 40% (45/110) of these video recordings underwent a secondary evaluation by an independent, unbiased rater. We assessed the interrater reliability for GAD diagnoses. Any discrepancies were addressed through discussions and eventual consensus. The interrater agreement for GAD diagnosis was outstanding (Cohen κ=1.00), whereas, for other comorbid diagnoses and rule-outs, it ranged from very good to outstanding (average Cohen κ=0.75-0.98).

GAD Diagnosis

The screening for GAD used the 14-item GAD-Q-IV [55], which included a combination of dichotomous responses ("yes" or "no") and continuous response formats (9-point Likert scales for items assessing the level of distress and interference attributed to GAD symptoms).

Intervention Arms

Mindfulness EMI App

During the first visit, assessors either exited the physical room (before the pandemic) or directed participants to turn off their Zoom audio and video before accessing the Qualtrics (Qualtrics International Inc) link to play the relevant intervention video (during the pandemic). In the mindfulness EMI, an instructional video featuring the principal investigator (a PhD-level clinical psychologist) was presented, delivering critical principles of evidence-based MBI protocols akin to the principles in MBSR [70]. Participants assigned to this condition were introduced to a precise conceptualization of mindfulness, with an explicit directive to immerse themselves entirely in their present circumstances and be engaged in activities. This segment was designed to endow individuals with chronic worry tendencies with proficiency in open monitoring, facilitating their capacity to attend to minute details. Following this, the video therapist taught techniques for deliberate, unhurried, rhythmic diaphragmatic breathing retraining, followed by a demonstration of its proper implementation. This element encompassed instruction in techniques for cultivating tranquility through controlled respiratory training and the cultivation of mindful qualities, such as observance without reactivity or judgment, drawing from the principles of MBCT [15]. Next, the video therapist underscored the significance and advantages of integrating mindfulness into daily routines. Subsequently, all assessors uniformly administered the 6-item Credibility and Expectancy Questionnaire [57]. Concluding this phase, assessors attentively addressed any inquiries about procedural but not the intervention aspects of the study (Multimedia Appendix 3). Participants were given a mindfulness EMI handout sent web-based in an automated manner via Qualtrics to preserve assessor blinding. The handout explicitly instructed them to review and practice its contents regularly.

The EMI encouraged individuals to engage in mindfulness exercises daily, precisely at 5 distinct intervals throughout the day: approximately 9 AM, noon, 3 PM, 6 PM, and 9 PM, over a span of 2 weeks. Within every instance of engagement with the mindfulness EMI, individuals participating in the program were provided with the following standard guidance instructions (Multimedia Appendices 1 and 3):

> *Pay attention to your breathing. Breathe in a slow, steady, and rhythmic manner. Stay focused on the sensations of the air coming into your lungs and then letting it out. As you are breathing, observe your experience as it is. Let go of judgments that do not serve you. Focus on the here and now. Attend to the small moments right now (e.g., reading a chapter, having a cool glass of water), as that is where enjoyment, peace, and serenity in life happen.*

Individuals assessed their current levels of mindfulness ("To what extent are you experiencing the present moment fully?"), depression ("To what degree do you feel depressed right now?"), and anxiety ("To what degree do you feel keyed up or on edge right now?") before and after receiving these instructions on a 9-point Likert scale, ranging from 1 (*not at all*) to 9 (*extremely*). Every mindfulness EMI notification concluded with the following motivating message to foster the enduring adoption of these skills:

> *Remember that the cultivation of mindfulness is lifelong. The goal of therapy is to be your own therapist. Practice mindfulness between the prompts and after you have completed this study.*

Self-Monitoring App

In the self-monitoring app, the standardized video commenced with the principal investigator delineating self-monitoring as the heightened awareness of cognitive processes and emotional states. Subsequently, the video advanced by positing that the mere act of vigilantly tracking one's thoughts and documenting any associated emotional distress could promote the cultivation of more adaptive cognitive patterns. Finally, the self-monitoring video conveyed the notion that the act of self-monitoring, in and of itself, possessed the potential to ameliorate feelings of anxiety. The foundational rationale for the self-monitoring condition was derived and adapted from the rationale used in a recent, brief app intervention [71]. This approach was crafted to closely mimic the mindfulness EMI protocol while excluding its hypothesized active, helpful components: acceptance, diaphragmatic breathing retraining, awareness toward subtle experiences, open monitoring, and sustained mindfulness practice. Consequently, the self-monitoring app deliberately abstained from any reference to the concept of mindfulness. This approach refrained from imparting explicit directives for participants to heighten their sensitivity and consciousness of their ongoing experiences; instead, its emphasis rested on vigilant monitoring of their thoughts and emotional responses. Moreover, participants were not tasked with focusing exclusively on their immediate activities because such a directive could inadvertently induce emotional state alterations. Although self-monitoring participants were directed to observe their cognitions and emotional states, there was a deliberate omission of instructions about accepting these thoughts and feelings as they manifested. Furthermore, the intervention did not include any guidance regarding breathing retraining techniques. It was not designed to elicit any sensations associated with relaxation typically linked to abdominal breathing. Participants were not instructed to engage in self-monitoring activities between the prescribed prompts or beyond the conclusion of the initial 2-week period. Consequently, self-monitoring, unlike mindfulness, was delimited to the active intervention period. The self-monitoring approach was also designed to control for credibility and expectancy effects, pre-empting regression to the mean and averting potential inflation of effect sizes that typically happen with a no-treatment or waitlist control (Multimedia Appendix 4).

In contrast to the more extensive mindfulness guidance provided by the mindfulness EMI, participants in the self-monitoring intervention received concise instructions 5 times a day (approximately 9 AM, noon, 3 PM, 6 PM, and 9 PM) over a span of 14 days (Multimedia Appendices 1 and 4): "Notice your thoughts and how distressing they may be." We assessed individuals' levels of mindfulness, depression, and anxiety using identical 9-point Likert scale questions both before and after each prompt in the self-monitoring sessions. Next, mirroring the mindfulness EMI, all assessors administered the 6-item Credibility and Expectancy Questionnaire. Furthermore, participants were provided a copy of the self-monitoring handout in an automated manner programmed via Qualtrics. Unlike the mindfulness EMI handout, this handout did not provide explicit instructions to review its contents routinely. In addition, the self-monitoring app was chosen as a placebo control as previous theory and research suggested that it would not be strong enough to elicit improvements in empathy and related social cognition relative to an MBI [72] yet could serve adequately as a placebo in an RCT [73].

Prerandomization, Postintervention, and 1MFU Measures

Overview

Table 2 summarizes descriptive statistics of the empathy domains and ToM scores.

Table 2. Descriptive data of empathy and ToM[a] variables across prerandomization, postintervention, and 1MFU[b] time points in the mindfulness EMI[c] (n=68) and self-monitoring (n=42) app arms (N=110).

	Values, mean (SE)[d]	Skewness	Kurtosis
Mindfulness EMI at the prerandomization time point			
ToM (BLERT[e])	17.85 (0.78)	−1.05	0.12
Perspective taking (IRI[f])	3.66 (0.28)	−0.22	−0.77
Fantasy (IRI)	3.53 (0.29)	−0.27	−1.05
Empathic concern (IRI)	3.97 (0.26)	−0.96	0.72
Personal distress (IRI)	3.13 (0.27)	−0.16	−0.88
S elf-monitoring at the prerandomization time point			
ToM (BLERT)	17.81 (0.34)	−0.07	−1.23
Perspective taking (IRI)	3.62 (0.12)	−0.11	−1.25
Fantasy (IRI)	3.71 (0.13)	−0.38	−0.04
Empathic concern (IRI)	3.93 (0.12)	−0.35	−0.69
Personal distress (IRI)	3.08 (0.12)	0.17	−0.79
Mindfulness EMI at the postintervention time point			
ToM (BLERT)	17.94 (1.66)	−1.12	2.46
Perspective taking (IRI)	3.68 (0.66)	−0.23	−0.94
Fantasy (IRI)	3.87 (0.68)	−0.65	−0.83
Empathic concern (IRI)	4.10 (0.60)	−1.31	1.71
Personal distress (IRI)	3.21 (0.63)	−0.10	−1.10
Self-monitoring at the postintervention time point			
ToM (BLERT)	17.95 (0.73)	−1.57	5.33
Perspective taking (IRI)	3.60 (0.29)	0.14	−0.81
Fantasy (IRI)	3.45 (0.30)	−0.57	−0.31
Empathic concern (IRI)	3.74 (0.26)	−0.46	−0.41
Personal distress (IRI)	2.80 (0.28)	0.01	−0.13
Mindfulness EMI at the 1MFU time point			
ToM (BLERT)	19.13 (1.54)	−1.12	2.46
Perspective taking (IRI)	3.90 (0.56)	−0.23	−0.94
Fantasy (IRI)	3.85 (0.57)	−0.65	−0.83
Empathic concern (IRI)	4.11 (0.51)	−1.31	1.71
Personal distress (IRI)	3.38 (0.57)	−0.10	−1.10
Self-monitoring at the 1MFU time point			
ToM (BLERT)	17.98 (0.68)	−0.04	−1.05
Perspective taking (IRI)	3.62 (0.25)	−0.31	−0.77
Fantasy (IRI)	3.57 (0.25)	−0.09	−0.43
Empathic concern (IRI)	3.80 (0.22)	−0.57	−0.21
Personal distress (IRI)	2.71 (0.25)	−0.11	−0.29

[a]ToM: theory-of-mind.

[b]1MFU: 1-month follow-up time point.

[c]EMI: ecological momentary intervention.

[d]These marginal means and SEs were derived from the hierarchical linear models.

[e]BLERT: Bell-Lysaker Emotion Recognition Task.

[f]IRI: Interpersonal Reactivity Index.

Empathy Domains

We used the 28-item Interpersonal Reactivity Index (IRI) [50] to assess 4 trait-level empathy domains, each with a subscale [74]. *Empathic concern* refers to experiencing affection and compassion when witnessing others' distress, focusing on the well-being of others' feelings (eg, "I am often quite touched by things that I see happen."). *Fantasy* pertains to peoples' imaginative ability to immerse themselves in fictional scenarios (eg, "I really get involved with the feelings of the characters in a novel."). *Personal distress* refers to individuals' feelings of anxiety and stress when witnessing others' negative experiences (eg, "In emergency situations, I feel apprehensive and ill-at-ease."). *Perspective-taking* entails comprehending others' viewpoint (eg, "I sometimes try to understand my friends better by imagining how things look from their perspective."). Each subscale comprised 4 items rated on a 5-point scale ranging from 1 (*never*) to 5 (*always*). Possible values of each subscale range from 1 to 5. The internal consistency (Cronbach α) of the IRI empathic concern scale was .85, .88, and .89 at prerandomization, postintervention, and 1MFU time points (fantasy: .82, .89, and .90; personal distress: .81, .90, and .89; and perspective tasking: .85, .88, and .89), respectively. Each IRI subscale has shown high convergent validity, acceptable discriminant validity [75], strong retest reliability, and good cross-cultural generalizability [76].

ToM Domains

We evaluated ToM using the Bell-Lysaker Emotion Recognition Task (BLERT) [51]. BLERT comprises 21 brief video clips where an actor enacts one of 3 dialogue options while expressing 7 distinct emotions (anger, disgust, fear, happiness, sadness, surprise, and no emotion). Participants selected the emotion that best corresponded to the emotion expressed by the actor from the 7 options displayed. An overall accuracy score was calculated (based on a score of 1 for each correct trial). Possible values range from 0 to 21. Cronbach α values of the BLERT total scale were .94, .89, and .95 at prerandomization, postintervention, and 1MFU time points, respectively. Furthermore, a prior review showed that the BLERT had high retest reliability, strong convergent validity, and good discriminant validity [77].

Procedures

At visit 1, participants initially underwent the structured ADIS-5 interview. Afterward, eligible individuals completed a battery of prerandomization self-reports and behavioral and neuropsychological assessments in a manner that was counterbalanced, thus minimizing any potential for order-related biases. Assessors were uninformed regarding assigned arms, ensuring that the treatment assignment remained concealed from them throughout all study visits. They either left the physical room (before the COVID-19 pandemic) or provided clear instructions to participants to deactivate their Zoom audio and video before activating the Qualtrics link to play the pertinent treatment video (amid the COVID-19 pandemic). Participants installed the Personal Analytics Companion app, preloaded with either the mindfulness EMI or self-monitoring tool, onto their respective smartphones. The assessor was present if participants had questions about study procedures (eg, future study visits and technical questions about installing the app on their phone) but was not present when the participants were notified about their assigned intervention arm and its components. Participants were told that they would receive prompts at 5 designated times each day, at approximately 9 AM, noon, 3 PM, 6 PM, and 9 PM, over the ensuing 14 days, and these prompts could be flexibly adjusted based on their schedules. The prompts provided specific instructions to guide them to use mindfulness or self-monitoring strategies, contingent upon their assigned intervention. After 14 days, all participants reconvened at the laboratory (or on Zoom) for postintervention assessments and again during the 1MFU time point, during which they completed the requisite self-reports and neuropsychological battery. Participants received compensation in credit hours, monetary remuneration, or a combination of both (Multimedia Appendix 1). In addition, the research team performed a compliance verification on the seventh day and extended invitations to those participants who successfully passed this compliance assessment to continue in the treatment.

Data Analyses

All analyses were based on the intent-to-treat principle. Of the 110 participants assigned to the mindfulness EMI (n=68, 61.8%) or self-monitoring app (n=42, 38.2%) arm, 98 (89.1%) participants conscientiously finished engaging with the 6-week study protocol, including all study visit assessments and ≥ 80% of the app prompts. A total of 12 (10.7%) out of 110 participants did not meet the seventh-day engagement threshold by failing to complete at least 80% of the EMI prompts from both arms over the 2-week intervention period. However, the data of these 12 (10.7%) participants were included in all analyses. The median number of prompts completed was 63 (range 0-70). To address missing values (29/1650, 1.76% of the entire data set), we conducted multiple imputation using chained equations with the predictive mean matching algorithm under the missing at-random assumption [78]. Specifically, we synthesized the data from 100 imputed data sets, each with a maximum of 10 iterations [79]. As another preprocessing step, we determined from a series of ordinary least square regressions that there were no significant between-intervention effects on BLERT ToM (β=.04, *P*=.92), IRI empathic concern (β=.05, *P*=.74), IRI fantasy (β=−.18, *P*=.26), IRI personal distress (β=.05, *P*=.74), and IRI perspective-taking (β=.04, *P*=.80) domains at prerandomization (baseline) time point. Furthermore, there were no significant differences in perceived credibility or expectancy between the interventions, with Cohen *d* values ranging from −0.05 to 0.19. Therefore, no covariates were included in the hierarchical linear modeling (HLM) predictive equations, given randomization success (Table 1).

HLM was used to compare interventions over time on all empathy-related outcomes. HLM considers the nesting of time points within participants, enabling the exploration of changes within and between participants over time (prerandomization, postintervention, and 1MFU time points) and by intervention (mindfulness EMI and self-monitoring app) [80,81]. Distinct models were used for each of the 5 outcomes. To assess efficacy of the mindfulness EMI versus self-monitoring app, we analyzed the time at level 1, delineating the first segment from pre-post intervention time points and a second segment from prerandomization to 1MFU (pre-1MFU) time points. The third segment from postintervention to 1MFU (post-1MFU) time points examined maintenance of any treatment gains. In each HLM model, fixed-effect predictors included intervention, time, and their interaction. The random-effect predictor was the intercept, which was coded for time (eg, prerandomization time point as "0" and postintervention time point as "1"), allowing for variability in average outcome values among participants. We used fitted models to calculate estimated mean scores at each time point. Cohen *d* effect sizes and their corresponding 95% CIs were computed to facilitate the interpretation of parameter estimates [82-84]. In this context, Cohen *d* values of 0.2, 0.5, and 0.8 represent small, moderate, and large effects, respectively.

Results

ToM: Propositional Knowledge of Others' Emotional and Mental States

Despite nonsignificant intervention differences from pre-post intervention time points (Cohen *d*=−0.01, 95% CI −0.21 to 0.19; *P*=.91), the mindfulness EMI (vs the self-monitoring app) led to significantly greater change in ToM (BLERT scores) from pre-1MFU (Cohen *d*=0.25, 95% CI 0.05-0.45; *P*=.01; Table 3). From pre-post intervention, no significant changes in ToM occurred in either the mindfulness EMI (Cohen *d*=0.03, 95% CI −0.17 to 0.23; *P*=.75) or self-monitoring (Cohen *d*=0.03, 95% CI −0.16 to 0.23; *P*=.74) interventions. From pre-1MFU, the mindfulness EMI (Cohen *d*=0.43, 95% CI 0.23-0.63; *P*<.001) but not the self-monitoring app (Cohen *d*=0.06, 95% CI −0.14 to 0.25; *P*=.56) significantly enhanced ToM (Table 4). No substantial between-intervention and within-intervention effects on ToM emerged from post-1MFU time points (Tables S1 and S2 in Multimedia Appendix 5).

Table 3. Hierarchical linear models of between-intervention mindfulness ecological momentary intervention versus self-monitoring app effects on empathy and ToM[a] variables from prerandomization to postintervention (pre-post intervention) and prerandomization to 1-month follow-up (pre-1MFU) time points.

	Pre-post intervention time points			Pre-1MFU time points		
	β[b]	*P* value	Cohen *d* (95% CI)	β	*P* value	Cohen *d* (95% CI)
ToM (BLERT[c])						
Intercept	17.81	<.001	4.94 (4.56 to 5.33)	17.81	<.001	4.94 (4.56 to 5.33)
Time	0.14	.71	0.04 (−0.16 to 0.23)	0.08	.62	0.05 (−0.15 to 0.24)
Intervention	0.04	.92	0.01 (−0.19 to 0.21)	0.04	.92	0.01 (−0.19 to 0.21)
Time × intervention	−0.05	.91	−0.01 (−0.21 to 0.19)	0.56	.01	0.25 (0.05 to 0.45)
Empathic concern (IRI[d])						
Intercept	3.93	<.001	3.24 (2.95 to 3.53)	3.93	<.001	3.24 (2.95 to 3.53)
Time	−0.18	.23	−0.12 (−0.31 to 0.08)	−0.06	.25	−0.11 (−0.31 to 0.09)
Intervention	0.05	.74	0.03 (−0.16 to 0.23)	0.05	.74	0.03 (−0.16 to 0.23)
Time × intervention	0.30	.11	0.15 (−0.04 to 0.35)	0.13	.06	0.18 (−0.02 to 0.38)
Fantasy (IRI)						
Intercept	3.71	<.001	2.78 (2.51 to 3.05)	3.71	<.001	2.78 (2.51 to 3.05)
Time	−0.26	.13	−0.14 (−0.34 to 0.05)	−0.07	.25	−0.11 (−0.31 to 0.09)
Intervention	−0.18	.26	−0.11 (−0.3 to 0.09)	−0.18	.26	−0.11 (−0.3 to 0.09)
Time × intervention	0.60	.007	0.26 (0.06 to 0.46)	0.23	.004	0.28 (0.09 to 0.48)
Personal distress (IRI)						
Intercept	3.08	<.001	2.46 (2.20 to 2.71)	3.08	<.001	2.46 (2.20 to 2.71)
Time	−0.28	.08	−0.17 (−0.37 to 0.03)	−0.18	.005	−0.27 (−0.47 to −0.08)
Intervention	0.05	.74	0.03 (−0.16 to 0.23)	0.05	.74	0.03 (−0.16 to 0.23)
Time × intervention	0.36	.083	0.17 (−0.03 to 0.36)	0.31	<.001	0.36 (0.16 to 0.56)
Perspective taking (IRI)						
Intercept	3.62	<.001	2.77 (2.50 to 3.04)	3.62	<.001	2.77 (2.50 to 3.04)
Time	−0.02	.89	−0.01 (−0.21 to 0.18)	0.00	.99	0.00 (−0.20 to 0.20)
Intervention	0.04	.80	0.02 (−0.17 to 0.22)	0.04	.80	0.02 (−0.17 to 0.22)
Time × intervention	0.04	.84	0.02 (−0.18 to 0.22)	0.12	.13	0.15 (−0.05 to 0.34)

[a]ToM: theory-of-mind.

[b]β: regression unstandardized parameter estimate.

[c]BLERT: Bell-Lysaker Emotion Recognition Task.

[d]IRI: Interpersonal Reactivity Index.

Table 4. Hierarchical linear models of within-intervention mindfulness EMI[a] and self-monitoring app effects on empathy and ToM[b] variables from prerandomization to postintervention (pre-post intervention) and prerandomization to 1-month follow-up (pre-1MFU) time points.

	Pre-post intervention time points			Pre-1MFU time points		
	β[c]	*P* value	Cohen *d* (95% CI)	β	*P* value	Cohen *d* (95% CI)
ToM (BLERT[d])						
Intercept (mindfulness EMI)	17.85	<.001	5.50 (5.08 to 5.92)	17.85	<.001	5.50 (5.08 to 5.92)
Time (mindfulness EMI)	0.09	.75	0.03 (−0.17 to 0.23)	0.64	<.001	0.43 (0.23 to 0.63)
Intercept (self-monitoring)	17.81	<.001	7.10 (6.58 to 7.62)	17.81	<.001	7.10 (6.58 to 7.62)
Time (self-monitoring)	0.14	.74	0.03 (−0.16 to 0.23)	0.08	.56	0.06 (−0.14 to 0.25)
Empathic concern (IRI[e])						
Intercept (mindfulness EMI)	3.97	<.001	4.29 (3.94 to 4.64)	3.97	<.001	4.29 (3.94 to 4.64)
Time (mindfulness EMI)	0.12	.26	0.11 (−0.09 to 0.31)	0.07	.06	0.18 (−0.01 to 0.38)
Intercept (self-monitoring)	3.93	<.001	3.11 (2.82 to 3.40)	3.93	<.001	3.11 (2.82 to 3.40)
Time (self-monitoring)	−0.18	.28	−0.11 (−0.30 to 0.09)	−0.06	.35	−0.09 (−0.29 to 0.11)
Fantasy (IRI)						
Intercept (mindfulness EMI)	3.53	<.001	3.08 (2.80 to 3.37)	3.53	<.001	3.08 (2.80 to 3.37)
Time (mindfulness EMI)	0.34	.02	0.22 (0.03 to 0.42)	0.16	.002	0.30 (0.11 to 0.5)
Intercept (self-monitoring)	3.71	<.001	3.36 (3.06 to 3.66)	3.71	<.001	3.36 (3.06 to 3.66)
Time (self-monitoring)	−0.26	.10	−0.16 (−0.36 to 0.03)	−0.07	.22	−0.12 (−0.31 to 0.08)
Personal distress (IRI)						
Intercept (mindfulness EMI)	3.13	<.001	3.00 (2.72 to 3.28)	3.13	<.001	3.00 (2.72 to 3.28)
Time (mindfulness EMI)	0.07	.60	0.05 (−0.15 to 0.25)	0.12	.03	0.21 (0.02 to 0.41)
Intercept (self-monitoring)	3.08	<.001	2.75 (2.48 to 3.02)	3.08	<.001	2.75 (2.48 to 3.02)
Time (self-monitoring)	−0.28	.03	−0.21 (−0.41 to −0.02)	−0.18	.001	−0.33 (−0.53 to −0.13)
Perspective taking (IRI)						
Intercept (mindfulness EMI)	3.66	<.001	3.43 (3.13 to 3.74)	3.66	<.001	3.43 (3.13 to 3.74)
Time (mindfulness EMI)	0.02	.89	0.01 (−0.18 to 0.21)	0.12	.01	0.25 (0.06 to 0.45)
Intercept (self-monitoring)	3.62	<.001	2.95 (2.67 to 3.23)	3.62	<.001	2.95 (2.67 to 3.23)
Time (self-monitoring)	−0.02	.89	−0.01 (−0.21 to 0.18)	0.00	.99	0.00 (−0.20 to 0.20)

[a]EMI: ecological momentary intervention.

[b]ToM: theory-of-mind.

[c]β: unstandardized regression parameter estimate.

[d]BLERT: Bell-Lysaker Emotion Recognition Task.

[e]IRI: Interpersonal Reactivity Index.

Empathic Concern: Care About Others' Psychological Well-Being

There was no significant difference between the mindfulness EMI and self-monitoring interventions in effects on empathic concern (IRI scores) from pre-post intervention (Cohen *d*=0.15, 95% CI −0.04 to 0.35; *P*=.11) and pre-1MFU (Cohen *d*=0.18, 95% CI −0.02 to 0.38; *P*=.06; Table 3) time points. From pre-post intervention, there were no significant changes in empathic concern in either the mindfulness EMI (Cohen *d*=0.11, 95% CI −0.09 to 0.31; *P*=.26) or self-monitoring (Cohen *d*=0.11, 95% CI −0.30 to 0.09; *P*=.28) interventions. Similarly, from pre-1MFU, no significant changes in empathic concern emerged in the mindfulness EMI (Cohen *d*=0.18, 95% CI −0.01 to 0.38; *P*=.06) or self-monitoring (Cohen *d*=−0.09, 95% CI −0.29 to 0.11; *P*=.35; Table 4) interventions. No significant between-intervention and within-intervention effects on empathic concern emerged from post-1MFU time points (Tables S1 and S2 in Multimedia Appendix 5).

Fantasy: the Ability to Imagine Others' Experiences

The mindfulness EMI (vs self-monitoring app) led to greater effects on fantasy (IRI scores) from pre-post intervention (Cohen *d*=0.26, 95% CI 0.06-0.46; *P*=.007) and pre-1MFU (Cohen *d*=0.28, 95% CI 0.09-0.48; *P*=.004; Table 3). From pre-post intervention, the mindfulness EMI (Cohen *d*=0.22, 95% CI

0.03-0.42; *P*=.02) but not the self-monitoring app (Cohen *d*=−0.16, 95% CI −0.36 to 0.03; *P*=.10) generated significant improvement in fantasy. Similarly, from pre-1MFU, the mindfulness EMI (Cohen *d*=0.30, 95% CI 0.11-0.50; *P*=.002) but not the self-monitoring app (Cohen *d*=−0.12, 95% CI −0.31 to 0.08; *P*=.22) significantly enhanced fantasy (Table 4). Maintenance of gains occurred from post-1MFU time points (Tables S1 and S2 in Multimedia Appendix 5).

Personal Distress: Feeling Distress When Observing Others' Adverse Experiences

There were no significant intervention differences In personal distress (IRI scores) from pre-post intervention (Cohen *d*=0.17, 95% CI −0.03 to 0.36; *P*=.08). At the same time, from pre-post intervention, the self-monitoring app (Cohen *d*=−0.21, 95% CI −0.41 to −0.02; *P*=.03) but not the mindfulness EMI (Cohen *d*=0.05, 95% CI −0.15 to 0.25; *P*=.60) significantly reduced personal distress. There were significant differences between mindfulness EMI and self-monitoring interventions from pre-1MFU time points (Cohen *d*=0.36, 95% CI 0.16-0.56; *P*<.001; Table 3). Although the mindfulness EMI significantly increased personal distress (Cohen *d*=0.21, 95% CI 0.02-0.41; *P*=.03), the self-monitoring app significantly decreased it (Cohen *d*=−0.33, 95% CI −0.53 to −0.13; *P*=.001; Table 4). No significant between-intervention and within-intervention effects on personal distress emerged from post-1MFU time points (Tables S1 and S2 in Multimedia Appendix 5).

Perspective Taking: Comprehending Others' Viewpoint

There were no significant differences between mindfulness EMI and self-monitoring on perspective-taking (IRI scores) from pre-post intervention (Cohen *d*=0.02, 95% CI −0.18 to 0.22; *P*=.84) and pre-1MFU (Cohen *d*=0.15, 95% CI −0.05 to 0.34; *P*=.13; Table 3). From pre-post intervention, no significant changes in perspective-taking emerged in the mindfulness EMI (Cohen *d*=0.01, 95% CI −0.18 to 0.21; *P*=.89) and self-monitoring (Cohen *d*=−0.01, 95% CI −0.21 to 0.18; *P*=.89) interventions. However, from pre-1MFU, the mindfulness EMI (Cohen *d*=0.25, 95% CI 0.06-0.45; *P*=.01) but not the self-monitoring app (Cohen *d*=0.00, 95% CI −0.20 to 0.20; *P*=.99; Table 4) yielded significant improvements in perspective-taking. No significant between-intervention and within-intervention effects on perspective taking emerged from post-1MFU time points (Tables S1 and S2 in Multimedia Appendix 5).

Discussion

Principal Findings and Comparison With Prior Work

Partially supporting our hypothesis, brief self-guided mindfulness EMI displayed longer-term efficacy from pre-1MFU time points compared to self-monitoring in improving 3 of the 5 examined social-cognitive domains: ToM, the capacity for imaginative immersion in various scenarios (*fantasy*), and experiencing distress when observing others' adverse situations (*personal distress*). Moreover, the mindfulness EMI, but not self-monitoring, enhanced perspective-taking from pre-1MFU time points despite the lack of between-intervention effects. An unexpected outcome was that neither of the interventions yielded effects on empathic concern. Our findings cannot be attributed to sociodemographic and comorbid psychiatric diagnoses, as those variables did not differ between compared conditions at prerandomization. On the whole, these outcomes suggest *specificity*, instead of globality, in the impact of brief mindfulness EMI and self-monitoring on ToM and trait-level empathy domains in the context of GAD. Potential theories are put forth to advance clinical science on this underinvestigated yet essential topic.

Why did brief mindfulness EMI but not self-monitoring enhance ToM (the ability to interpret social cues and represent abstract propositional knowledge about others' emotional and mental states), the capacity to envision being in another person's shoes (*fantasy*), and perspective-taking across 6 weeks? These outcomes extended reports of positive correlations between mindfulness and the ability to understand the emotions of self and others cognitively (cognitive empathy) [85-87]. The facets of nonjudgmental contemplation and a focus on the present moment emphasized by the mindfulness EMI could wield significant influence in cultivating empathic responses. Such mindfulness EMI instructions might augment cognitive capabilities, such as adopting another's perspective [9] and fostering a pragmatic understanding of human distress [88]. Collectively, our findings and these testable interpretations aligned harmoniously with previous 8- to 12-week MBI RCTs [89,90], highlighting the benefits of the mindfulness EMI's brevity on these social-cognitive domains.

Moreover, although the brief mindfulness EMI improved negative affective empathy (*distress*) or anxiety and stress experienced when confronted with the negative experiences of others, self-monitoring worsened it from pre-1MFU time points. This finding was inconsistent with a study showing that a brief mindfulness EMI was associated with increased *cognitive empathy* but decreased *affective empathy* (the capacity to share in the emotional experiences of others vicariously) [91]. Perhaps the brief mindfulness EMI evaluated in this study might have safeguarded against desensitization to suffering or developing callousness, enabling one to witness others' adverse experiences without feeling overwhelmed [92]. The opposite pattern occurred in the self-monitoring intervention, probably due to instructions to focus inwardly on one's stressful feelings and thoughts without asking the participant to have regard for others' difficulties.

Contrary to expectations, the brief mindfulness EMI did not outperform the self-monitoring app in improving empathic concern (experiencing affection and compassion when witnessing others' distress). Furthermore, no changes in empathic concern occurred from pre-post intervention and pre-1MFU time points. Similarly, a prior experiment found that a 5-minute mindfulness induction, compared to 2 control conditions, did not yield any discernible impact on empathic responses [93]. On the basis of prior RCTs that tested empirically-supported 8-week MBIs [94,95], a tenable account is a need for lengthier, higher-intensity (fully self-guided or coach-guided), multicomponent mindfulness EMIs focused on interpersonal relationships to evoke measurable improvement in empathic concern [96,97]. Future studies could test this conjecture and evaluate dose-response relationships between mindfulness EMI duration or intensity and its impact on various empathy domains. Alternatively, as prior research showed that people diagnosed with GAD experienced notably above-average levels of empathic concern [98], perhaps other empathy domains apart from empathic concern were more malleable to change via mindfulness EMIs in this population. A future RCT that recruited participants with both GAD and below-average levels of empathic concern would provide a direct test of this hypothesis.

As far as we know, there has been no other study that examined the effects of MBI on empathy in individuals with GAD. One prior study found that CBT but not MBI led to changes in positive affective empathy in individuals with social anxiety disorder compared to a waitlist control condition [33]. Furthermore, a change in positive affective empathy was a mediator of change in CBT but not in MBSR [33]. Given that we found changes in empathy and ToM domains relative to a control condition, perhaps enhancements in ToM and empathy domains are mechanisms of change of the effects of MBIs on GAD symptom alteration and related outcomes but not on social anxiety disorder. Conducting a mediation analysis to determine if changes in empathy and ToM domains were potential mechanisms of how brief mindfulness EMIs might improve GAD symptoms was beyond the scope of this study. However, a future report using the same data set will aim to explore this possibility.

Limitations and Strengths

Interpreting the findings within the context of the limitations herein is imperative for an in-depth understanding. First, it is worth noting that the 2-week intervention duration might be inadequate to elicit immediate enhancements in empathy domains; nevertheless, the results showed greater promise for habitual worriers from pre-1MFU time points. Second, we did not incorporate assessments to gauge mindfulness EMI participants' ongoing mindfulness skill use during the post-1MFU evaluation. Future investigations should, therefore, test the potential impact of continuous mindfulness practices, even in the absence of recurring guidance via the mindfulness EMI, on any discernible between-intervention effects during follow-up assessments. Third, we only assessed ToM and 4 empathy domains; extensive theories have suggested that MBIs are helpful in increasing other social-cognitive outcomes, such as altruism and prosocial behaviors [99], and these theories should be explored in GAD and related disorders. On that note, clinical science can profit from examining the effect of mindfulness EMIs on multimodal measures of empathy, such as proinflammatory cytokines, oxytocin, and neuroimaging indexes [100,101]. Finally, recruiting more diverse samples is required to establish cross-cultural generalizability [102,103].

Nonetheless, this study demonstrated notable strengths, including the gold standard RCT design featuring an active control intervention, a commendably high compliance rate, the recruitment of a clinically diagnosed sample with ample statistical power, and the incorporation of comprehensive assessments. Furthermore, it is noteworthy that our attrition rate, standing at a mere 11%, fell below the typical 24% to 50% range observed in RCTs delivered through smartphones [104,105]. Our relatively minimal attrition may be attributed to the prorated reimbursement schedule design.

Conclusions

In summary, these findings suggest that brief self-guided mindfulness EMIs such as ours hold promise as scalable solutions to enhance specific empathy domains (*perspective-taking* and *fantasy* or the ability to immerse oneself imaginatively in diverse situations experienced by others), *personal distress* (feeling distressed when witnessing unfavorable circumstances affecting others), and *ToM* (the skill to decipher social signals) for persons with GAD. Nonetheless, mindfulness EMI did not elicit between-intervention and within-intervention effects on *empathic concern* (experiencing warmth, compassion, and sincere concern while observing another person's distress) for this population. To advance our understanding of the genuine therapeutic efficacy of mindfulness EMIs (brief or prolonged) on social-cognitive domains, it becomes crucial to discern the specific subinterventions for whom it benefits, the contexts in which it exerts its effects, and when it may be contraindicated or inadvisable. Consequently, these efforts might address global challenges to social integration, anxiety disorders, and other psychological well-being aspects encountered by communities.

Acknowledgments

This trial received partial funding from the National Institute of Mental Health (R01 MH115128), the Penn State Office of Research and Graduate Studies Dissertation Award, the Penn State Susan Welch or Arthur Nagle Family Graduate Fellowship, the National University of Singapore Development Grant, and the Association for Behavioral and Cognitive Therapies Leonard Krasner Student Dissertation Award.

NHZ's research assistants (Ally Atkinson, Ana Luiza Clever Galvao, Drishti Sanghvi, Emma Win, Emily Forcht, Jamie Gensbauer, Kate Miller, Madeline Noelle Lamonica, Madeleine Miller, Marietta Kocher, Natalie Gottret, Natalie Suzette Marr, and Vivian Vo Nguyen) assisted heavily with the data collection and various logistical aspects of this study.

Authors' Contributions

All authors have (1) made substantial contributions to the analysis and interpretation of the study and its findings, (2) drafted and revised the paper for intellectual content, and (3) given their final approval of the version to be submitted. The manuscript has been read and approved by all authors. NHZ was involved in the conceptualization, methodology (design and creation of models), funding acquisition, formal analysis, coding with the R software (R Foundation for Statistical Computing) for data analysis, visualization, and writing (original draft, reviewing, and editing) of this work. MGN assisted with the manuscript's conceptualization, study design, data curation, validation, supervision, and writing (original draft, reviewing, and editing). NHZ and MGN take full responsibility for the data, accuracy of analyses and interpretation, and conduct of the research.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Study methodology details.
[DOCX File , 45 KB - mental_v11i1e54412_app1.docx]

Multimedia Appendix 2
CONSORT E-HEALTH checklist.
[PDF File (Adobe PDF File), 391 KB - mental_v11i1e54412_app2.pdf]

Multimedia Appendix 3
Screenshots of the intervention arm.
[DOCX File , 1349 KB - mental_v11i1e54412_app3.docx]

Multimedia Appendix 4
Screenshots of the control arm.
[DOCX File , 722 KB - mental_v11i1e54412_app4.docx]

Multimedia Appendix 5
Postintervention to 1-month follow-up outcomes.
[DOCX File , 26 KB - mental_v11i1e54412_app5.docx]

References

1. de Waal FB, Preston SD. Mammalian empathy: behavioural manifestations and neural basis. Nat Rev Neurosci 2017 Aug;18(8):498-509. [doi: 10.1038/nrn.2017.72] [Medline: 28655877]
2. Preston SD, de Waal FB. Only the PAM explains the personalized nature of empathy. Nat Rev Neurosci 2017 Dec;18(12):769. [doi: 10.1038/nrn.2017.140] [Medline: 29097789]
3. Preckel K, Kanske P, Singer T. On the interaction of social affect and cognition: empathy, compassion and theory of mind. Curr Opin Behav Sci 2018 Feb;19:1-6. [doi: 10.1016/j.cobeha.2017.07.010]
4. Schurz M, Radua J, Tholen MG, Maliske L, Margulies DS, Mars RB, et al. Toward a hierarchical model of social cognition: a neuroimaging meta-analysis and integrative review of empathy and theory of mind. Psychol Bull 2021 Mar;147(3):293-327. [doi: 10.1037/bul0000303] [Medline: 33151703]
5. Morelli SA, Lieberman MD, Zaki J. The emerging study of positive empathy. Soc Personal Psychol Compass 2015 Feb 05;9(2):57-68. [doi: 10.1111/spc3.12157]
6. Morelli SA, Ong DC, Makati R, Jackson MO, Zaki J. Empathy and well-being correlate with centrality in different social networks. Proc Natl Acad Sci U S A 2017 Sep 12;114(37):9843-9847 [FREE Full text] [doi: 10.1073/pnas.1702155114] [Medline: 28851835]
7. Zaki J. Integrating empathy and interpersonal emotion regulation. Annu Rev Psychol 2020 Jan 04;71:517-540. [doi: 10.1146/annurev-psych-010419-050830] [Medline: 31553672]
8. Bayot M, Vermeulen N, Kever A, Mikolajczak M. Mindfulness and empathy: differential effects of explicit and implicit Buddhist teachings. Mindfulness 2018 Jun 2;11(1):5-17. [doi: 10.1007/s12671-018-0966-4]
9. Block-Lerner J, Adair C, Plumb JC, Rhatigan DL, Orsillo SM. The case for mindfulness-based approaches in the cultivation of empathy: does nonjudgmental, present-moment awareness increase capacity for perspective-taking and empathic concern? J Marital Fam Ther 2007 Oct;33(4):501-516. [doi: 10.1111/j.1752-0606.2007.00034.x] [Medline: 17935532]
10. Malin Y. Others in mind: a systematic review and meta-analysis of the relationship between mindfulness and prosociality. Mindfulness 2023 Jun 13;14(7):1582-1605. [doi: 10.1007/s12671-023-02150-3]
11. Feldman C, Kuyken W. Compassion in the landscape of suffering. Contemporary Buddhism 2011 Jun 14;12(1):143-155. [doi: 10.1080/14639947.2011.564831]
12. Cooper D, Yap K, O'Brien M, Scott I. Mindfulness and empathy among counseling and psychotherapy professionals: a systematic review and meta-analysis. Mindfulness 2020 Jun 08;11(10):2243-2257. [doi: 10.1007/s12671-020-01425-3]

13. Bruce NG, Shapiro SL, Constantino MJ, Manber R. Psychotherapist mindfulness and the psychotherapy process. Psychotherapy (Chic) 2010 Mar;47(1):83-97. [doi: 10.1037/a0018842] [Medline: 22402003]
14. Cheang R, Gillions A, Sparkes E. Do mindfulness-based interventions increase empathy and compassion in children and adolescents: a systematic review. J Child Fam Stud 2019 Apr 20;28(7):1765-1779. [doi: 10.1007/s10826-019-01413-9]
15. Segal ZJ, Williams MG, Teasdale JD. Mindfulness-Based Cognitive Therapy for Depression: A New Approach to Preventing Relapse. New York, NY: The Guilford Press; 2002.
16. Centeno RP. Effect of mindfulness on empathy and self-compassion: an adapted MBCT program on Filipino college students. Behav Sci (Basel) 2020 Feb 27;10(3):61 [FREE Full text] [doi: 10.3390/bs10030061] [Medline: 32120924]
17. Hu Z, Wen Y, Wang Y, Lin Y, Shi J, Yu Z, et al. Effectiveness of mindfulness-based interventions on empathy: a meta-analysis. Front Psychol 2022;13:992575 [FREE Full text] [doi: 10.3389/fpsyg.2022.992575] [Medline: 36337535]
18. Kabat-Zinn J. An outpatient program in behavioral medicine for chronic pain patients based on the practice of mindfulness meditation: theoretical considerations and preliminary results. Gen Hosp Psychiatry 1982 Apr;4(1):33-47. [doi: 10.1016/0163-8343(82)90026-3] [Medline: 7042457]
19. Shapiro SL, Schwartz GE, Bonner G. Effects of mindfulness-based stress reduction on medical and premedical students. J Behav Med 1998 Dec;21(6):581-599. [doi: 10.1023/a:1018700829825] [Medline: 9891256]
20. Trautwein F, Kanske P, Böckler A, Singer T. Differential benefits of mental training types for attention, compassion, and theory of mind. Cognition 2020 Jan;194:104039 [FREE Full text] [doi: 10.1016/j.cognition.2019.104039] [Medline: 31450018]
21. López-Navarro E, Fonseca-Pedrero E, Errasti J, Al-Halabí S. Mindfulness improves theory of mind in people experiencing psychosis: a pilot randomized clinical trial. Psychiatry Res 2022 Apr;310:114440 [FREE Full text] [doi: 10.1016/j.psychres.2022.114440] [Medline: 35180612]
22. Creswell JD. Mindfulness Interventions. Annu Rev Psychol 2017 Jan 03;68:491-516. [doi: 10.1146/annurev-psych-042716-051139] [Medline: 27687118]
23. Wu J, Ma Y, Zuo Y, Zheng K, Zhou Z, Qin Y, et al. Effects of mindfulness exercise guided by a smartphone app on negative emotions and stress in non-clinical populations: a systematic review and meta-analysis. Front Public Health 2021;9:773296 [FREE Full text] [doi: 10.3389/fpubh.2021.773296] [Medline: 35155341]
24. Tan LB, Lo BC, Macrae CN. Brief mindfulness meditation improves mental state attribution and empathizing. PLoS One 2014;9(10):e110510 [FREE Full text] [doi: 10.1371/journal.pone.0110510] [Medline: 25329321]
25. Zheng D, Berry DR, Brown KW. Effects of brief mindfulness meditation and compassion meditation on parochial empathy and prosocial behavior toward ethnic out-group members. Mindfulness (N Y) 2023 Mar 27:1-17 [FREE Full text] [doi: 10.1007/s12671-023-02100-z] [Medline: 37362189]
26. Schreiter S, Pijnenborg GH, Aan Het Rot M. Empathy in adults with clinical or subclinical depressive symptoms. J Affect Disord 2013 Aug 15;150(1):1-16. [doi: 10.1016/j.jad.2013.03.009] [Medline: 23668900]
27. Kerr-Gaffney J, Harrison A, Tchanturia K. Cognitive and affective empathy in eating disorders: a systematic review and meta-analysis. Front Psychiatry 2019;10:102 [FREE Full text] [doi: 10.3389/fpsyt.2019.00102] [Medline: 30886590]
28. Baez S, Tangarife MA, Davila-Mejia G, Trujillo-Güiza M, Forero DA. Performance in emotion recognition and theory of mind tasks in social anxiety and generalized anxiety disorders: a systematic review and meta-analysis. Front Psychiatry 2023;14:1192683 [FREE Full text] [doi: 10.3389/fpsyt.2023.1192683] [Medline: 37275989]
29. Xu X, Dai J, Liu C, Chen Y, Xin F, Zhou F, et al. Common and disorder-specific neurofunctional markers of dysregulated empathic reactivity in major depression and generalized anxiety disorder. Psychother Psychosom 2020;89(2):114-116. [doi: 10.1159/000504180] [Medline: 31770760]
30. Tone EB, Tully EC. Empathy as a "risky strength": a multilevel examination of empathy and risk for internalizing disorders. Dev Psychopathol 2014 Nov;26(4 Pt 2):1547-1565 [FREE Full text] [doi: 10.1017/S0954579414001199] [Medline: 25422978]
31. Nair TK, Waslin SM, Rodrigues GA, Datta S, Moore MT, Brumariu LE. A meta-analytic review of the relations between anxiety and empathy. J Anxiety Disord 2024 Jan;101:102795. [doi: 10.1016/j.janxdis.2023.102795] [Medline: 38039916]
32. Pittelkow M, Aan Het Rot M, Seidel LJ, Feyel N, Roest AM. Social anxiety and empathy: a systematic review and meta-analysis. J Anxiety Disord 2021 Mar;78:102357 [FREE Full text] [doi: 10.1016/j.janxdis.2021.102357] [Medline: 33588287]
33. Morrison AS, Mateen MA, Brozovich FA, Zaki J, Goldin PR, Heimberg RG, et al. Changes in empathy mediate the effects of cognitive-behavioral group therapy but not mindfulness-based stress reduction for social anxiety disorder. Behav Ther 2019 Nov;50(6):1098-1111. [doi: 10.1016/j.beth.2019.05.005] [Medline: 31735245]
34. Bora E. Social cognition and empathy in adults with obsessive compulsive disorder: a meta-analysis. Psychiatry Res 2022 Oct;316:114752. [doi: 10.1016/j.psychres.2022.114752] [Medline: 35961153]
35. Stevens JS, Jovanovic T. Role of social cognition in post-traumatic stress disorder: a review and meta-analysis. Genes Brain Behav 2019 Jan;18(1):e12518 [FREE Full text] [doi: 10.1111/gbb.12518] [Medline: 30221467]
36. Newman MG, Przeworski A. J Anxiety Disord 2018 May;56:11-13 [FREE Full text] [doi: 10.1016/j.janxdis.2018.04.006] [Medline: 29859657]

37. Zainal NH, Newman MG. Worry amplifies theory-of-mind reasoning for negatively valenced social stimuli in generalized anxiety disorder. J Affect Disord 2018 Feb;227:824-833 [FREE Full text] [doi: 10.1016/j.jad.2017.11.084] [Medline: 29254067]
38. Aydın O, Balıkçı K, Çökmüş FP, Ünal Aydın P. The evaluation of metacognitive beliefs and emotion recognition in panic disorder and generalized anxiety disorder: effects on symptoms and comparison with healthy control. Nord J Psychiatry 2019;73(4-5):293-301. [doi: 10.1080/08039488.2019.1623317] [Medline: 31157577]
39. Erickson TM, Granillo MT, Crocker J, Abelson JL, Reas HE, Quach CM. Compassionate and self-image goals as interpersonal maintenance factors in clinical depression and anxiety. J Clin Psychol 2018 Apr;74(4):608-625 [FREE Full text] [doi: 10.1002/jclp.22524] [Medline: 28898407]
40. Borkovec TD, Newman MG, Pincus AL, Lytle R. A component analysis of cognitive-behavioral therapy for generalized anxiety disorder and the role of interpersonal problems. J Consult Clin Psychol 2002 Sep 01;70(2):288-298. [doi: 10.1037/0022-006X.70.2.288]
41. Hara KM, Westra HA, Constantino MJ, Antony MM. The impact of resistance on empathy in CBT for generalized anxiety disorder. Psychother Res 2018 Jul;28(4):606-615. [doi: 10.1080/10503307.2016.1244616] [Medline: 27756184]
42. Watson J, Timulak L, Greenberg LS. Emotion-focused therapy for generalized anxiety disorder. In: Greenberg LS, Goldman RN, editors. Clinical Handbook of Emotion-Focused Therapy. New York, NY: American Psychological Association; 2019:315-336.
43. Goetter EM, Frumkin MR, Palitz SA, Swee MB, Baker AW, Bui E, et al. Barriers to mental health treatment among individuals with social anxiety disorder and generalized anxiety disorder. Psychol Serv 2020 Feb;17(1):5-12. [doi: 10.1037/ser0000254] [Medline: 30070552]
44. Zainal NH, Newman MG. A randomized controlled trial of a 14-day mindfulness ecological momentary intervention (MEMI) for generalized anxiety disorder. Eur Psychiatry 2023 Jan 16;66(1):e12 [FREE Full text] [doi: 10.1192/j.eurpsy.2023.2] [Medline: 36645098]
45. Blanck P, Perleth S, Heidenreich T, Kröger P, Ditzen B, Bents H, et al. Effects of mindfulness exercises as stand-alone intervention on symptoms of anxiety and depression: systematic review and meta-analysis. Behav Res Ther 2018 Mar;102:25-35. [doi: 10.1016/j.brat.2017.12.002] [Medline: 29291584]
46. de Abreu Costa M, D'Alò de Oliveira GS, Tatton-Ramos T, Manfro GG, Salum GA. Anxiety and stress-related disorders and mindfulness-based interventions: a systematic review and multilevel meta-analysis and meta-regression of multiple outcomes. Mindfulness 2018 Nov 15;10(6):996-1005. [doi: 10.1007/s12671-018-1058-1]
47. Ghahari S, Mohammadi-Hasel K, Malakouti SK, Roshanpajouh M. Mindfulness-based cognitive therapy for generalised anxiety disorder: a systematic review and meta-analysis. East Asian Arch Psychiatry 2020 Jun;30(2):52-56 [FREE Full text] [doi: 10.12809/eaap1885] [Medline: 32611828]
48. Haller H, Breilmann P, Schröter M, Dobos G, Cramer H. A systematic review and meta-analysis of acceptance- and mindfulness-based interventions for DSM-5 anxiety disorders. Sci Rep 2021 Oct 14;11(1):20385 [FREE Full text] [doi: 10.1038/s41598-021-99882-w] [Medline: 34650179]
49. Zainal NH, Newman MG. Mindfulness enhances cognitive functioning: a meta-analysis of 111 randomized controlled trials. Health Psychol Rev 2023 Aug 30:1-27. [doi: 10.1080/17437199.2023.2248222] [Medline: 37578065]
50. Davis MH. Measuring individual differences in empathy: evidence for a multidimensional approach. J Pers Soc Psychol 1983 Jan;44(1):113-126. [doi: 10.1037/0022-3514.44.1.113]
51. Bell M, Bryson G, Lysaker P. Positive and negative affect recognition in schizophrenia: a comparison with substance abuse and normal control subjects. Psychiatry Res 1997 Nov 14;73(1-2):73-82 [FREE Full text] [doi: 10.1016/s0165-1781(97)00111-x] [Medline: 9463840]
52. Bonfils KA, Lysaker PH, Minor KS, Salyers MP. Empathy in schizophrenia: a meta-analysis of the interpersonal reactivity index. Psychiatry Res 2017 Mar;249:293-303 [FREE Full text] [doi: 10.1016/j.psychres.2016.12.033] [Medline: 28142103]
53. Zainal NH, Newman MG. Does a brief mindfulness ecological momentary intervention (MEMI) improve empathy and theory-of-mind for generalized anxiety disorder? A randomized controlled trial. Open Science Framework. 2023. URL: https://osf.io/vbsjd [accessed 2024-03-23]
54. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 5th edition. Arlington, VA: American Psychiatric Association; 2013.
55. Newman MG, Zuellig AR, Kachin KE, Constantino MJ, Przeworski A, Erickson T, et al. Preliminary reliability and validity of the generalized anxiety disorder questionnaire-IV: a revised self-report diagnostic measure of generalized anxiety disorder. Behav Ther 2002;33(2):215-233. [doi: 10.1016/S0005-7894(02)80026-0]
56. Brown TA, Barlow DH. Anxiety and Related Disorders Interview Schedule for DSM-5 (ADIS-5L): Client Interview Schedule. Oxford, UK: Oxford University Press; 2014.
57. Devilly GJ, Borkovec TD. Psychometric properties of the credibility/expectancy questionnaire. J Behav Ther Exp Psychiatry 2000 Jun;31(2):73-86. [doi: 10.1016/s0005-7916(00)00012-4] [Medline: 11132119]
58. Arend MG, Schäfer T. Statistical power in two-level models: a tutorial based on Monte Carlo simulation. Psychol Methods 2019 Feb;24(1):1-19. [doi: 10.1037/met0000195] [Medline: 30265048]

59. Homack S, Lee D, Riccio CA. Test review: delis-Kaplan executive function system. J Clin Exp Neuropsychol 2005 Jul;27(5):599-609. [doi: 10.1080/13803390490918444] [Medline: 16019636]
60. powerlmm: power analysis for longitudinal multilevel models. The Comprehensive R Archive Network. URL: https://cran.r-project.org/web/packages/powerlmm/index.html [accessed 2024-04-05]
61. Martínez-Loredo V, Fernández-Hermida JR, Carballo LS, Fernández-Artamendi S. Long-term reliability and stability of behavioral measures among adolescents: the Delay Discounting and Stroop tasks. J Adolesc 2017 Jul;58:33-39. [doi: 10.1016/j.adolescence.2017.05.003] [Medline: 28494412]
62. Moore MT, Anderson NL, Barnes JM, Haigh EA, Fresco DM. Using the GAD-Q-IV to identify generalized anxiety disorder in psychiatric treatment seeking and primary care medical samples. J Anxiety Disord 2014 Jan;28(1):25-30. [doi: 10.1016/j.janxdis.2013.10.009] [Medline: 24334213]
63. Rutter LA, Brown TA. Reliability and validity of the dimensional features of generalized anxiety disorder. J Anxiety Disord 2015 Jan;29:1-6. [doi: 10.1016/j.janxdis.2014.10.003] [Medline: 25465882]
64. Shankman SA, Funkhouser CJ, Klein DN, Davila J, Lerner D, Hee D. Reliability and validity of severity dimensions of psychopathology assessed using the structured clinical interview for DSM-5 (SCID). Int J Methods Psychiatr Res 2018 Mar;27(1):e1590. [doi: 10.1002/mpr.1590] [Medline: 29034525]
65. Spitzer RL, Williams JB, Kroenke K, Linzer M, deGruy FV, Hahn SR, et al. Utility of a new procedure for diagnosing mental disorders in primary care. The PRIME-MD 1000 study. JAMA 1994 Dec 14;272(22):1749-1756. [Medline: 7966923]
66. Wade EC, Cohen RT, Loftus P, Ruscio AM. A novel measure of real-time perseverative thought. Clin Psychol Sci 2022 May;10(3):534-552. [doi: 10.1177/21677026211038017] [Medline: 35959247]
67. Boutron I, Altman DG, Moher D, Schulz KF, Ravaud P, CONSORT NPT Group. CONSORT statement for randomized trials of nonpharmacologic treatments: a 2017 update and a CONSORT extension for nonpharmacologic trial abstracts. Ann Intern Med 2017 Jul 04;167(1):40-47 [FREE Full text] [doi: 10.7326/M17-0046] [Medline: 28630973]
68. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of Web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
69. American Psychiatric Association. Diagnostic And Statistical Manual Of Mental Disorders, 5th Edition, Text Revision (DSM-5-TR). New York, NY: American Psychiatric Association; 2022.
70. Kabat-Zinn J. Full Catastrophe Living: Using the Wisdom of Your Body and Mind to Face Stress, Pain and Illness. New York, NY: Delacorte Press; 1990.
71. LaFreniere LS, Newman MG. A brief ecological momentary intervention for generalized anxiety disorder: a randomized controlled trial of the worry outcome journal. Depress Anxiety 2016 Sep;33(9):829-839. [doi: 10.1002/da.22507] [Medline: 27062682]
72. Clarke J, Proudfoot J, Whitton A, Birch MR, Boyd M, Parker G, et al. Therapeutic alliance with a fully automated mobile phone and web-based intervention: secondary analysis of a randomized controlled trial. JMIR Ment Health 2016 Feb 25;3(1):e10 [FREE Full text] [doi: 10.2196/mental.4656] [Medline: 26917096]
73. Lutz J, Offidani E, Taraboanta L, Lakhan SE, Campellone TR. Appropriate controls for digital therapeutic clinical trials: a narrative review of control conditions in clinical trials of digital therapeutics (DTx) deploying psychosocial, cognitive, or behavioral content. Front Digit Health 2022 Aug 18;4:823977 [FREE Full text] [doi: 10.3389/fdgth.2022.823977] [Medline: 36060538]
74. Chrysikou EG, Thompson WJ. Assessing cognitive and affective empathy through the interpersonal reactivity index: an argument against a two-factor model. Assessment 2016 Dec;23(6):769-777. [doi: 10.1177/1073191115599055] [Medline: 26253573]
75. Bonfils KA, Longenecker JM, Hammer LA, Luther L. Measuring empathy in groups with high schizotypy: psychometric evaluation of the interpersonal reactivity index. Psychol Assess 2022 May;34(5):459-466. [doi: 10.1037/pas0001111] [Medline: 35099198]
76. Fernández AM, Dufey M, Kramp U. Testing the psychometric properties of the interpersonal reactivity index (IRI) in Chile. Eur J Psychol Assess 2011 Jan;27(3):179-185. [doi: 10.1027/1015-5759/a000065]
77. Pinkham AE, Penn DL, Green MF, Harvey PD. Social cognition psychometric evaluation: results of the initial psychometric study. Schizophr Bull 2016 Mar;42(2):494-504 [FREE Full text] [doi: 10.1093/schbul/sbv056] [Medline: 25943125]
78. Hughes RA, Heron J, Sterne JA, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. Int J Epidemiol 2019 Aug 01;48(4):1294-1304 [FREE Full text] [doi: 10.1093/ije/dyz032] [Medline: 30879056]
79. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med 2011 Feb 20;30(4):377-399. [doi: 10.1002/sim.4067] [Medline: 21225900]
80. Myers ND, Brincks AM, Ames AJ, Prado GJ, Penedo FJ, Benedict C. Multilevel modeling in psychosomatic medicine research. Psychosom Med 2012;74(9):925-936 [FREE Full text] [doi: 10.1097/PSY.0b013e3182736971] [Medline: 23107843]
81. Faraway JJ. Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models 2nd edition. Boca Raton, FL: CRC Press; 2016.

82. Feingold A. A regression framework for effect size assessments in longitudinal modeling of group differences. Rev Gen Psychol 2013 Mar;17(1):111-121 [FREE Full text] [doi: 10.1037/a0030048] [Medline: 23956615]
83. Feingold A. Confidence interval estimation for standardized effect sizes in multilevel and latent growth modeling. J Consult Clin Psychol 2015 Feb;83(1):157-168 [FREE Full text] [doi: 10.1037/a0037721] [Medline: 25181028]
84. Feingold A. Meta-analysis with standardized effect sizes from multilevel and latent growth models. J Consult Clin Psychol 2017 Mar;85(3):262-266 [FREE Full text] [doi: 10.1037/ccp0000162] [Medline: 28068118]
85. Beitel M, Ferrer E, Cecero JJ. Psychological mindedness and awareness of self and others. J Clin Psychol 2005 Jun;61(6):739-750. [doi: 10.1002/jclp.20095] [Medline: 15546143]
86. Fulton CL, Cashwell CS. Mindfulness‐based awareness and compassion: predictors of counselor empathy and anxiety. Couns Educ Superv 2015 Jun 02;54(2):122-133. [doi: 10.1002/ceas.12009]
87. Thomas JT, Otis MD. Intrapsychic correlates of professional quality of life: mindfulness, empathy, and emotional separation. J Soc Social Work Res 2010 Jan;1(2):83-98. [doi: 10.5243/jsswr.2010.7]
88. Raab K. Mindfulness, self-compassion, and empathy among health care professionals: a review of the literature. J Health Care Chaplain 2014 Jun 13;20(3):95-108. [doi: 10.1080/08854726.2014.913876] [Medline: 24926896]
89. Chan SH, Yu CK, Li AW. Impact of mindfulness-based cognitive therapy on counseling self-efficacy: a randomized controlled crossover trial. Patient Educ Couns 2021 Feb;104(2):360-368. [doi: 10.1016/j.pec.2020.07.022] [Medline: 32798079]
90. Schonert-Reichl KA, Oberle E, Lawlor MS, Abbott D, Thomson K, Oberlander TF, et al. Enhancing cognitive and social-emotional development through a simple-to-administer mindfulness-based school program for elementary school children: a randomized controlled trial. Dev Psychol 2015 Jan;51(1):52-66 [FREE Full text] [doi: 10.1037/a0038454] [Medline: 25546595]
91. Borghi O, Mayrhofer L, Voracek M, Tran US. Differential associations of the two higher-order factors of mindfulness with trait empathy and the mediating role of emotional awareness. Sci Rep 2023 Feb 24;13(1):3201 [FREE Full text] [doi: 10.1038/s41598-023-30323-6] [Medline: 36828879]
92. Leonard HD, Campbell K, Gonzalez VM. The relationships among clinician self-report of empathy, mindfulness, and therapeutic alliance. Mindfulness 2018 Mar 12;9(6):1837-1844. [doi: 10.1007/s12671-018-0926-z]
93. Ridderinkhof A, de Bruin EI, Brummelman E, Bögels SM. Does mindfulness meditation increase empathy? An experiment. Self Identity 2017 Jan 02;16(3):251-269. [doi: 10.1080/15298868.2016.1269667]
94. Shapiro SL, Brown KW, Thoresen C, Plante TG. The moderation of mindfulness-based stress reduction effects by trait mindfulness: results from a randomized controlled trial. J Clin Psychol 2011 Mar;67(3):267-277. [doi: 10.1002/jclp.20761] [Medline: 21254055]
95. Wallmark E, Safarzadeh K, Daukantaitė D, Maddux RE. Promoting altruism through meditation: an 8-week randomized controlled pilot study. Mindfulness 2012 Jun 4;4(3):223-234. [doi: 10.1007/s12671-012-0115-4]
96. Goldberg SB, Imhoff-Smith T, Bolt DM, Wilson-Mendenhall CD, Dahl CJ, Davidson RJ, et al. Testing the efficacy of a multicomponent, self-guided, smartphone-based meditation app: three-armed randomized controlled trial. JMIR Ment Health 2020 Nov 27;7(11):e23825 [FREE Full text] [doi: 10.2196/23825] [Medline: 33245288]
97. Silveira S, Godara M, Singer T. Boosting empathy and compassion through mindfulness-based and socioemotional dyadic practice: randomized controlled trial with app-delivered trainings. J Med Internet Res 2023 Jul 26;25:e45027 [FREE Full text] [doi: 10.2196/45027] [Medline: 37494106]
98. Cordaro M, Grigsby TJ, Howard JT, Deason RG, Haskard-Zolnierek K, Howard K. Pandemic-specific factors related to generalized anxiety disorder during the initial COVID-19 protocols in the United States. Issues Ment Health Nurs 2021 Aug;42(8):747-757. [doi: 10.1080/01612840.2020.1867675] [Medline: 33480832]
99. Van Doesum NJ, Van Lange DA, Van Lange PA. Social mindfulness: skill and will to navigate the social world. J Pers Soc Psychol 2013 Jul;105(1):86-103. [doi: 10.1037/a0032540] [Medline: 23647176]
100. Trofimova IN, Gaykalova AA. Emotionality vs. other biobehavioural traits: a look at neurochemical biomarkers for their differentiation. Front Psychol 2021;12:781631 [FREE Full text] [doi: 10.3389/fpsyg.2021.781631] [Medline: 34987450]
101. Weisz E, Zaki J. Motivated empathy: a social neuroscience perspective. Curr Opin Psychol 2018 Dec;24:67-71. [doi: 10.1016/j.copsyc.2018.05.005] [Medline: 29966924]
102. Zainal NH, Newman MG, Hong RY. Cross-cultural and gender invariance of transdiagnostic processes in the United States and Singapore. Assessment 2021 Mar;28(2):485-502 [FREE Full text] [doi: 10.1177/1073191119869832] [Medline: 31538795]
103. Van Doren N, Zainal NH, Newman MG. Cross-cultural and gender invariance of emotion regulation in the United States and India. J Affect Disord 2021 Dec 01;295:1360-1370 [FREE Full text] [doi: 10.1016/j.jad.2021.04.089] [Medline: 34706449]
104. Linardon J, Fuller-Tyszkiewicz M. Attrition and adherence in smartphone-delivered interventions for mental health problems: a systematic and meta-analytic review. J Consult Clin Psychol 2020 Jan;88(1):1-13. [doi: 10.1037/ccp0000459] [Medline: 31697093]
105. Torous J, Lipschitz J, Ng M, Firth J. Dropout rates in clinical trials of smartphone apps for depressive symptoms: a systematic review and meta-analysis. J Affect Disord 2020 Feb 15;263:413-419. [doi: 10.1016/j.jad.2019.11.167] [Medline: 31969272]

Abbreviations

1MFU: 1-month follow-up
ADIS: Anxiety and Related Disorders Interview Schedule
BLERT: Bell-Lysaker Emotion Recognition Task
CBT: cognitive behavioral therapy
CONSORT: Consolidated Standards of Reporting Trials
CONSORT-EHEALTH: Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth
DSM-5: Diagnostic and Statistical Manual, Fifth Edition
EMI: ecological momentary intervention
GAD: generalized anxiety disorder
GAD-Q-IV: Generalized Anxiety Disorder Questionnaire, Fourth Version
HLM: hierarchical linear modeling
IRI: Interpersonal Reactivity Index
MBCT: mindfulness-based cognitive therapy
MBI: mindfulness-based intervention
MBSR: mindfulness-based stress reduction
RCT: randomized controlled trial
ToM: theory-of-mind

Edited by J Torous; submitted 08.11.23; peer-reviewed by E Bernstein, N Van Doren; comments to author 23.11.23; revised version received 26.12.23; accepted 09.01.24; published 24.05.24.

Please cite as:
Zainal NH, Newman MG
Examining the Effects of a Brief, Fully Self-Guided Mindfulness Ecological Momentary Intervention on Empathy and Theory-of-Mind for Generalized Anxiety Disorder: Randomized Controlled Trial
JMIR Ment Health 2024;11:e54412
URL: https://mental.jmir.org/2024/1/e54412
doi:10.2196/54412
PMID:38787613

Original Paper

Effects of a Chatbot-Based Intervention on Stress and Health-Related Parameters in a Stressed Sample: Randomized Controlled Trial

Christine Schillings[1], MSc; Echo Meißner[2], MSc; Benjamin Erb[2], Dr rer nat; Eileen Bendig[3], DPhil; Dana Schultchen[1], DPhil; Olga Pollatos[1], MD, DPhil

[1]Department of Clinical and Health Psychology, Institute of Psychology and Education, Ulm University, Ulm, Germany

[2]Institute of Distributed Systems, Ulm University, Ulm, Germany

[3]Department of Clinical Psychology and Psychotherapy, Institute of Psychology and Education, Ulm University, Ulm, Germany

Corresponding Author:
Christine Schillings, MSc
Department of Clinical and Health Psychology
Institute of Psychology and Education
Ulm University
Albert-Einstein-Allee 43
Ulm, 89081
Germany
Phone: 49 731 50 31738
Email: christine.schillings@uni-ulm.de

Abstract

Background: Stress levels and the prevalence of mental disorders in the general population have been rising in recent years. Chatbot-based interventions represent novel and promising digital approaches to improve health-related parameters. However, there is a lack of research on chatbot-based interventions in the area of mental health.

Objective: The aim of this study was to investigate the effects of a 3-week chatbot-based intervention guided by the chatbot ELME, specifically with respect to the ability to reduce stress and improve various health-related parameters in a stressed sample.

Methods: In this multicenter two-armed randomized controlled trial, 118 individuals with medium to high stress levels were randomized to the intervention group (n=59) or the treatment-as-usual control group (n=59). The ELME chatbot guided participants of the intervention group through 3 weeks of training based on the topics stress, mindfulness, and interoception, with practical and psychoeducative elements delivered in two daily interactive intervention sessions via a smartphone (approximately 10-20 minutes each). The primary outcome (perceived stress) and secondary outcomes (mindfulness; interoception or interoceptive sensibility; subjective well-being; and emotion regulation, including the subfacets reappraisal and suppression) were assessed preintervention (T1), post intervention (T2; after 3 weeks), and at follow-up (T3; after 6 weeks). During both conditions, participants also underwent ecological momentary assessments of stress and interoceptive sensibility.

Results: There were no significant changes in perceived stress (β_{03}=–.018, SE=.329; *P*=.96) and momentary stress. Mindfulness and the subfacet reappraisal significantly increased in the intervention group over time, whereas there was no change in the subfacet suppression. Well-being and momentary interoceptive sensibility increased in both groups over time.

Conclusions: To gain insight into how the intervention can be improved to achieve its full potential for stress reduction, besides a longer intervention duration, specific sample subgroups should be considered. The chatbot-based intervention seems to have the potential to improve mindfulness and emotion regulation in a stressed sample. Future chatbot-based studies and interventions in health care should be designed based on the latest findings on the efficacy of rule-based and artificial intelligence–based chatbots.

Trial Registration: German Clinical Trials Register DRKS00027560; https://drks.de/search/en/trial/DRKS00027560

International Registered Report Identifier (IRRID): RR2-doi.org/10.3389/fdgth.2023.1046202

(JMIR Ment Health 2024;11:e50454) doi:10.2196/50454

KEYWORDS

chatbot; intervention; stress; interoception; interoceptive sensibility; mindfulness; emotion regulation; RCT; randomized controlled trial

Introduction

Stress levels and the prevalence of mental disorders in the general population have been rising in recent years, which have been further accelerated by the COVID-19 pandemic [1]. Digital mindfulness-based interventions were indicated as promising approaches to improve mental health outcomes such as stress (eg, [2-4]), mindfulness (eg, [2,3]), or subjective well-being (eg, [3]), highlighting the crucial role of emotion regulation [5]. In particular, guided online interventions are of high relevance, as they are associated with higher adherence rates than unguided interventions [6-9]. Novel digital approaches of increasing interest include support from chatbots [10-12], which can be used anonymously, regardless of time and location, and are easily integrated into individuals' everyday lives [6,13-16].

Studies of chatbot-based interventions aiming to improve mental health outcomes have provided evidence for decreases in distress [17-20] or increased subjective and psychological well-being [18,21,22]. Importantly, randomized controlled trials (RCTs) in the context of mental health are sparse and inconsistent [12,20,23-25] and there is a lack of research on the efficacy of chatbot-based interventions [20,23,24], especially for emotion regulation and interoception [26,27]. Given the impairment of interoceptive abilities in mental disorders (eg, [28,29]) or under long-term stress [30], approaches that can help to train interoceptive abilities are essential, which can be achieved via mindfulness-based interventions (eg, [31,32]).

Overall, there is a need for research on chatbot-based interventions considering standardized characteristics (eg, intervention duration, samples, outcome assessments) and guidelines. Furthermore, interoception has not been the focus of previous research on chatbot-based interventions, neither being included as part of the intervention contents nor implemented as ecological momentary assessment (EMA) measures. EMA represents a flexible approach to measure real-time data, including health data, in everyday life [33]. Therefore, to fill these gaps, we developed a new chatbot-based intervention fostering the abilities of interoception, mindfulness, and stress management in everyday life.

The aim of this study was to investigate the effects of a 3-week chatbot-based intervention on stress, mindfulness, interoception, subjective well-being, and emotion regulation in individuals with medium to high stress levels. Based on previous findings, perceived stress was chosen as the primary outcome. Further details are described in the study protocol [34].

We hypothesized that: (1) the primary outcome (perceived stress) will be reduced in the intervention group compared to the treatment-as-usual control group over time, as assessed at preintervention (T1), post intervention (T2), and at the 3-week follow-up (T3) and via EMA; and (2) the secondary outcomes (mindfulness; interoception, including interoceptive sensibility; subjective well-being, and emotion regulation) will be improved in the intervention group compared to the control group, as assessed at T1, T2, and T3. Momentary interoceptive sensibility and stress were also assessed via EMA.

Furthermore, adherence, dropout reasons, usability, and user feedback regarding the intervention were assessed to potentially further improve the intervention for future research.

Methods

Setting and Recruitment

The data collection took place between February and September 2022. German-speaking people were recruited via offline and online recruitment strategies. Participants were included in the study if they (1) were 18 years or older, (2) had sufficient knowledge of the German language, (3) owned a smartphone (Android or iOS) with internet access, (4) possessed a valid smartphone number, (5) possessed a valid mailing address, (6) experienced a middle to high level of perceived stress (according to a 10-item Perceived Stress Scale [PSS-10] score≥14, assessed at screening [T0]), (7) were not diagnosed with any mental disorder, (8) did not undertake psychotherapy, and (9) were not currently participating in another online mental health intervention.

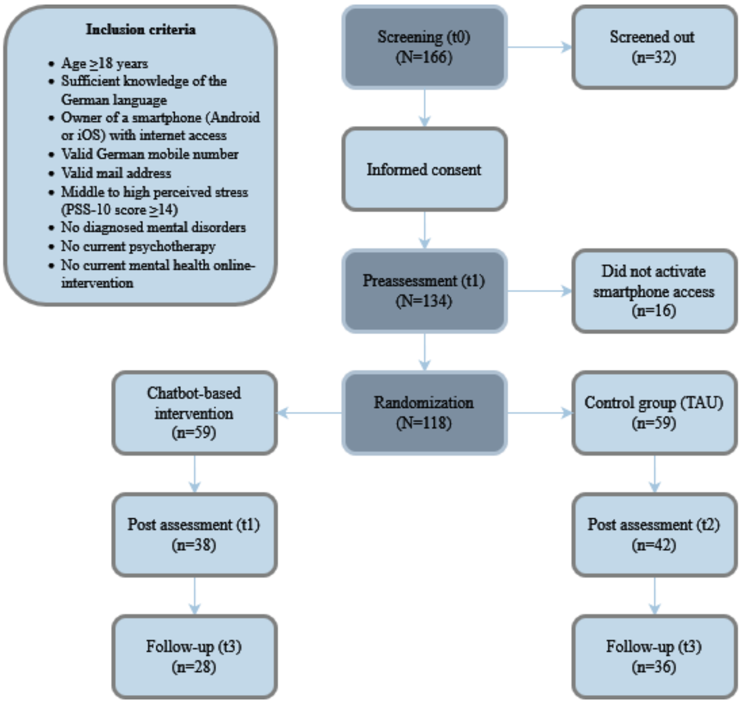
Study Design

The intervention group received a 3-week online-based intervention guided by the chatbot ELME. The control group received treatment as usual (ie, no content and only answered the questionnaires and the EMAs). Primary and secondary outcomes were assessed in both groups at T0, T1, daily during the intervention (between T1 and T2), T2, and T3. The design of the study and the usability of the chatbot were successfully tested in a previous feasibility study [35]. The trial was registered a priori at the World Health Organization (WHO) International Clinical Trials Registry Platform via the German Clinical Studies Trial Register (DRKS00027560) on January 6, 2022. The detailed design of this two-armed, parallel RCT is presented in the published study protocol [34].

Study Procedure

Figure 1 provides a schematic of the study procedure including the final numbers of participants.

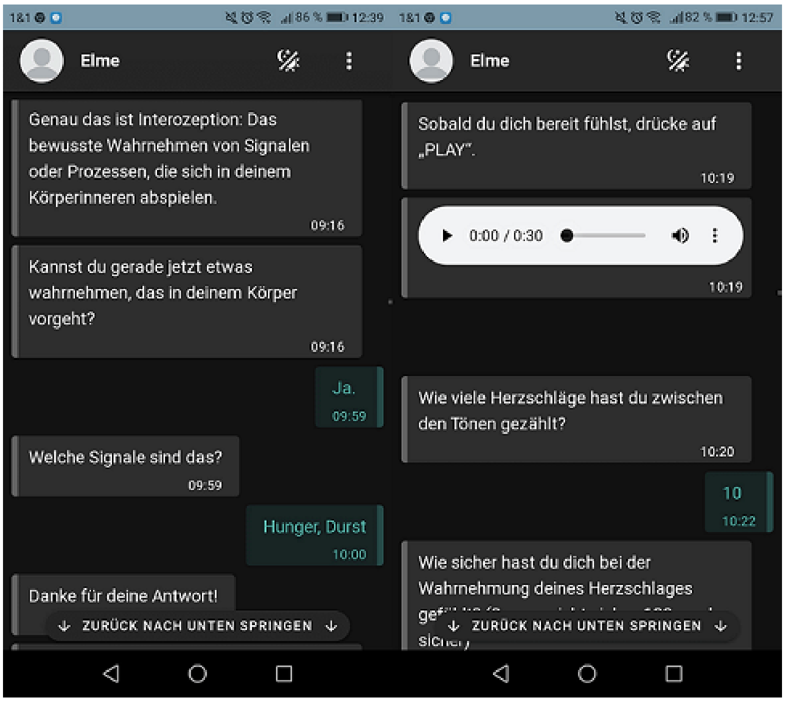
Figure 1. Flowchart of the study procedure. PSS-10: 10-item Perceived Stress Scale; TAU: treatment as usual.



Intervention

ELME is a rule-based chatbot, implemented as a web-based mobile app. ELME offers psychoeducation, exercises in real-time dialogues with the chatbot, audio files, and individual feedback. Sessions were held twice a day (for approximately 10-20 minutes each) over 3 weeks and with flexible timing. Participants could postpone exercises and receive SMS text message reminders. For more detailed intervention information and the detailed procedure, see descriptions in the study protocol [34]. Examples of representative dialogues of the interaction between the chatbot and a participant are depicted in Figure 2.

Figure 2. Screenshots of representative dialogues in the interaction between the chatbot and a participant (in German).



Ethical Considerations

All study procedures were approved by the ethics committee of Ulm University (application number 401/20). Written informed consent was obtained from all participants prior to their participation. As an incentive, participants could take part in the intervention for free and received the chance to win a €25 (approximately US $26) gift card from an online shop or, as a student participant, to receive 5 course credits as expense allowance for completing the questionnaires. A further incentive was the possible access to two relaxing exercises and to obtain individual summaries regarding the change in the participants' health-related parameters from preintervention to follow-up after completing the T3 questionnaire.

Outcome Assessments

Primary Outcome: Perceived Stress

The PSS-10 [36] was used as a screening questionnaire. At T1 to T3, perceived stress was assessed via the 4-item short scale (PSS-4). The ratings on both scales, ranging from 0="never" to 4="very often," were calculated as sum scores, with higher scores representing higher perceived stress.

Secondary Outcomes

Mindfulness

The 14-item short version of the Freiburg Mindfulness Inventory [37] was used to assess mindfulness. Answers were rated on a 4-point Likert scale ranging from 1="rarely" to 4="almost always." A sum score (range 14-56) was calculated, with higher scores indicating higher mindfulness.

Interoceptive Sensibility

Interoceptive sensibility was measured by German versions of the Interoceptive Accuracy Scale (IAS) [38] and the "Awareness" subscale of the Body Perception Questionnaire (BPQ) [39]. The 21-item IAS was rated on a 5-point Likert scale ranging from 1="strongly disagree" to 5="strongly agree." Higher sum scores (range 21-105) indicate greater interoceptive sensibility. The "Awareness" subscale of the BPQ consists of 45 items rated on a 5-point Likert scale ranging from 1="never" to 5="always."

Subjective Well-Being

The 5-item WHO Well-Being Index [40,41] was used to assess subjective well-being. Participants responded on a 5-point Likert scale ranging from 5="all of the time" to 0="at no time." A total sum score (range 0-100, with 100 indicating the best well-being) was calculated from raw scores (range 0-25) and multiplied by 4.

Emotion Regulation

The German version [42] of the Emotion Regulation Questionnaire [43] was used to assess emotion regulation. The 10-item questionnaire included 6 items representing emotion regulation strategy reappraisal and 4 items assessing emotion regulation strategy suppression, rated on a 7-point Likert scale ranging from 1="strongly disagree" to 7="strongly agree." Accordingly, the mean scores reflect the use of and preferences for various emotion regulation strategies.

Ecological Momentary Assessment

Momentary perceived stress and momentary interoceptive sensibility were measured via EMAs twice a day (in the morning and in the afternoon). Momentary perceived stress was assessed via two adapted items for the momentary use of the PSS-4 [36]: "Do you feel that things are going your way?" and "Do you find you can cope with all the things that you have to do?" The items "How present do you feel at the moment?" and "How aware are you of your own body at the moment?" [31,44] were used to measure momentary body awareness. To assess interoceptive sensibility, we used a self-developed question, taking the heartbeat perception task developed by Schandry [45] into account: "How intense do you perceive your heartbeat in the moment?" All rating scales were presented as visual analog scales ranging from 0="not at all" to 100="very much."

Mental Health App Usability Questionnaire

To assess the usability of the chatbot as a mental health app, a self-translated German version of the 18-item Mental Health App Usability Questionnaire [46] was used, rated on a scale ranging from 1="strongly agree" to 7="strongly disagree." The questionnaire comprises the following three subscales: ease of use (5 items), interface and satisfaction (7 items), and usefulness (6 items). Mean scores for each subscale were calculated as a total mean score, with lower scores reflecting higher usability.

Adherence, Potential Dropout Reasons, and User Feedback

Adherence to the intervention was operationalized by the percentage of completed modules of the intervention. Reasons for potential dropout were assessed via the Dropout Reasons Questionnaire for Internet Interventions [47]. User feedback questions asking the participants if they liked the training (range 1-10) and judging the extent of the training (1="too short" to 12="too long") were also assessed.

Data Analysis

Data analyses were performed according to the intention-to-treat principle. Due to the nested longitudinal data structure, hierarchical linear regression models were constructed to investigate the intervention effects over time. The measurement points (level 1) were nested within the participants (level 2). The regression analyses include the 3 measurement points preintervention (T1), post intervention (T2), and follow-up (T3). We analyzed hierarchical linear models and model comparisons in R using the packages lme4 [48], lmerTest [49], and r2mlm [50]. The predictor variable time had an interpretable 0 point and the dichotomous predictor group was dummy-coded. Due to assumed interindividual and intraindividual differences in all outcome variables, random-intercept, random-slope models were calculated. The restricted maximum-likelihood estimator was applied for parameter estimation, as it is generally considered to be less biased compared to the maximum-likelihood estimation [51]. We here report the main results that address hypotheses (1) and (2). The significance level for all analyses was set to $P \leq .05$.

Results

Participant Characteristics

A total sample of 118 participants was randomized to the intervention group (n=59; 72% female) and to the control group (n=59; 81% female). The relevant descriptive statistics at T1 are summarized in Table 1; there were no significant differences between the groups at T1.

Table 1. Comparison of relevant participant characteristics at baseline.

Characteristics	Intervention group (n=59), mean (SD)	Control group (n=59), mean (SD)	*t* (*df*=116)	*P* value
Age	33.117 (11.778)	33.085 (13.853)	–0.014	.99
Perceived stress	8.017 (2.701)	7.627 (2.355)	–0.836	.41
Mindfulness	34.271 (5.825)	34.051 (6.957)	0.187	.85
IS[a] (IAS[b])	82.220 (10.992)	80.475 (9.233)	–0.934	.35
IS (BPQ[c])	3.252 (.793)	3.305 (.561)	0.443	.66
Well-being	41.153 (17.341)	42.780 (17.297)	0.510	.61
Emotion regulation: reappraisal	4.130 (1.139)	4.429 (.934)	1.561	.12
Emotion regulation: suppression	3.725 (1.260)	3.339 (1.227)	–1.684	.10

[a]IS: interoceptive sensibility.

[b]IAS: Interoceptive Accuracy Scale.

[c]BPQ: Body Perception Questionnaire.

Perceived Stress

According to the model regarding perceived stress (Table 2), the nonsignificant fixed effect of the level-1 predictor time indicated that the stress levels did not change over time (from T1 to T3). The fixed effect of the level-2 predictor group and the cross-level interaction of the variables time and group were also not significant. The results of the two models predicting momentary perceived stress showed neither significant main effects of time and group nor their interactions (see Tables S1 and S2 in Multimedia Appendix 1).

Table 2. Random-intercept, random-slope model for perceived stress with the predictors time, group, and their interaction.

Effects	Coefficient[a] (SE or SD)	*df*	*t*	*P* value
Fixed effects				
Intercept	8.135 (.337)	122.119	24.141	<.001
Level 1: time	–.341 (.222)	141.471	–1.538	.13
Level 2: group	.172 (.478)	123.037	0.359	.72
Cross-level interaction (time×group)	–.018 (.329)	140.501	–0.055	.96
Random effects (variance components)				
σ_{u0j} (Intercept)	3.243 (1.801)	—[b]	—	—
σ^2_{u01j} (Time)	0.017 (0.132)	—	—	—
σ^2_{rij} (Residual)	3.916 (1.979)	—	—	—

[a]Fixed-effects coefficients are β values reported with SEs; random-effects coefficients are σ² (variance) values reported with SDs.

[b]Not applicable.

Mindfulness

The results of the model regarding mindfulness (Table 3) showed no significant fixed effects for time and group. However, the cross-level interaction of time and group was significant.

Table 3. Random-intercept, random-slope model for mindfulness with the predictors time, group, and their interaction.

Effects	Coefficient[a] (SE or SD)	*df*	*t*	*P* value
Fixed effects				
Intercept	34.362 (.826)	115.647	41.589	<.001
Level 1: time	.078 (.350)	68.640	0.223	.84
Level 2: group	–.080 (1.170)	116.058	–0.069	.95
Cross-level-interaction (time×group)	1.130	71.363	2.171	.03
Random effects (variance components)				
σ^2_{u0j} (Intercept)	31.906 (5.649)	—[b]	—	—
σ^2_{u01j} (Time)	0.113 (0.336)	—	—	—
σ^2_{rij} (Residual)	9.461 (3.076)	—	—	—

[a]Fixed-effects coefficients are β values reported with SEs; random-effects coefficients are σ² (variance) values reported with SDs.
[b]Not applicable.

Interoceptive Sensibility

The results of the model predicting interoceptive sensibility (assessed via the IAS) revealed neither significant main effects of time or group nor their interaction (see Table S3 in Multimedia Appendix 1). Similarly, assessments via the BPQ showed no significant effects (see Table S4 in Multimedia Appendix 1).

Momentary interoceptive sensibility increased on average over time in both groups. However, the main effect for group and the cross-level interaction of time and group were not significant (see Table S5 in Multimedia Appendix 1).

Well-Being

As shown in Table 4, subjective well-being improved over time in both groups on average. However, there were neither significant differences in well-being for the groups nor over both time and group.

Table 4. Random-intercept, random-slope model for well-being with the predictors time, group, and their interaction.

Effects	Coefficient[a] (SE or SD)	*df*	*t*	*P* value
Fixed effects				
Intercept	42.150 (2.236)	115.591	18.848	< .001
Level 1 (time)	4.237 (1.479)	74.662	2.865	.005
Level 2 (group)	–.032 (3.168)	116.363	–0.010	.99
Cross-level interaction (time×group)	2.312 (2.189)	78.034	1.056	.29
Random effects (variance components)				
σ^2_{u0j} (Intercept)	185.17 (13.608)	—[b]	—	—
σ^2_{u01j} (Time)	26.580 (5.156)	—	—	—
σ^2_{rij} (Residual)	124.180 (11.144)	—	—	—

[a]Fixed-effects coefficients are β values reported with SEs; random-effects coefficients are σ² (variance) values reported with SDs.
[b]Not applicable.

Emotion Regulation: Reappraisal Subfacet

The results of the model concerning the subfacet reappraisal of emotion regulation (Table 5) revealed neither a significant effect of time nor of group. However, the cross-level interaction of time and group was significant.

Table 5. Random-intercept, random-slope model for the emotion regulation reappraisal subfacet with the predictors time, group, and their interaction.

Effects	Coefficient[a] (SE or SD)	*df*	*t*	*P* value
Fixed effects				
Intercept	4.426 (.133)	116.328	33.297	< .001
Level 1: time	.022 (.064)	127.10	0.347	.73
Level 2: group	–.255 (.188)	116.819	–1.353	.18
Cross-level interaction (time×group)	.223 (.096)	127.085	2.331	.02
Random effects (variance components)				
σ^2_{u0j} (Intercept)	0.775 (0.880)	—[b]	—	—
σ^2_{u01j} (Time)	0.006 (0.075)	—	—	—
σ^2_{rij} (Residual)	0.302 (0.550)	—	—	—

[a]Fixed-effects coefficients are β values reported with SEs; random-effects coefficients are σ² (variance) values reported with SDs.
[b]Not applicable.

Emotion Regulation: Suppression Subfacet

Results regarding the suppression subfacet of emotion regulation revealed no significant changes (see Table S6 in Multimedia Appendix 1).

Adherence, Dropout Reasons, and User Feedback

The mean adherence (percentage of completed modules) was 58% for the 59 participants in the intervention group; 23 participants skipped intervention units, with "no time" cited as the main reason (n=19). In addition, 22 participants reported technical problems. The average answer rate of the EMA questions was 48% in the intervention group and 66% in the control group. In response to the question if the participants liked the training, the mean score was 6.95 (SD 1.86). The extent of the training was rated a mean score of 7.62.

Usability

The mean usability score (total score) was 2.55 (SD 0.68), with means of the subscales “ease of use,” “interface and satisfaction,” and “usefulness” of 1.85 (SD 1.01), 2.62 (SD 1.08), and 3.2 (SD 0.94), respectively.

Discussion

Principal Results

The aim of this study was to examine the effects of a 3-week chatbot-based intervention on perceived stress and various health-related parameters in stressed individuals. The results show no significant changes in perceived stress after the intervention. There was a significant increase in mindfulness and in emotion regulation as assessed by the subfacet reappraisal in the intervention group over time, whereas there was no change in the suppression subfacet of emotion regulation. Well-being and momentary interoceptive sensibility increased in both groups over time.

Comparison With Prior Work

Effects on Perceived Stress

The nonsignificant reduction in perceived stress is in line with the findings of a similar intervention study [52] and a pilot study [22]; however, considering statistical power problems of these studies, the intervention duration or intensity might be one factor to consider for interpreting the missing effects of the present study. Another explanation could be that there might have been greater initial focus on stress perception, which would potentially buffer the stress-reducing effects due to the intervention. This is supported by findings from psychotherapeutic interventions [53,54], in which the hypothesized effects on psychological outcomes were only detected later because of the confrontation with emotionally charging topics. Furthermore, the results of studies by Baer et al [55] and Venkatesan et al [56] indicated that the effects on perceived stress might become (more) visible after a longer duration of the intervention.

The results regarding momentary perceived stress are in line with previous studies evaluating the effects of 3-month mindfulness-based interventions [57,58]. Moreover, considering the mean adherence of 58% for the present intervention, the mean answer rates of the EMA questions need to be considered when interpreting the results.

Effects on Mindfulness

The significant increase in mindfulness is in line with previous findings from online mindfulness-based interventions (eg, [2,3]), indicating that the 3-week chatbot-based intervention comprising mindfulness-based content has the potential to increase mindfulness over time in a stressed sample. A possible mechanism might be that the contents of the intervention addressing mindfulness, stress, and interoception support mindfulness. Nevertheless, mindfulness needs to be interpreted as a secondary outcome in this study.

Effects on Interoceptive Sensibility

The missing effects of interoceptive sensibility in this study are in contrast to previous positive effects found for diverse mindfulness-based interventions (eg, [31,32,59]). However, these effects were found in the context of interventions lasting at least 8 weeks. In particular, and in line with the present findings, a 1-week mindfulness-based intervention [32] or a 3-week heartbeat perception training [60] could not improve interoceptive abilities. The findings of this study support the conclusions put forth by Fischer et al [59], Bornemann and Singer [31], and Schillings et al [60] that a longer intervention might be necessary to effectively improve interoceptive abilities. Moreover, previous studies differed in the methods used to assess diverse dimensions of interoceptive abilities (eg, [61,62]). Finally, a longer intervention design of such an innovative chatbot-based intervention might only be reasonable after initial trials with a shorter intervention design such as that of 3 weeks used in this study.

Due to the innovative EMA questions in this study and another study design not including an intervention, the results are not comparable to the previous EMA study by Höller et al [63]. The significant increase in momentary interoceptive sensibility could be explained by a training effect of frequent EMAs, which took place twice a day over 3 weeks.

Effects on Emotion Regulation

In line with the results regarding reappraisal, a recent systematic review and meta-analysis on mental health apps to promote emotion regulation and positive mental health in the general population [64] found a medium effect size (g=0.49) for emotion regulation compared to control conditions. However, it must be emphasized that this effect was based on only 6 studies, reflecting the lack of RCTs on chatbot-based interventions addressing emotion regulation.

Effects on Well-Being

The increase in well-being is in line with comparable previous studies [18,21,22,65] considering differences in the study designs and samples. However, well-being also improved in the control group of this study, which might have also been induced by the daily EMAs as potential positive triggers or observational processes.

Strengths and Limitations

To the best of our knowledge, this is the first chatbot-based intervention study including contents and assessments on interoception, as well as its association with mindfulness and stress. Further strengths of the study are the highly standardized design in line with the CONSORT (Consolidated Standards of Reporting Trials) guidelines [66,67] and EMAs of interoceptive sensibility [44,57,63]. Furthermore, the design and the usability of the chatbot were successfully tested in a previous feasibility study [35]. Therefore, the chatbot fulfills the required standards of chatbots for mental health support [6]. Finally, the results indicate the high usability of the chatbot.

Limitations of this study should be mentioned and considered for the design of future studies. First, the adherence of the intervention was relatively low at only 58%. Nevertheless, this adherence rate is on average as compared to other online mindfulness-based interventions with adherence rates ranging from 35% to 92% [68]. It should also be noted that adherence rates of digital or chatbot-based interventions were often not reported or operationalized by diverse assessments [20,69] and lack of long-term user engagement in eHealth is a common problem [70,71]. Second, there was a majority of female participants in this study, representing 77% of the sample. Therefore, future intervention studies should consider diverse strategies to specifically address male participants. Third, this study exclusively assessed self-report data. Due to potential differences to objective physiological data [72], future studies should assess both subjective and objective data, especially regarding stress and interoception. Lastly, a text- and rule-based chatbot as used in this study might lack human-like characteristics, such as those regarding the type of interaction between the chatbot and the user. Recent meta-analyses [73,74] showed that chatbot-based studies are more effective when diverse input and output modalities are combined. A multimodal chatbot might be superior because it will appear to be more lively and flexible in dialogues with the user [75] and ready to interact.

Conclusions and Future Research

To gain insight into how such interventions can be improved to achieve their full potential for stress reduction, besides a longer intervention duration, specific sample groups should be considered, such as employees, diverse age groups, and clinical or subclinical populations, aiming to adapt to individual needs and preferences in everyday life. A chatbot-based intervention seems to have the potential to improve mindfulness and emotion regulation in a stressed sample. Additional factors such as the participants' social motivation regarding the guidance by the chatbot and the personality of the chatbot [70,76] would be of further interest to foster the alliance or a therapeutic relationship between the user or a patient and the chatbot. Future studies should also investigate the specific elements that have the greatest effects to improve diverse health parameters, such as psychoeducation or exercises. Future research should implement large language models to provide and further develop diverse artificial intelligence (AI) chatbots in digital mental health interventions [77,78]. Recent findings such as those showing that AI-based chatbots are more effective in clinical or subclinical populations [74] need to be considered. Nevertheless, besides the potential of AI-based chatbots for a professional mental health service, emerging reputational risks of AI-based chatbots such as safety and data privacy issues [79,80]; gender, ethnic, and socioeconomic biases [81]; limited empathy and emotional awareness as compared to a human counterpart [82]; and hallucinations [83] should be discussed extensively.

In summary, based on the numerous prospects of chatbots in the psychological and medical field, such as counselling, psychotherapy, diagnostic assessment, and interventions [23,84,85], future studies are needed to derive robust implications in these fields.

Acknowledgments

We would like to thank all students and student assistants for their contributions in the development of the chatbot, especially Julia Kegelmann, Sina Gorhan, Ria Matapurkar, and Luisa Braitsch, and the further collaboration of the project with Anna-Lena Schäling, Vanessa Runft, Lena Gilbert, and Sandra Hoppe.

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

CS initiated the study. ELME was developed by the Department of Clinical and Health Psychology and the Institute of Distributed Systems at Ulm University (lead developers CS, EM, and BE). CS, EM, BE, DS, and OP designed and planned the study. EB, DS, BE, and OP supervised the study. CS was responsible for the recruitment and conduction of the study. EM was responsible for the technical implementation of the chatbot. CS wrote the first draft of the manuscript. All authors discussed the results of the study, edited the manuscript, and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Model results for ecological momentary assessments of perceived stress "coping with things" (Table S1) and "feeling on top of things" (Table S2) items. Model results for interoceptive sensibility, including Interoceptive Accuracy Scale (Table S3), Body Perception Questionnaire (Table S4), and momentary assessment (Table S5). Model results for the emotion regulation suppression subfacet (Table S6).

[DOCX File , 39 KB - mental_v11i1e50454_app1.docx]

Multimedia Appendix 2

CONSORT-EHEALTH checklist (V 1.6.1).

[PDF File (Adobe PDF File), 381 KB - mental_v11i1e50454_app2.pdf]

References

1. Salari N, Hosseinian-Far A, Jalali R, Vaisi-Raygani A, Rasoulpoor S, Mohammadi M, et al. Prevalence of stress, anxiety, depression among the general population during the COVID-19 pandemic: a systematic review and meta-analysis. Global Health 2020 Jul 06;16(1):57 [FREE Full text] [doi: 10.1186/s12992-020-00589-w] [Medline: 32631403]
2. Zhang Y, Xue J, Huang Y. A meta-analysis: internet mindfulness-based interventions for stress management in the general population. Medicine 2020 Jul 10;99(28):e20493 [FREE Full text] [doi: 10.1097/MD.0000000000020493] [Medline: 32664060]
3. Spijkerman MPJ, Pots WTM, Bohlmeijer ET. Effectiveness of online mindfulness-based interventions in improving mental health: a review and meta-analysis of randomised controlled trials. Clin Psychol Rev 2016 Apr;45:102-114 [FREE Full text] [doi: 10.1016/j.cpr.2016.03.009] [Medline: 27111302]
4. Nguyen-Feng VN, Greer CS, Frazier P. Using online interventions to deliver college student mental health resources: evidence from randomized clinical trials. Psychol Serv 2017 Nov;14(4):481-489. [doi: 10.1037/ser0000154] [Medline: 29120206]
5. Ma Y, She Z, Siu AF, Zeng X, Liu X. Effectiveness of online mindfulness-based interventions on psychological distress and the mediating role of emotion regulation. Front Psychol 2018;9:2090 [FREE Full text] [doi: 10.3389/fpsyg.2018.02090] [Medline: 30429816]
6. Kretzschmar K, Tyroll H, Pavarini G, Manzini A, Singh I, NeurOx Young People's Advisory Group. Can your phone be your therapist? Young people's ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support. Biomed Inform Insights 2019;11:1178222619829083 [FREE Full text] [doi: 10.1177/1178222619829083] [Medline: 30858710]
7. Musiat P, Johnson C, Atkinson M, Wilksch S, Wade T. Impact of guidance on intervention adherence in computerised interventions for mental health problems: a meta-analysis. Psychol Med 2022 Jan;52(2):229-240. [doi: 10.1017/S0033291721004621] [Medline: 34802474]
8. Domhardt M, Geßlein H, von Rezori RE, Baumeister H. Internet- and mobile-based interventions for anxiety disorders: a meta-analytic review of intervention components. Depress Anxiety 2019 Mar;36(3):213-224. [doi: 10.1002/da.22860] [Medline: 30450811]

9. Ebert DD, Buntrock C, Lehr D, Smit F, Riper H, Baumeister H, et al. Effectiveness of web- and mobile-based treatment of subthreshold depression with adherence-focused guidance: a single-blind randomized controlled trial. Behav Ther 2018 Jan;49(1):71-83. [doi: 10.1016/j.beth.2017.05.004] [Medline: 29405923]
10. Adamopoulou E, Moussiades L. Chatbots: history, technology, and applications. Machine Learn Applic 2020 Dec;2:100006. [doi: 10.1016/j.mlwa.2020.100006]
11. Brandtzaeg P, Følstad A. Why people use chatbots. 2017 Presented at: Internet Science. 4th International Conference, INSCI 2017; November 22-24, 2017; Thessaloniki, Greece. [doi: 10.1007/978-3-319-70284-1_30]
12. Bendig E, Erb B, Schulze-Thuesing L, Baumeister H. Die nächste generation: chatbots in der klinischen psychologie und psychotherapie zur förderung mentaler gesundheit – ein scoping-review. Verhaltenstherapie 2019 Mar 27;29(4):266-280. [doi: 10.1159/000499492]
13. Hill J, Randolph Ford W, Farreras IG. Real conversations with artificial intelligence: a comparison between human–human online conversations and human–chatbot conversations. Comput Hum Behav 2015 Aug;49:245-250. [doi: 10.1016/j.chb.2015.02.026]
14. Müschenich M, Wamprecht L. Wie gehts uns denn morgen? [Health 4.0 - how are we doing tomorrow?]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2018 Mar;61(3):334-339. [doi: 10.1007/s00103-018-2702-6] [Medline: 29411045]
15. Gamble A. Artificial intelligence and mobile apps for mental healthcare: a social informatics perspective. Aslib J Inf Manag 2020 Jun 02;72(4):509-523 [FREE Full text] [doi: 10.1108/ajim-11-2019-0316]
16. Stieger M, Nißen M, Rüegger D, Kowatsch T, Flückiger C, Allemand M. PEACH, a smartphone- and conversational agent-based coaching intervention for intentional personality change: study protocol of a randomized, wait-list controlled trial. BMC Psychol 2018 Sep 04;6(1):43 [FREE Full text] [doi: 10.1186/s40359-018-0257-9] [Medline: 30180880]
17. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry 2019 Jul;64(7):456-464 [FREE Full text] [doi: 10.1177/0706743719828977] [Medline: 30897957]
18. Suganuma S, Sakamoto D, Shimoyama H. An embodied conversational agent for unguided internet-based cognitive behavior therapy in preventative mental health: feasibility and acceptability pilot trial. JMIR Ment Health 2018 Jul 31;5(3):e10454 [FREE Full text] [doi: 10.2196/10454] [Medline: 30064969]
19. Gaffney H, Mansell W, Tai S. Conversational agents in the treatment of mental health problems: mixed-method systematic review. JMIR Ment Health 2019 Oct 18;6(10):e14166 [FREE Full text] [doi: 10.2196/14166] [Medline: 31628789]
20. Vaidyam AN, Linggonegoro D, Torous J. Changes to the psychiatric chatbot landscape: a systematic review of conversational agents in serious mental illness. Can J Psychiatry 2021 Apr;66(4):339-348 [FREE Full text] [doi: 10.1177/0706743720966429] [Medline: 33063526]
21. Potts C, Lindström F, Bond R, Mulvenna M, Booth F, Ennis E, et al. A multilingual digital mental health and well-being chatbot (ChatPal): pre-post multicenter intervention study. J Med Internet Res 2023 Jul 06;25:e43051 [FREE Full text] [doi: 10.2196/43051] [Medline: 37410537]
22. Ly KH, Ly A, Andersson G. A fully automated conversational agent for promoting mental well-being: a pilot RCT using mixed methods. Internet Interv 2017 Dec;10:39-46 [FREE Full text] [doi: 10.1016/j.invent.2017.10.002] [Medline: 30135751]
23. Bendig E, Erb B, Schulze-Thuesing L, Baumeister H. The next generation: chatbots in clinical psychology and psychotherapy to foster mental health – a scoping review. Verhaltenstherapie 2019 Aug 20;32(Suppl 1):64-76. [doi: 10.1159/000501812]
24. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018 Sep 01;25(9):1248-1258 [FREE Full text] [doi: 10.1093/jamia/ocy072] [Medline: 30010941]
25. Abd-Alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M. An overview of the features of chatbots in mental health: a scoping review. Int J Med Inform 2019 Dec;132:103978 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.103978] [Medline: 31622850]
26. Khalsa SS, Adolphs R, Cameron OG, Critchley HD, Davenport PW, Feinstein JS, Interoception Summit 2016 participants. Interoception and mental health: a roadmap. Biol Psychiatry Cogn Neurosci Neuroimaging 2018 Jun;3(6):501-513 [FREE Full text] [doi: 10.1016/j.bpsc.2017.12.004] [Medline: 29884281]
27. Craig AD. How do you feel? Interoception: the sense of the physiological condition of the body. Nat Rev Neurosci 2002 Aug;3(8):655-666. [doi: 10.1038/nrn894] [Medline: 12154366]
28. Fischer D, Berberich G, Zaudig M, Krauseneck T, Weiss S, Pollatos O. Interoceptive processes in anorexia nervosa in the time course of cognitive-behavioral therapy: a pilot study. Front Psychiatry 2016;7:199 [FREE Full text] [doi: 10.3389/fpsyt.2016.00199] [Medline: 28018249]
29. Eggart M, Lange A, Binser MJ, Queri S, Müller-Oerlinghausen B. Major depressive disorder is associated with impaired interoceptive accuracy: a systematic review. Brain Sci 2019 Jun 06;9(6):131 [FREE Full text] [doi: 10.3390/brainsci9060131] [Medline: 31174264]
30. Schultchen D, Bayer J, Kühnel J, Melchers KG, Pollatos O. Interoceptive accuracy is related to long-term stress via self-regulation. Psychophysiology 2019 Oct 23;56(10):e13429. [doi: 10.1111/psyp.13429] [Medline: 31231829]

31. Bornemann B, Singer T. Taking time to feel our body: steady increases in heartbeat perception accuracy and decreases in alexithymia over 9 months of contemplative mental training. Psychophysiology 2017 Mar;54(3):469-482. [doi: 10.1111/psyp.12790] [Medline: 27925645]
32. Parkin L, Morgan R, Rosselli A, Howard M, Sheppard A, Evans D, et al. Exploring the relationship between mindfulness and cardiac perception. Mindfulness 2013 Feb 5;5(3):298-313. [doi: 10.1007/s12671-012-0181-7]
33. Balaskas A, Schueller SM, Cox AL, Doherty G. Ecological momentary interventions for mental health: a scoping review. PLoS One 2021;16(3):e0248152 [FREE Full text] [doi: 10.1371/journal.pone.0248152] [Medline: 33705457]
34. Schillings C, Meissner D, Erb B, Schultchen D, Bendig E, Pollatos O. A chatbot-based intervention with ELME to improve stress and health-related parameters in a stressed sample: study protocol of a randomised controlled trial. Front Digit Health 2023 Mar 1;5:1046202 [FREE Full text] [doi: 10.3389/fdgth.2023.1046202] [Medline: 36937250]
35. Feasibility study of a chatbot-based training for stress reduction and health improvement. German Clinical Trials Register. URL: https://drks.de/search/en/trial/DRKS00025446 [accessed 2024-05-14]
36. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385. [doi: 10.2307/2136404]
37. Walach H, Buchheld N, Buttenmüller V, Kleinknecht N, Schmidt S. Measuring mindfulness—the Freiburg Mindfulness Inventory (FMI). Person Individ Diff 2006 Jun;40(8):1543-1555. [doi: 10.1016/j.paid.2005.11.025]
38. Murphy J, Brewer R, Plans D, Khalsa SS, Catmur C, Bird G. Testing the independence of self-reported interoceptive accuracy and attention. Q J Exp Psychol 2020 Jan;73(1):115-133. [doi: 10.1177/1747021819879826] [Medline: 31519137]
39. The Body Perception Questionnaire. Traumatic Stress Research Consortium. URL: https://www.traumascience.org/body-perception-questionnaire [accessed 2024-05-13]
40. World Health Organization. WHO Info Package: Mastering depression in Primary Care. Frederiksborg: World Health Organization, Regional Office for Europe, Psychiatric Research Unit, Frederiksborg General Hospital; 1998.
41. Topp CW, Østergaard SD, Søndergaard S, Bech P. The WHO-5 Well-Being Index: a systematic review of the literature. Psychother Psychosom 2015;84(3):167-176 [FREE Full text] [doi: 10.1159/000376585] [Medline: 25831962]
42. Abler B, Kessler H. Emotion Regulation Questionnaire – Eine deutschsprachige Fassung des ERQ von Gross und John. Diagnostica 2009 Jul;55(3):144-152. [doi: 10.1026/0012-1924.55.3.144]
43. Gross JJ, John OP. Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. J Pers Soc Psychol 2003 Aug;85(2):348-362. [doi: 10.1037/0022-3514.85.2.348] [Medline: 12916575]
44. Singer T, Kok B, Bornemann B, Zurborg S, Bolz M, Bochow C. The ReSource Project: Background, Design, Samples, and Measurements. Second edition. Leipzig, Germany: Max Planck Institute for Human Cognitive and Brain; 2016.
45. Schandry R. Heart beat perception and emotional experience. Psychophysiology 1981 Jul;18(4):483-488 [FREE Full text] [doi: 10.1111/j.1469-8986.1981.tb02486.x] [Medline: 7267933]
46. Zhou L, Bao J, Setiawan IMA, Saptono A, Parmanto B. The mHealth App Usability Questionnaire (MAUQ): development and validation study. JMIR Mhealth Uhealth 2019 Apr 11;7(4):e11500 [FREE Full text] [doi: 10.2196/11500] [Medline: 30973342]
47. Bendig E, Braun L, Simon L. Dropout Reasons Questionnaire for Internet Interventions (DRQi) - Questionnaire for the Systematic Recording of Dropout Reasons at Different Stages in the Implementation of an Online Intervention. Ulm, Germany: Department of Clinical Psychology and Psychotherapy, University of Ulm; 2020.
48. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Soft 2015;67(1):1-48 [FREE Full text] [doi: 10.18637/jss.v067.i01]
49. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: tests in linear mixed effects models. J Stat Soft 2017;82(13):1-26 [FREE Full text] [doi: 10.18637/jss.v082.i13]
50. Rights JD, Sterba SK. Quantifying explained variance in multilevel models: an integrative framework for defining R-squared measures. Psychol Methods 2019 Jun;24(3):309-338. [doi: 10.1037/met0000184] [Medline: 29999378]
51. Finch W, Bolin J, Kelley K. Multilevel Modeling Using R. Second edition. Boca Raton, FL: CRC Press; 2019.
52. Gardiner PM, McCue KD, Negash LM, Cheng T, White LF, Yinusa-Nyahkoon L, et al. Engaging women with an embodied conversational agent to deliver mindfulness and lifestyle recommendations: a feasibility randomized control trial. Patient Educ Couns 2017 Sep;100(9):1720-1729 [FREE Full text] [doi: 10.1016/j.pec.2017.04.015] [Medline: 28495391]
53. Schauenburg H, Sammet I, Strack M. Course of symptom severity and prediction of outcome in inpatient psychotherapy. Z Psychosom Med Psychother 2001;47(4):380-395. [doi: 10.13109/zptm.2001.47.4.380] [Medline: 11731990]
54. Owen J, Adelson J, Budge S, Wampold B, Kopta M, Minami T, et al. Trajectories of change in psychotherapy. J Clin Psychol 2015 Sep;71(9):817-827. [doi: 10.1002/jclp.22191] [Medline: 26235730]
55. Baer RA, Carmody J, Hunsinger M. Weekly change in mindfulness and perceived stress in a mindfulness-based stress reduction program. J Clin Psychol 2012 Jul;68(7):755-765. [doi: 10.1002/jclp.21865] [Medline: 22623334]
56. Venkatesan A, Krymis H, Scharff J, Waber A. Changes in perceived stress following a 10-week digital mindfulness-based stress reduction program: retrospective study. JMIR Form Res 2021 May 25;5(5):e25078 [FREE Full text] [doi: 10.2196/25078] [Medline: 34032571]
57. Linz R, Puhlmann LMC, Engert V, Singer T. Investigating the impact of distinct contemplative mental trainings on daily life stress, thoughts and affect-Evidence from a nine-month longitudinal ecological momentary assessment study.

Psychoneuroendocrinology 2022 Aug;142:105800 [FREE Full text] [doi: 10.1016/j.psyneuen.2022.105800] [Medline: 35598494]
58. Aguilar-Raab C, Stoffel M, Hernández C, Rahn S, Moessner M, Steinhilber B, et al. Effects of a mindfulness-based intervention on mindfulness, stress, salivary alpha-amylase and cortisol in everyday life. Psychophysiology 2021 Dec;58(12):e13937. [doi: 10.1111/psyp.13937] [Medline: 34525214]
59. Fischer D, Messner M, Pollatos O. Improvement of interoceptive processes after an 8-week body scan intervention. Front Hum Neurosci 2017;11:452 [FREE Full text] [doi: 10.3389/fnhum.2017.00452] [Medline: 28955213]
60. Schillings C, Karanassios G, Schulte N, Schultchen D, Pollatos O. The effects of a 3-week heartbeat perception training on interoceptive abilities. Front Neurosci 2022 May 9;16:838055 [FREE Full text] [doi: 10.3389/fnins.2022.838055] [Medline: 35615275]
61. Garfinkel SN, Seth AK, Barrett AB, Suzuki K, Critchley HD. Knowing your own heart: distinguishing interoceptive accuracy from interoceptive awareness. Biol Psychol 2015 Jan;104:65-74 [FREE Full text] [doi: 10.1016/j.biopsycho.2014.11.004] [Medline: 25451381]
62. Murphy J, Catmur C, Bird G. Classifying individual differences in interoception: implications for the measurement of interoceptive awareness. Psychon Bull Rev 2019 Oct;26(5):1467-1471 [FREE Full text] [doi: 10.3758/s13423-019-01632-7] [Medline: 31270764]
63. Höller I, Stenzel J, Rath D, Forkmann T. Listen to your heart-ecological momentary assessment of interoceptive accuracy, awareness and sensibility: a pilot study. Int J Environ Res Public Health 2021 May 04;18(9):4893 [FREE Full text] [doi: 10.3390/ijerph18094893] [Medline: 34064438]
64. Eisenstadt M, Liverpool S, Infanti E, Ciuvat RM, Carlsson C. Mobile apps that promote emotion regulation, positive mental health, and well-being in the general population: systematic review and meta-analysis. JMIR Ment Health 2021 Nov 08;8(11):e31170 [FREE Full text] [doi: 10.2196/31170] [Medline: 34747713]
65. Bendig E, Erb B, Meißner D, Bauereiß N, Baumeister H. Feasibility of a Software agent providing a brief Intervention for Self-help to Uplift psychological wellbeing ("SISU"). A single-group pretest-posttest trial investigating the potential of SISU to act as therapeutic agent. Internet Interv 2021 Apr;24:100377 [FREE Full text] [doi: 10.1016/j.invent.2021.100377] [Medline: 33816127]
66. Eysenbach G, CONSORT-EHEALTH Group. CONSORT-EHEALTH: improving and standardizing evaluation reports of web-based and mobile health interventions. J Med Internet Res 2011 Dec 31;13(4):e126 [FREE Full text] [doi: 10.2196/jmir.1923] [Medline: 22209829]
67. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 2010 Mar 23;340:c332 [FREE Full text] [doi: 10.1136/bmj.c332] [Medline: 20332509]
68. Sommers-Spijkerman M, Austin J, Bohlmeijer E, Pots W. New evidence in the booming field of online mindfulness: an updated meta-analysis of randomized controlled trials. JMIR Ment Health 2021 Jul 19;8(7):e28168 [FREE Full text] [doi: 10.2196/28168] [Medline: 34279240]
69. Beintner I, Vollert B, Zarski A, Bolinski F, Musiat P, Görlich D, et al. Adherence reporting in randomized controlled trials examining manualized multisession online interventions: systematic review of practices and proposal for reporting standards. J Med Internet Res 2019 Aug 15;21(8):e14181 [FREE Full text] [doi: 10.2196/14181] [Medline: 31414664]
70. Koulouri T, Macredie RD, Olakitan D. Chatbots to support young adults' mental health: an exploratory study of acceptability. ACM Trans Interact Intell Syst 2022 Jul 20;12(2):1-39. [doi: 10.1145/3485874]
71. Torous J, Nicholas J, Larsen ME, Firth J, Christensen H. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. Evid Based Ment Health 2018 Aug;21(3):116-119 [FREE Full text] [doi: 10.1136/eb-2018-102891] [Medline: 29871870]
72. Zuniga Gonzalez DA, Richards D, Bilgin AA. Making it real: a study of augmented virtuality on presence and enhanced benefits of study stress reduction sessions. Int J Hum Comput Stud 2021 Mar;147:102579. [doi: 10.1016/j.ijhcs.2020.102579]
73. Lim SM, Shiau CWC, Cheng LJ, Lau Y. Chatbot-delivered psychotherapy for adults with depressive and anxiety symptoms: a systematic review and meta-regression. Behav Ther 2022 Mar;53(2):334-347. [doi: 10.1016/j.beth.2021.09.007] [Medline: 35227408]
74. Li H, Zhang R, Lee Y, Kraut RE, Mohr DC. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med 2023 Dec 19;6(1):236. [doi: 10.1038/s41746-023-00979-5] [Medline: 38114588]
75. Montenegro JLZ, da Costa CA, da Rosa Righi R. Survey of conversational agents in health. Exp Syst Appl 2019 Sep;129:56-67 [FREE Full text] [doi: 10.1016/j.eswa.2019.03.054]
76. Grové C. Co-developing a mental health and wellbeing chatbot with and for young people. Front Psychiatry 2020;11:606041 [FREE Full text] [doi: 10.3389/fpsyt.2020.606041] [Medline: 33597898]
77. Thakur S, Rastogi D, Singh L. MOODY: a natural language processing-based chatbot for mental health care. 2021 Presented at: International Conference on Sustainable Technology and Advanced Computing in Electrical Engineering (ICSTACE) 2021; November 11-12, 2021; Surat, Gujarat, India. [doi: 10.1007/978-981-19-4364-5_64]

78. Boucher EM, Harake NR, Ward HE, Stoeckl SE, Vargas J, Minkel J, et al. Artificially intelligent chatbots in digital mental health interventions: a review. Expert Rev Med Devices 2021 Dec;18(sup1):37-49. [doi: 10.1080/17434440.2021.2013200] [Medline: 34872429]
79. Li J. Security implications of AI chatbots in health care. J Med Internet Res 2023 Nov 28;25:e47551 [FREE Full text] [doi: 10.2196/47551] [Medline: 38015597]
80. De Freitas J, Uğuralp AK, Oğuz - Uğuralp Z, Puntoni S. Chatbots and mental health: insights into the safety of generative AI. J Consum Psychol 2023 Dec 19:online early. [doi: 10.1002/jcpy.1393]
81. Kim J, Cai ZR, Chen ML, Simard JF, Linos E. Assessing biases in medical decisions via clinician and AI chatbot responses to patient vignettes. JAMA Netw Open 2023 Oct 02;6(10):e2338050 [FREE Full text] [doi: 10.1001/jamanetworkopen.2023.38050] [Medline: 37847506]
82. Pham KT, Nabizadeh A, Selek S. Artificial intelligence and chatbots in psychiatry. Psychiatr Q 2022 Mar;93(1):249-253 [FREE Full text] [doi: 10.1007/s11126-022-09973-8] [Medline: 35212940]
83. Giuffrè M, You K, Shung DL. Evaluating ChatGPT in medical contexts: the imperative to guard against hallucinations and partial accuracies. Clin Gastroenterol Hepatol 2024 May;22(5):1145-1146. [doi: 10.1016/j.cgh.2023.09.035] [Medline: 37863408]
84. Pryss R, Kraft R, Baumeister H, Winkler J, Probst T, Reichert M, et al. Using chatbots to support medical and psychological treatment procedures: challenges, opportunities, technologies, reference architecture. In: Montag C, Baumeister H, editors. Digital Phenotyping and Mobile Sensing. Cham: Springer; 2019:249-260.
85. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. J Med Internet Res 2023 Feb 24;25:e40789 [FREE Full text] [doi: 10.2196/40789] [Medline: 36826990]

Abbreviations

AI: artificial intelligence
BPQ: Body Perception Questionnaire
CONSORT: Consolidated Standards of Reporting Trials
EMA: ecological momentary assessment
IAS: Interoceptive Accuracy Scale
PSS: Perceived Stress Scale
RCT: randomized controlled trial
T0: screening
T1: preintervention
T2: post intervention (3 weeks)
T3: follow-up (6 weeks)
WHO: World Health Organization

Edited by T de Azevedo Cardoso, J Torous; submitted 01.07.23; peer-reviewed by Y Bafna, I Liu, E Smoktunowicz; comments to author 19.01.24; revised version received 09.02.24; accepted 26.03.24; published 28.05.24.

Please cite as:
Schillings C, Meißner E, Erb B, Bendig E, Schultchen D, Pollatos O
Effects of a Chatbot-Based Intervention on Stress and Health-Related Parameters in a Stressed Sample: Randomized Controlled Trial
JMIR Ment Health 2024;11:e50454
URL: https://mental.jmir.org/2024/1/e50454
doi:10.2196/50454
PMID:38805259

Original Paper

Toward Tailoring Just-in-Time Adaptive Intervention Systems for Workplace Stress Reduction: Exploratory Analysis of Intervention Implementation

Jina Suh[1], PhD; Esther Howe[2], MA; Robert Lewis[3], MSc; Javier Hernandez[1], PhD; Koustuv Saha[4], PhD; Tim Althoff[5], PhD; Mary Czerwinski[6], PhD

[1]Microsoft Research, Redmond, WA, United States
[2]Idiographic Dynamics Lab, Department of Psychology, University of California, Berkeley, CA, United States
[3]MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA, United States
[4]Siebel School of Computing and Data Science, University of Illinois Urbana-Champaign, Urbana, IL, United States
[5]Paul G Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, United States
[6]Human-Centered Design and Engineering, University of Washington, Seattle, WA, United States

Corresponding Author:
Jina Suh, PhD
Microsoft Research
1 Microsoft Way
Redmond, WA, 98052
United States
Phone: 1 4254211584
Email: jinsuh@microsoft.com

Abstract

Background: Integrating stress-reduction interventions into the workplace may improve the health and well-being of employees, and there is an opportunity to leverage ubiquitous everyday work technologies to understand dynamic work contexts and facilitate stress reduction wherever work happens. Sensing-powered just-in-time adaptive intervention (JITAI) systems have the potential to adapt and deliver tailored interventions, but such adaptation requires a comprehensive analysis of contextual and individual-level variables that may influence intervention outcomes and be leveraged to drive the system's decision-making.

Objective: This study aims to identify key tailoring variables that influence momentary engagement in digital stress reduction microinterventions to inform the design of similar JITAI systems.

Methods: To inform the design of such dynamic adaptation, we analyzed data from the implementation and deployment of a system that incorporates passively sensed data across everyday work devices to send just-in-time stress reduction microinterventions in the workplace to 43 participants during a 4-week deployment. We evaluated 27 trait-based factors (ie, individual characteristics), state-based factors (ie, workplace contextual and behavioral signals and momentary stress), and intervention-related factors (ie, location and function) across 1585 system-initiated interventions. We built logistical regression models to identify the factors contributing to momentary engagement, the choice of interventions, the engagement given an intervention choice, the user rating of interventions engaged, and the stress reduction from the engagement.

Results: We found that women (odds ratio [OR] 0.41, 95% CI 0.21-0.77; *P*=.03), those with higher neuroticism (OR 0.57, 95% CI 0.39-0.81; *P*=.01), those with higher cognitive reappraisal skills (OR 0.69, 95% CI 0.52-0.91; *P*=.04), and those that chose calm interventions (OR 0.43, 95% CI 0.23-0.78; *P*=.03) were significantly less likely to experience stress reduction, while those with higher agreeableness (OR 1.73, 95% CI 1.10-2.76; *P*=.06) and those that chose prompt-based (OR 6.65, 95% CI 1.53-36.45; *P*=.06) or video-based (OR 5.62, 95% CI 1.12-34.10; *P*=.12) interventions were substantially more likely to experience stress reduction. We also found that work-related contextual signals such as higher meeting counts (OR 0.62, 95% CI 0.49-0.78; *P*<.001) and higher engagement skewness (OR 0.64, 95% CI 0.51-0.79; *P*<.001) were associated with a lower likelihood of engagement, indicating that state-based contextual factors such as being in a meeting or the time of the day may matter more for engagement than efficacy. In addition, a just-in-time intervention that was explicitly rescheduled to a later time was more likely to be engaged with (OR 1.77, 95% CI 1.32-2.38; *P*<.001).

Conclusions: JITAI systems have the potential to integrate timely support into the workplace. On the basis of our findings, we recommend that individual, contextual, and content-based factors be incorporated into the system for tailoring as well as for monitoring ineffective engagements across subgroups and contexts.

(JMIR Ment Health 2024;11:e48974) doi:10.2196/48974

KEYWORDS

workplace stress; just-in-time; just-in-time adaptive intervention; JITAI; engagement; microintervention; stress reduction; psychotherapy

Introduction

Background

Work is a major source of stress in the United States, affecting more than half of Americans throughout most of the day [1]. Workplace stress leads to increased risk of mental and physical health disorders, decreased productivity and job satisfaction, and higher rates of accidents and employee costs [2-5]. Integrating stress reduction strategies directly into the workplace has proven to be effective and is widely recommended [6,7]. However, incorporating these individual-based techniques (eg, cognitive behavioral skills training, meditation, and exercise) [7,8] into the workday can be challenging due to work culture [9,10] or psychological barriers [9-11].

Information workers who frequently use computing technology face challenges, such as prolonged desk-bound and sedentary behaviors, that contribute to chronic physical and mental health issues [12,13]. High computer use without adequate breaks, high levels of multitasking, and constant connectivity demanded by information and communications technology has been found to be associated with increased stress and burnout [14-17]. Despite information and communications technology being associated with increased stress, workplace computing tools can also be leveraged to understand and reduce stress [18,19].

Passive sensing capabilities via ubiquitous devices have shown potential in health and well-being domains through monitoring and assessing individuals over time [20], and such data may be harnessed to provide precision mental health support. However, introducing new devices (eg, wearables) to an organization can be costly (eg, ≥US $100 per worker) and could be impractical for real-world, daily functional use with compliance and quality issues [21,22]. In contrast, everyday technologies commonly used at work (eg, webcam, keyboard, and software telemetry) can be harnessed for passing sensing and behavioral analysis, offering a more feasible approach to infer affect [23], physiology [24], attention [25], or stress [18,26].

Recently, technology-mediated support for mental health has generated interest for its ability to provide flexible and always-available access. However, these systems often provide generalized support that does not account for individual variabilities or contexts [27]. Despite an abundance of mental health apps [28,29], digital interventions that fit the specific workplace context are still highly sought after [27,30].

Digital mental health interventions (DMHIs), known as digital microinterventions, leverage technology to adapt existing evidence-based psychosocial interventions and leverage technology affordances to provide individual components of traditional psychotherapy focused on managing proximal symptoms (eg, relaxation for stress) in the hopes of achieving broad, distal objectives (eg, overcoming depression) [31]. Just-in-time adaptive interventions (JITAIs) have been introduced to deliver personalized, contextualized, and adaptable interventions using dynamic human behavior data captured through ubiquitous sensing technologies [27,32-34]. This concept has been explored in various health contexts, such as promoting physical activity, stress management, and weight management, with recent interest in applying them to positive coping skill use [35].

Despite their promises, JITAI systems are not yet pervasive with many applications still relying on ecological momentary assessments (EMAs [36]). Recent developments in algorithmic and machine learning approaches to dynamic adaptation have begun to show improvements in timing, receptivity, and engagement [37-40]. Operationalizing the adaptation of the system requires choosing appropriate tailoring variables and intervention options to drive the system's decision-making regarding intervention timing and content. Prior research has examined factors associated with engagement in DMHIs [41,42], including personal-related (eg, demographics and personalities), content-related (eg, perceived fit and usefulness), and technology-related factors (eg, technical issues and privacy). However, most prior studies investigate study-long engagement, rather than examining "in-the-moment" engagement factors, which are crucial for improving the usability of interventions in real-world contexts [31], particularly because the integration of intervention use into life is a core facilitator for engagement [42].

A recent meta-analysis [36] has also shown that tailoring is an important aspect of JITAI design associated with greater efficacy. Despite JITAI systems' potential to provide precision support [27], few studies demonstrate the value of just-in-time (JIT) support in improving user engagement [43,44]. Although the use of passively sensed data for contextual understanding and system adaptation is often recommended for the design of JITAI systems [27], many still rely mostly on EMAs, app use, or simple temporal features primarily from a single modality (ie, mobile devices) or without considering individual or intervention-related factors [36,44]. Passive sensing technologies offer numerous sets of contextual variables (eg, location, calendar, movement, and activity), presenting a challenge in designing sensing-capable JITAI systems: identifying a core set of tailoring variables among all possible variables that the system should consider for optimizing effective engagement [45]. Therefore, our primary goal and key contribution is to identify crucial tailoring variables that influence momentary engagement in digital stress-reduction microinterventions to inform the design of similar JITAI systems.

Summary

In this study, we analyze engagement data from a 4-week longitudinal deployment of a workplace stress-reduction intervention system [19]. We leverage everyday workplace devices as unobtrusive and passive sensors to gain a glimpse into participants' daily work activities (eg, emails, meetings, and computer activity). This system leverages the Cloud to integrate passively collected data and EMAs to deliver JIT nudges to engage in digital microinterventions across devices. Unlike laboratory experiments or controlled studies, our study allows observing users' moment-by-moment interactions with the system in naturalistic work environments. We combine passively sensed work contexts, the system use, including which intervention participants chose and liked, individual demographics, personality traits, and coping styles to understand how the work context influences engagement patterns and to understand the appropriate conditions that lead to momentary intervention engagement and positive outcomes. We leverage such data from 43 participants to contextualize 1585 system-initiated interventions. From statistical modeling of the impact of individual, contextual, and intervention-related factors on engagement outcomes, we confirmed that individual factors (eg, age, gender, personality traits, and coping skills), as well as contextual and content-related factors (eg, availability and intervention modality), significantly influenced momentary intervention engagement, intervention choice, user ratings, and stress reduction outcomes. These findings suggest tailoring guidelines for JITAI systems whereby contextual and personalized factors can be used to find a positive balance between user preferences and maximal intervention efficacy.

In preparing this paper, we referred to the Guidelines and Checklist for the Reporting on Digital Health Implementations (iCHECK-DH) [46], which can be found in Multimedia Appendix 1.

Methods

Objectives

There are multiple considerations that influence JITAI systems' decision points (ie, "a time at which an intervention decision is made" [32]), such as the right timing for a prompt, the right intervention for the moment, or the intervention likely to be engaged in. Therefore, this study aimed to identify factors that may contribute to improving participant engagement for JITAI systems. The study builds on a pilot implementation and deployment of a JIT stress-reduction microinterventions in a real-world workplace setting, targeting information workers who spend most of their working hours at processing information with computing devices [47]. The deployed system did not adapt the intervention content based on changing state or user context and, therefore, is not a full JITAI system. Instead, the study conducted a retrospective analysis of the deployment data to understand factors that influence the engagement and efficacy of workplace stress-reduction JITAI systems for future development.

Participants

Information workers were recruited from a large technology organization via randomly distributed email advertisements. We recruited primarily US-based workers for the ease of system troubleshooting and to minimize any country-specific organizational factors. We enrolled participants on a first-come, first-served basis as long as we could satisfy system compatibility. To ensure system compatibility, interested participants completed a brief screener survey about their work setup (eg, primary device specification and operating system and web camera availability). Eligible participants, whose primary device specification met our sensing software requirements, were asked to install and run the study system on their primary desktop for 30 minutes. Only 43 participants that could run the sensing software on their desktop for ≥30 minutes were selected to participate in the study. The sample size of approximately 40 participants was determined for statistical power within a larger experimental study [19] that compared the JIT condition with a baseline condition for intervention effectiveness.

Our intake survey included several demographics measures such as age, gender, and role. Of the 43 participants, 29 (67%) identified as man. This distribution of gender closely aligns with the current industry demographics for large technology companies [48], and therefore, we consider this gender representation acceptable for our analysis. Of the 43 participants, 3 (7%) self-reported being aged 18 to 25 years old, 11 (25%) were 26 to 35 years old, 18 (42%) were 36 to 45 years old, 8 (19%) were 46 to 55 years old, 2 (5%) were 56 to 65 years old, and 1 (2%) was >66 years old. Moreover, 24 (56%) of 43 participants reported as being in engineering or development role, and 14 (33%) reported as being in sales or business strategy.

Other intake measures included the Depression, Anxiety, and Stress Scale-21 (DASS-21) [49]), the brief Big Five Personality Inventory-10 [50], the Emotional Regulation Questionnaire [51] designed to measure the tendency to regulate emotions through cognitive reappraisal and expressive suppression, and the 6-item brief resilience scale [52] that measures the ability to bounce back from stress. The average stress level of the stress subscale of DASS-21 reported by participants was 5 out of 21 (SD 3.8) which is within normal ranges. Our participants scored an average of 3.8 for agreeableness, 4.1 for conscientiousness, 2.6 for extraversion, 2.8 for neuroticism, and 3.5 for openness out of 5. Cognitive reappraisal was scored at 4.7 out of 7 (SD 1.1), expressive suppression was at 3.7 out of 7 (SD 1.3), and resilience was at 3.5 out of 5 (SD 0.9) on average.

Study Implementation and Procedure

We deployed our system on the 43 participants who consented to the study, for a period of 4 weeks. During every workday of the 4-week study period, the system asked users to complete 5 EMAs per day during their reported work hours to capture their subjective stress ratings from the past 30 minutes (ie, "How would you rate your level of stress during the last 30 minutes?"). When the system determined that the user's stress level may be high, the system sent JIT nudges via a chatbot, asking users to engage in a stress reduction microintervention. In the background, the system captured use data as well as passively sensed contextual data. Detailed description of the system architecture is provided in section 1 of Multimedia Appendix 2 [5,15,19,33,53-72].

JIT Heuristics

The system determined higher-than-baseline stress levels based on our JIT heuristics informed by computed stress scores and self-reported stress levels. Stress scores were computed in real time per individual as an average of 5 components ranging between 0 and 1, each representing 5 components that previous work has identified as sources of stress: (1) the number of emails received [15,53], (2) the total number of meetings in a given day [54], (3) the percentage of time into the day [55], (4) the amount of facial expressions (via the Facial Action Coding System [56]) from corrugator (ie, brow furrowing) and lip depressor (ie, frowning) minus zygomatic major (ie, smiling) [55,57,58], and (5) heart rate [59,60]. Self-reported stress levels were obtained from EMAs.

In our JIT heuristics, first, we compute each user's baselines as the average of the computed stress scores and self-reported stress levels based on the data from the first week of using the system. These individualized baselines (captured at week 1 of the 4-week study) are used as thresholds for delineating high stress from low stress during subsequent weeks (weeks 2 to 4). During the first week, we use the default baseline at the middle of the score range. Then, we send intervention nudges only if it is during the working hours that the users have stated at intake, if they have not explicitly scheduled an intervention at a later time that day, if they have not completed an intervention in the past hour, if there has not been an intervention nudge in the past 2 hours, and if there have not been ≥4 intervention nudges that day. Detailed description of our stress inference and heuristics is provided in section 1 of Multimedia Appendix 2.

Microinterventions

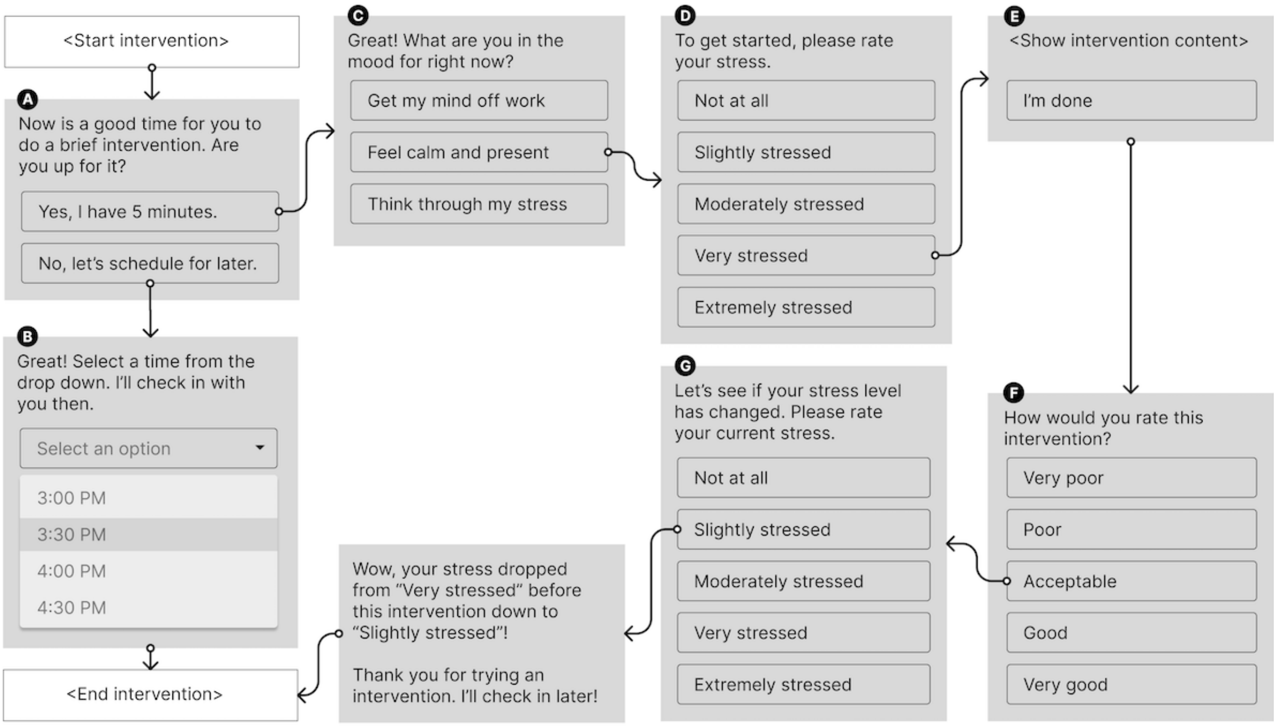
Microinterventions [31] used in our study were translated from components of cognitive behavioral therapy and dialectical behavioral therapy, 2 empirically supported and widely used psychotherapy modalities [61,62]). These were under 5-minute interventions that were either a short video, a single-turn text prompt, or a brief therapeutic conversation with the chatbot. The microinterventions used in our study can be categorized by (1) the function served for users, (2) the modality in which the intervention was delivered, and (3) the intended location to perform the intervention.

Microinterventions are primarily categorized into 3 functional categories that align approximately with the amount of effort required. "Get my mind off work" interventions are low-effort interventions designed to help users take their mind away from work with positive activities that promote emotion regulation [63]. "Feel calm and present" interventions are medium-effort interventions that help users feel calm and present by drawing inspiration from the mindfulness practices. "Think through my stress" interventions are high-effort interventions that help users think through their stress and directly address and resolve stress-inducing components of their lives. For simplicity, we refer to these 3 intervention categories as "distract," "calm," and "address," respectively, in the rest of the paper. Overall, there were 18 interventions per functional category. Section 2 of Multimedia Appendix 2 provides more details about the microinterventions, including all categories and examples for each category.

User Engagement Flow

Figure 1 illustrates a series of user engagement steps when a system nudge is sent. When the system sends a nudge to perform an intervention (Figure 1A), users can choose to delay it to a later time that day (Figure 1B). The system does not send another nudge until that time. Users may ignore the nudge, and the system expires the nudge after 30 minutes of inactivity. If the user decides to engage in an intervention, they can choose from 3 intervention categories (Figure 1C), and then the system randomly selects a new intervention within the category. Before beginning the intervention, the users are asked to subjectively rate their current stress level (Figure 1D). Then, the intervention content is shown to the user (Figure 1E). Once the intervention is done, the users are asked to rate the intervention (Figure 1F) and rate their stress level again (Figure 1G) before concluding the intervention flow. An example screenshot of the full user engagement flow can be found in section 3 of Multimedia Appendix 2.

Figure 1. (A) System sends a nudge to users to perform an intervention. (B) Users can opt to postpone the intervention at a later time. (C) If users choose to do an intervention, they can select from 1 of 3 intervention categories. (D) Users first self-report their current stress level. (E) The system shows the intervention content for users to interact with. (F) User rates the intervention. (G) Users self-report their stress levels after the intervention.



Other Study Tasks

Beyond JIT-based interactions with the system, participants could also access interventions on demand, where they could perform the intervention at that moment or schedule it to a later time that day. Participants were also asked to complete morning surveys to log their sleep quality and evening surveys to log their food and beverage intake. We additionally asked participants to rate their sleep quality the night before via a morning survey and log their food and drink consumptions throughout the day via an evening survey. They also completed weekly surveys including the DASS-21 and the brief resilience scales. The exit survey solicited feedback about the usability of the system and the perceived helpfulness and impact of the interventions.

Poststudy Analysis

This section describes how we processed and analyzed the data collected from the above study to understand the factors contributing to improving participant engagement in a system-initiated intervention and the effectiveness of interventions at a given moment. Table 1 describes the full set of variables selected for our analysis and their descriptive statistics. Section 3 of Multimedia Appendix 2 includes additional detail, including descriptive statistics, inclusion or exclusion of data points, and correlational analyses conducted between variables.

Table 1. List of the per-participant, per-half hour, per-nudge, and per-intervention variables and their descriptive statistics.

Variable	Distribution
Per-participant	
Age (y)	18-35 (n=14), 36-45 (n=18), >46 (n=11)
Gender	Men (n=29), women (n=14)
Cognitive reappraisal	Mean 4.69 (SD 1.09); range 2-7
Expressive suppression	Mean 3.74 (SD 1.26); range 2-6
Resilience	Mean 3.51 (SD 0.89); range 2-5
Agreeableness	Mean 3.79 (SD 0.74); range 2-5
Conscientiousness	Mean 4.12 (SD 0.83); range 2-5
Extraversion	Mean 2.62 (SD 0.86); range 1-4
Neuroticism	Mean 2.85 (SD 1.04); range 1-5
Openness	Mean 3.48 (SD 0.79); range 2-5
Engagement skewness	Mean –0.09 (SD 0.58); range –1.41 to 1.41
Per-half hour	
Nudge probability	Mean 0.06 (SD 0.04); range 0.02-0.30
Per-nudge	
Meeting counts	Mean 0.30 (SD 0.51); range 0-3
No meeting minutes	Mean 5.43 (SD 11.79); range 0-50
Self-event counts	Mean 0.12 (SD 0.37); range 0-3
Email messages sent	Mean 0.39 (SD 0.88); range 0-8
Email messages read	Mean 3.22 (SD 5.09); range 0-54
Chat messages count	Mean 3.88 (SD 5.90); range 0-64
Ad hoc call count	Mean 0.04 (SD 0.21); range 0-2
Number of attention signals	Mean 1442.64 (SD 1316.68); range 0-5705
Nudge source	JIT[a] algorithm (n=1337), rescheduled (n=248)
Engaged	True (n=563), False (n=1022)
Per-intervention (chosen)	
Category	Address (n=112), calm (n=338), distract (n=200)
Engaged	True (n=563), false (n=87)
Modality	Video (n=113), prompt (n=384), conversation (n=24)
Location	At desk (n=422), inside (n=84), outside (n=15)
Per-intervention (completed)	
Stress reduction	Mean 0.29 (SD 0.53); range –1 to 3
Rating	Mean 3.61 (SD 1.02); range 1-5
Improved	True (n=150), false (n=371)
Liked	True (n=289), false (n=232)
Stress before	Mean 2.14 (SD 0.97); range 1-5
Category	Address (n=77), calm (n=275), distract (n=72)
Modality	Video (n=60), prompt (n=340), conversation (n=24)
Location	At desk (n=325), inside (n=84), outside (n=15)

[a]JIT: just-in-time.

Engagement Outcome Variables

From participants' interactions with the system, we reconstructed each participant's step-by-step interaction with the system as represented in Figure 1 to extract several outcome variables, such as whether a system-initiated intervention was completed and the effectiveness and rating of interventions once engaged. We note that the number of data points varies based on where in the engagement flow the participant exits.

Engagement Label

We labeled each system-initiated intervention as "engaged" in an intervention (ie, *engaged*=true) if the participant explicitly marked the intervention as done (ie, clicking on "I'm done" button in Figure 1E), regardless of whether they completed any subsequent prompts (ie, Figures 1F and G). Any ignored, incomplete, or timed-out nudges were considered not engaged. Each system-initiated intervention that was triggered at a postponed time (Figure 1B) was categorized as "rescheduled" (ie, *nudge source*=rescheduled). Although these rescheduled nudges looked identical to JIT nudges, we hypothesized that the participants would be more likely to engage in a system-initiated intervention if they postponed the intervention to a time that is more suitable for engagement.

Intervention Choice

The characteristics of interventions, such as location, modality, and effort level, are important to consider because certain interventions might not be feasible in certain situations (eg, the participant cannot go outside) [73,74]. If a participant chose an intervention category (Figure 1C), we marked that nudge with binary labels of *distract chosen*, *calm chosen*, or *address chosen*. Each of these interventions was further labeled with *modality* and *location* based on the specific intervention that the system chose within the category.

Intervention Effectiveness and Rating

We looked at 2 outcome metrics—momentary *stress reduction* and intervention *rating*. We further binarized these outcome metrics to determine if a certain intervention engagement *improved* stress (ie, self-reported stress rating was lowered after the intervention use) and if the participant *liked* the intervention (ie, rated as "good" or "very good").

Tailoring Variables for Postdeployment Analysis

We used per-participant demographics and validated scale responses (coping skills and personality traits) to explain individual differences. We harnessed passively sensed data streams to explain the context surrounding the system initiation and intervention engagement. On the basis of the participants' interactions with the system, we also account for the probability distribution of system nudges throughout the day and the participant's likelihood of engaging in interventions in the morning or the afternoon.

Participant Characteristics

We included demographic variables, such as age and gender, and personality traits that have been shown to impact engagement [42]. Because of the small sample sizes on either end of the age groups, we combined the lower 2 and upper 3 age groups to create more balanced age groups (ie, 18-35, 36-45, and >46 years). We included emotion regulation [51] and resilience [52] as measures of a person's ability to cope with stressors. We also included the Big Five personality traits [75-77] because they are known to impact stress [78,79] and engagement in mental health treatment [42].

Passively Sensed Context

To understand the context surrounding a system-initiated intervention, we leveraged 2 sources of passively sensed data: custom sensing software built by Microsoft Research and Viva Insights by Microsoft.

The custom sensing software ran on participants' desktops and can capture activities that may not be associated with work, such as browsing the internet or using non–work related software. It captured data from the 5 components used to compute stress scores in real time, as described earlier. The sensing software also captured general user computer activity events such as mouse and keyboard interactions into a single metric, *number of attention signals*, which could be an important indicator for presence. We hypothesized that presence at the computer could lead to higher engagement in interventions as the nudges were designed to grab the attention of participants at work. The range of values for *number of attention signals* was fairly large (maximum=5705) compared with other variables, so we divided the values by 2000 to estimate a comparable coefficient and CIs during modeling (ie, to have odds ratios [ORs] within 2 decimal points). When interpreting the effect sizes, we corrected for this factor of 2000. We also hypothesized that the likelihood of engaging in a stress-reduction intervention during active participation in a meeting is low. Therefore, we included *no meeting minutes* to represent the total number of minutes without a scheduled meeting with others and *self-event count* as the total number of calendar events with only the participant as the attendee.

Viva Insights captures deidentified activity aggregates in 30-minute windows for Microsoft tools across all devices associated with an individual's work account. From Viva Insights, we included *meeting counts*, *ad hoc call count*, and *email messages sent or read*. We excluded *chat messages count* from our analysis because the nudges were delivered through Microsoft Teams, and our data source cannot be used to discern if the messages were coming from the bot. Because Viva Insights data are limited to half-hourly windows, we associated the contextual metrics with each system-initiated intervention by taking the half-hour window that holds the nudge time stamp.

Both data sources had several overlaps or similar metrics, such as ones related to meetings or emails. Correlational analysis between 2 data sources is described in section 3 of Multimedia Appendix 2. It is important to note that contextual data at work can be noisy due to individual differences in the use of work tools for nonwork purposes (eg, personal use and subscriptions). We do not differentiate work versus nonwork data because it is challenging, for example, to isolate work-related emails as the only source of stress.

System-Initiated Intervention Probability

Although each of the system nudges could be considered as an independent, repeated observation, the timing of the system-initiated interventions was sometimes dependent on when EMAs were administered. Because the system's JIT heuristic runs every 5 minutes to check if a nudge needs to be sent based on the stress score and the EMA stress ratings, the most likely hours for receiving a system nudge is shortly after the EMA, leading to each participant receiving more nudges during certain hours of the day than others. To account for such variability in receiving system-initiated interventions, we incorporate the momentary nudge probability in our analysis. Because the range of nudge probabilities is small (mean 0.06, SD 0.04), we multiply the measure by 100 to represent it in percentages.

Temporal Engagement Skewness

Prior research has found that different hours of the day were seen as good or not-so-good timing for stress interventions [33]. To examine if a certain participant has a temporal tendency to engage, we computed the Fisher-Pearson coefficient of skewness, or *engagement skewness*, on the hourly intervention engagements per participant. A positive *engagement skewness* means that the participants tend to engage at the beginning of their workday, and a negative *engagement skewness* means that the participants tend to engage toward the end of their workday. We used the skewness metric instead of simply looking at the engagement during the morning and the afternoon to account for individual differences in working hours. We incorporated this skewness per participant in our analysis.

Analysis

Taking all the tailoring and outcome variables into account, our analysis focused on estimating the effects of contextual, individual, and intervention characteristics on binary outcome variables (eg, *engaged*, *distract chosen*, and *liked*). Thus, we built a logistic regression model predicting each outcome based on a combination of per-participant characteristics, per-half-hour nudge probability, per-nudge contextual metrics, or per-intervention metrics as fixed effects. The outputs of the logistical regression models are presented as ORs, representing a ratio of odds (eg, probability of engaging vs probability of not engaging) under 2 different conditions (eg, being a woman vs not being a woman). The data processing was conducted using Python packages (eg, *numpy*, *pandas*, *scipy*, and *seaborn*) and the models were tested using R libraries (eg, *lme4*, *car*, and *performance*).

We determined the significance of the fitted model against the null hypothesis model using the analysis of deviance. We conducted ANOVA to estimate the significance of fixed effects. Because the data are unbalanced (ie, unequal number of observations for each level of a factor), we obtain ANOVA type II sums of squares [80]. For categorical variables of ≥3 levels (eg, *age group* and *category*), we estimated pairwise differences using Tukey HSD (honestly significant difference) procedure. Multicollinearity in fixed effects was tested using the variance of inflation factor, and none of our models exhibited a multicollinearity issue. Results for ANOVA, multicollinearity, and Tukey analyses are presented in section 4 of Multimedia Appendix 2.

Ethical Considerations

The study was reviewed by the Microsoft Research Institutional Review Board (OHRP IORG #0008066, IRB #IRB00009672) before the research activities and was formally approved. In addition to the ethics review, our study obtained approvals from Microsoft's privacy, security, and legal review officers before conducting the study. All participants provided consent as part of the onboarding process and agreed for their deidentified data to be used for research purposes. The consent described the installation of our system, interaction with interventions, surveys, and joining the study data with cloud-based, device-independent telemetry data. All data, collected anonymously or otherwise, were collected and stored in a secured and access-controlled location. All data were joined and deidentified before analysis by the research team. For their participation and data, each participant was compensated with a US $400 Amazon gift card.

Implementation Considerations

The implementation of the system used in the deployment study has dependencies that restrict interoperability and sustainability. The system requires that the participants use Microsoft platforms (eg, M365, Windows, Viva Insights, and Teams) and own a decent desktop that can perform vision-based data processing on device. Our current implementation focused on the understanding of individual, contextual, and intervention-related factors to drive the tailoring of JITAI systems and does not allow generalizability beyond the supported architecture. However, we believe that recent advancements in generative artificial intelligence technologies and the implementation of interoperability layers across different technology ecosystems can enable more accessible implementation of the system. Because this study was a pilot implementation, any budget planning, sustainability model, or interoperability for sustained deployment were not in scope for this paper.

Results

Intervention Engagement

To see which factors influenced engagement, we modeled *engaged* as a function of per-participant characteristics, per-half hour nudge probability, and per-nudge contextual metrics. The logistic regression model of *engaged* with all fixed effects was significantly different from the null hypothesis model (ie, engage ~1; χ^2_{21}=115.5; P<.001). Table 2 outlines the ORs and CIs for each predictor. Predictors *engagement skewness*, *meeting count*, *number of attention signals*, and *nudge source* remained significant after applying the Benjamini-Hochberg correction for multiple comparisons. Reviewing the coefficients of our fixed effects, we found that being aged >46 years, being a woman, higher *cognitive reappraisal*, higher *number of attention signals*, and receiving a rescheduled nudge were associated with a higher likelihood of intervention engagement. In contrast, higher *engagement skewness* and *meeting count* were associated with to a lower likelihood of intervention engagement.

Table 2. Odds ratios (ORs) and CIs for each predictor of *engaged* for all system-initiated nudges[a].

Predictors	Engaged, OR (95% CI)
Intercept	0.1[b] (0.03-0.30)
Age group (y), reference (18-35)	
36-45	1.17 (0.88-1.56)
>46	1.47[c] (1.07-2.02)
Gender reference (man)	
Gender (woman)	1.37[c] (1.05-1.79)
Cognitive reappraisal	1.14[c] (1.01-1.29)
Expressive suppression	0.92 (0.83-1.02)
Resilience	1.05 (0.88-1.25)
Agreeableness	0.87 (0.72-1.05)
Conscientiousness	1.13 (0.96-1.33)
Extraversion	1.03 (0.89-1.21)
Neuroticism	0.97 (0.83-1.12)
Openness	1.15 (0.99-1.34)
Engagement skewness	0.64[b] (0.51-0.79)
Nudge probability	1.02 (1.00-1.05)
Meeting counts	0.62[b] (0.49-0.78)
No meeting minutes	1 (0.99-1.01)
Self-event counts	1.15 (0.85-1.54)
Email messages sent	1.05 (0.92-1.20)
Email messages read	1.01 (0.99-1.03)
Ad hoc call count	0.89 (0.52-1.49)
Number of attention signals	1.39[b] (1.17-1.66)
Trigger source reference (system)	
Trigger source (rescheduled)	1.77[b] (1.32-2.38)

[a]Observations=1585; Tjur R^2=0.07.

[b]Statistically significant (*P*<.05) after Benjamini-Hochberg correction.

[c]Statistically significant (*P*<.05) before any correction.

Next, we modeled each of the 3 choice outcome measures—*distract chosen*, *calm chosen*, and *address chosen*—as a function of per-participant characteristics, per-half hour nudge probability, and per-nudge contextual metrics. Table 3 outlines the ORs and CIs for all models.

Table 3. Odds ratios (ORs) and CIs for each predictor of *distract chosen, calm chosen,* and *address chosen* for all participant-chosen interventions.

Predictors	Distract chosen[a], OR (95% CI)	Calm chosen[b], OR (95% CI)	Address chosen[c], OR (95% CI)
Intercept	0.11[d] (0.01-0.73)	1.84 (0.32-10.62)	0.44 (0.04-4.52)
Age group (y), reference (18-35)			
36-45	1.65[d] (1.03-2.67)	0.62[d] (0.40-0.96)	0.88 (0.46-1.65)
>46	0.71 (0.40-1.26)	0.95 (0.57-1.57)	1.6 (0.85-3.07)
Gender reference (man)			
Gender (woman)	0.62[d] (0.39-0.98)	1.31 (0.87-1.96)	1.29 (0.73-2.26)
Cognitive reappraisal	1.01 (0.82-1.25)	1.27[d] (1.06-1.53)	0.67[e] (0.53-0.84)
Expressive suppression	0.99 (0.83-1.19)	1.06 (0.90-1.25)	0.95 (0.75-1.18)
Resilience	0.93 (0.69-1.25)	0.87 (0.67-1.13)	1.42[d] (1.00-2.02)
Agreeableness	0.75 (0.55-1.03)	1.02 (0.77-1.36)	1.46 (0.99-2.20)
Conscientiousness	1.30 (0.98-1.75)	0.92 (0.71-1.18)	0.74 (0.51-1.05)
Extraversion	1.18 (0.91-1.53)	1.00 (0.79-1.27)	0.79 (0.57-1.11)
Neuroticism	0.88 (0.68-1.14)	1.00 (0.79-1.26)	1.36 (0.99-1.89)
Openness	1.67[e] (1.30-2.16)	0.71[d] (0.57-0.88)	0.83 (0.60-1.13)
Engagement skewness	1.12 (0.74-1.68)	1.11 (0.78-1.59)	0.86 (0.53-1.38)
Nudge probability	1.00 (0.96-1.05)	1.01 (0.97-1.05)	0.97 (0.92-1.03)
Meeting counts	0.84 (0.55-1.25)	1.07 (0.76-1.52)	1.14 (0.71-1.77)
No meeting minutes	1.00 (0.99-1.02)	1.00 (0.99-1.01)	1.00 (0.98-1.02)
Self-event counts	1.08 (0.63-1.79)	0.83 (0.53-1.30)	1.19 (0.65-2.05)
Email messages sent	1.05 (0.84-1.30)	1.00 (0.82-1.22)	0.99 (0.74-1.30)
Email messages read	1.03 (0.99-1.07)	0.99 (0.96-1.03)	0.96 (0.90-1.01)
Ad hoc call count	0.88 (0.36-1.94)	0.93 (0.45-1.93)	1.26 (0.45-3.05)
Number of attention signals	0.65[e] (0.48-0.87)	1.18 (0.92-1.53)	1.33 (0.94-1.87)
Trigger source reference (system)			
Trigger source (rescheduled)	1.01 (0.61-1.63)	0.76 (0.49-1.16)	1.39 (0.81-2.35)

[a]Observations=650; Tjur R^2=0.115.
[b]Observations=650; Tjur R^2=0.059.
[c]Observations=650; Tjur R^2=0.089.
[d]Statistically significant (P<.05) before any correction.
[e]Statistically significant (P<.05) after Benjamini-Hochberg correction.

For choosing distract interventions, we found that higher *openness* was associated with a higher likelihood. In contrast, being a woman and higher *number of attention signals* were associated with a lower likelihood of choosing distract interventions. We found that higher *cognitive reappraisal* was associated with a higher likelihood and lower *openness* was associated with a lower likelihood of choosing calm interventions. We also found that higher *resilience* was associated with a higher likelihood and lower *cognitive reappraisal* was associated with a lower likelihood of choosing address interventions.

To understand the effect of intervention choice on engagement, we modeled *engaged* as a function of per-participant characteristics, per-half hour nudge probability, per-nudge contextual metrics, and per-intervention characteristics. The logistic regression model of *engaged* was significant (χ^2_{27}=69.6; P<.001). Table 4 outlines the ORs and CIs for each predictor. We found that a higher *resilience*, choosing a prompted-based intervention, and choosing a video-based intervention were associated with a higher likelihood of engagement. In contrast, choosing an intervention that could be performed inside was associated with a lower likelihood of engagement.

Table 4. Odds ratios (ORs), CI, and *P* values for each predictor of *engaged after chosen*, that is, engaged for all interventions after participants chose a category[a].

Predictors	Engaged after chosen, OR (95% CI)
Intercept	0.05[b] (0.00-0.964)
Age group (y), reference (18-35)	
36-45	1.06 (0.54-2.05)
>46	1.65 (0.71-4.02)
Gender (woman)	1.65 (0.88-3.16)
Cognitive reappraisal	1.11 (0.83-1.48)
Expressive suppression	1.25 (0.95-1.57)
Resilience	1.63[b] (1.05-2.59)
Agreeableness	1.24 (0.77-2.00)
Conscientiousness	0.81 (0.55-1.17)
Extraversion	0.72 (0.48-1.08)
Neuroticism	1.39 (0.96-2.07)
Openness	1.24 (0.88-1.76)
Engagement skewness	1.01 (0.60-1.73)
Nudge probability	1.01 (0.95-1.07)
Meeting counts	0.65 (0.41-1.06)
No meeting minutes	1 (0.98-1.02)
Self-event counts	1.94 (0.86-1.67)
Email messages sent	1.16 (0.84-1.67)
Email messages read	0.97 (0.93-1.03)
Ad hoc call count	0.54 (0.23-1.33)
Number of attention signals	0.9 (0.61-1.33)
Trigger source reference (system)	
Trigger source (rescheduled)	1.79 (0.85-4.10)
Engagement skewness	1.01 (0.60-1.73)
Category reference (distract)	
Category (calm)	1.14 (0.60-2.14)
Category (address)	0.55 (0.22-1.43)
Modality reference (conversation)	
Modality (prompt)	3.53[b] (1.35-9.60)
Modality (video)	5.86[b] (1.69-21.68)
Location reference (at desk)	
Location (inside)	0.43[b] (0.23-0.83)
Location (outside)	0.34 (0.11-1.18)

[a]Observations=650; Tjur R^2=0.12.

[b]Statistically significant (P<.05) before any correction.

Intervention Effectiveness

Finally, we analyzed the factors associated with higher intervention rating and intervention effectiveness. The logistic regression model of *liked* as a function of per-participant characteristics, per-half hour nudge probability, per-nudge contextual metrics, and per-intervention metrics was significant (χ^2_{28}=92.4; P<.001). The ORs and CIs for each predictor can be seen in Table 5. We found that being a woman, higher *cognitive reappraisal*, higher *extraversion*, and higher *stress* *reduction* were associated with a higher likelihood of liking the intervention. In contrast, being aged 36 to 45 years, higher *expressive suppression* and higher *nudge probability* were associated with a lower likelihood of liking the intervention.

Table 5. Odds ratios (ORs), CIs, and *P* values for each predictor of *liked* for all interventions that participants engaged in[a].

Predictors	Liked, OR (95% CI)
Intercept	5.95 (0.51-70.91)
Age group (y), reference (18-35)	
36-45	0.51[b] (0.30-0.86)
>46	1.01 (0.55-1.83)
Gender reference (man)	
Gender (woman)	2.51[c] (1.51-4.20)
Cognitive reappraisal	1.35[b] (1.08-1.69)
Expressive suppression	0.82[b] (0.68-0.99)
Resilience	1.10 (0.80-1.51)
Agreeableness	0.72 (0.50-1.02)
Conscientiousness	0.80 (0.58-1.09)
Extraversion	1.46[b] (1.10-1.95)
Neuroticism	0.84 (0.64-1.11)
Openness	0.99 (0.75-1.30)
Engagement skewness	0.73 (0.47-1.12)
Nudge probability	0.94[b] (0.90-0.99)
Meeting counts	0.93 (0.61-1.42)
No meeting minutes	1.00 (0.99-1.02)
Self-event counts	0.66 (0.40-1.10)
Email messages sent	1.01 (0.79-1.28)
Email messages read	1.00 (0.96-1.04)
Ad hoc call count	0.66 (0.24-1.69)
Number of attention signals	1.16 (0.84-1.59)
Trigger source reference (system)	
Trigger source (rescheduled)	0.78 (0.47-1.30)
Category reference (distract)	
Category (calm)	0.8 (0.49-1.29)
Category (address)	0.94 (0.44-2.01)
Stress reduction	2.36[c] (1.60-3.54)
Modality reference (conversation)	
Modality (prompt)	0.52 (0.17-1.53)
Modality (video)	0.75 (0.22-2.47)
Location reference (at desk)	
Location (inside)	0.69 (0.40-1.19)
Location (outside)	2.56 (0.79-9.27)

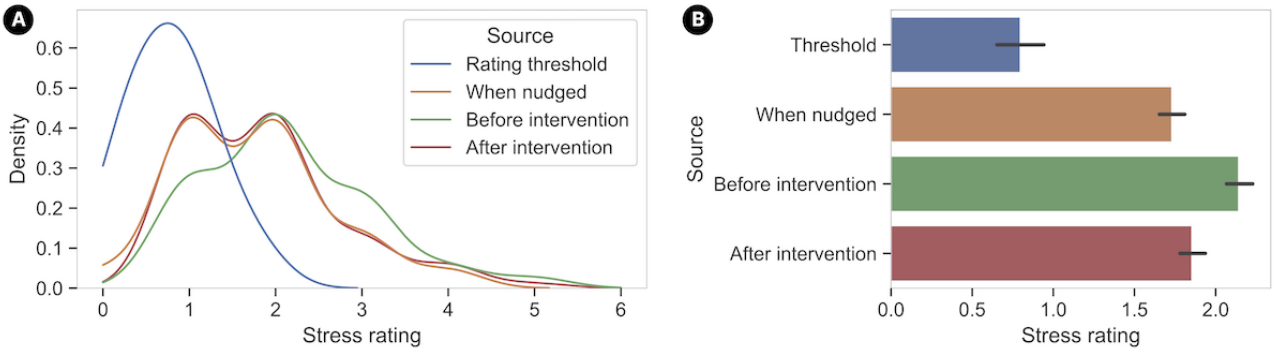
[a]Observations=521; Tjur R^2=0.166.

[b]Statistically significant (*P*<.05) before any correction.

[c]Statistically significant (*P*<.05) after Benjamini-Hochberg correction.

When we examine the momentary stress ratings surrounding the system nudges and the intervention use (Figure 2A), we see that the distributions of EMA and preintervention and postintervention stress ratings sit higher than the individual thresholds used for sending the nudges. The average stress rating measured just before the intervention use (mean 2.14) is higher than the average stress rating used for system nudges (mean 1.73), indicating that the participants were more likely to engage in interventions when their stress ratings were higher than average (Figure 2B).

Figure 2. (A) Distribution of momentary stress ratings. A kernel density estimate plot of subjective stress ratings shows that the distribution of stress ratings when the system sent intervention nudges is higher than the individual thresholds. Stress ratings before intervention are distributed at a higher rating than stress ratings after intervention and the ratings when the nudges were sent. (B) Comparison of momentary stress ratings. Average momentary stress ratings at different points in time with 95% CIs.



We then examined the factors associated with stress reduction. The logistic regression model of *improved* was significant (χ^2_{30}=186.7; *P*<.001), and the resulting ORs and CIs per predictor can be found in Table 6. We found that higher *agreeableness*, higher *nudge probability*, higher *stress before*, higher *rating*, and getting prompt-based or video interventions were associated with a higher likelihood of improvement on their stress rating. In contrast, being a woman, higher *cognitive reappraisal*, higher *neuroticism*, and choosing calm or address interventions were associated with a lower likelihood of improvement on their stress rating.

Table 6. Odds ratios (ORs), CIs, and *P* values for each predictor of *improved* for all interventions that participants engaged in[a].

Predictors	Improved, OR (95% CI)
Intercept	0.00[b] (0.00-0.00)
Age group (y), reference (18-35)	
36-45	0.58 (0.30-1.13)
>46	0.98 (0.49-1.98)
Gender reference (man)	
Gender (woman)	0.41[b] (0.21-0.77)
Cognitive reappraisal	0.69[b] (0.52-0.91)
Expressive suppression	1.01 (0.79-1.28)
Resilience	0.84 (0.55-1.25)
Agreeableness	1.73[c] (1.10-2.76)
Conscientiousness	1.27 (0.86-1.89)
Extraversion	0.89 (0.62-1.29)
Neuroticism	0.57[b] (0.39-0.81)
Openness	1.09 (0.78-1.55)
Engagement skewness	1.18 (0.68-2.03)
Nudge probability	1.09[b] (1.03-1.16)
Meeting counts	0.85 (0.48-1.42)
No meeting minutes	1.01 (0.99-1.03)
Self-event counts	1.29 (0.69-2.36)
Email messages sent	1.04 (0.78-1.39)
Email messages read	0.96 (0.90-1.01)
Ad hoc call count	0.94 (0.28-2.63)
Number of attention signals	1.4 (0.95-2.05)
Trigger source reference (system)	
Trigger source (rescheduled)	1.08 (0.57-2.02)
Category reference (distract)	
Category (calm)	0.43[b] (0.23-0.78)
Category (address)	0.40[c] (0.16-0.97)
Stress before	5.76[b] (3.98-8.64)
Rating	2.47[b] (1.84-3.39)
Modality reference (conversation)	
Modality (prompt)	6.65[c] (1.53-36.45)
Modality (video)	5.62[c] (1.12-34.10)
Location reference (at desk)	
Location (inside)	1.17 (0.60-2.25)
Location (outside)	0.23 (0.03-1.10)

[a]Observations=521; Tjur R^2=0.338.

[b]Statistically significant (*P*<.05) after Benjamini-Hochberg correction.

[c]Statistically significant (*P*<.05) before any correction.

Discussion

Principal Findings

In this paper, we presented a comprehensive and systematic study that simultaneously encompassed trait-based factors (ie, individual characteristics), state-based factors (ie, workplace contextual and behavioral signals and momentary stress), and intervention-related factors (ie, location and function) to identify what drives JIT intervention engagement and efficacy. This study particularly focuses on momentary outcomes to inform the design of *dynamic tailoring*, which is a key component of JITAI systems. In this study, we leveraged surveys, EMA, and passively-sensed data from the deployment of the JIT stress-reduction intervention system to identify significant factors that influence the momentary engagement, the choice of interventions, the engagement given an intervention choice, the user rating of interventions engaged, and the stress reduction from the engagement.

We found that stress ratings immediately after the JIT interventions were significantly lower than those reported in the moments immediately before the interventions (Figure 2B). While keeping in mind the study sample, we found that women, those with higher *neuroticism*, those with higher *cognitive reappraisal* skills, and those that chose *calm* and *address* interventions were significantly less likely to experience stress reduction, while those with higher *agreeableness* and those that chose prompt-based or video-based interventions were significantly more likely to experience stress reduction. Surprisingly, contextual signals, such as meeting or email load, did not lead to a significant increase or decrease in stress ratings, which may indicate that trait-based or intervention-related factors matter more for efficacy or that the contextual signals may need finer granularity.

In contrast, we found that contextual signals such as lower *meeting counts* and lower *engagement skewness* were associated with a higher likelihood of engagement, indicating that state-based contextual factors such as being in a meeting or time of the day may matter more for engagement than efficacy. In addition, a JIT intervention that was explicitly rescheduled to a later time was more likely to be engaged. This implies that intervention engagement can be improved by giving some level of control to the users over complete automation.

With respect to the choice of interventions and liking the interventions, factors that significantly influenced the choice were primarily trait-based. Higher *openness* was associated with a higher likelihood of choosing *distract* but a lower likelihood of choosing *calm*, while higher *cognitive reappraisal* was associated with a higher likelihood of choosing *calm* but lower likelihood of choosing *address*. Higher *resilience* was associated with a higher likelihood of choosing *address* as well as the subsequent engagement after the choice. Interventions were liked more by women, those with higher *cognitive reappraisal*, and those with higher *extraversion*, and when the interventions led to stress reduction. One state-based factor that stood out is the *number of attention signals*, which negatively impacted the likelihood of choosing *distract* interventions. Higher *number of attention signals* indicates higher desk-bound work activity (eg, keyboard and mouse) where getting the mind off work (ie, how *distract* interventions were communicated) during intense work activities may not be feasible or desired.

In summarizing our findings, we first categorize these factors into (1) nonmodifiable individual factors, (2) modifiable individual factors, (3) contextual factors, and (4) content factors.

The distinction between modifiable and nonmodifiable individual factors is important for intervention design. Once modifiable factors are identified, strategies can be deployed to directly influence those factors if those strategies can lead to a greater impact on the engagement or the efficacy of the interventions. Nonmodifiable factors are also important to determine which subset of populations can benefit from additional targeted support through organizational or policy-level changes [81]. In our analysis, nonmodifiable individual factors include gender and age, and modifiable individual factors, although debatable [82], include emotion regulation skills and personality traits. Our findings revealed that these individual factors not only influence study-long engagement [42] but also influence momentary engagement and can be useful for tailoring JITAI systems. Although our findings for individual factors corroborate with prior studies, it is important to highlight that our study evaluates instance-level engagement metrics through play-by-play analysis of app-use behaviors rather than study-long engagement metrics. Contextual factors such as workload (eg, meeting counts and email counts) and availability (eg, presence or activity at the computer) are helpful in the JITAI system's decision-making process for when to interrupt the user. Content factors include intervention-related information such as the amount of effort required, the modality of intervention delivery, and the ideal location for intervention engagement, and these factors can inform JITAI systems in determining which intervention to present to the user given the understanding of the current context.

In this section, we summarize and discuss the findings organized by these 4 categories of factors with recommendations for design and future research.

Nonmodifiable Individual Factors

In our analysis, we found several significant effects of gender and age on our outcome measures. Participants who self-reported as being a woman had more than twice the likelihood of liking interventions than being a man, despite having less likelihood of improving from engaging in interventions. Being a woman also had 38% less likelihood of choosing *distract* interventions than being a man, suggesting that there may be an unobserved motivational factor. For example, although not statistically significant, women reported higher momentary stress on average compared with men (2.15 for women vs 1.84 for men; $t_{20.756}$=–1.545; P=.14), which may contribute to choosing more *calm* or *address* interventions to reduce their stress. Although the general findings from DMHI studies that women are more likely to engage in digital interventions than men are corroborated by our analysis [42], the fact that women improve less despite engaging more, liking interventions more, and choosing more *address* (ie, high reward) interventions is a concern for systems design that only take engagement metrics into account.

Participants aged >46 years were 47% more likely to engage in interventions while being aged 36 to 45 years increased the likelihood of choosing *distract* interventions by 65% and decreased the likelihood of liking interventions by 49% compared with those aged 18 to 35 years. Although the effects of different age groups on engagement have mixed results across prior studies, the higher engagement rate for participants aged >46 years in the study could be explained by a higher rate of interest in digital interventions for older populations [83]. Participants in the 36 to 45 years age group choosing *distract* interventions and not liking them highlight an opportunity for finding different types of interventions that they might enjoy.

Tailoring JITAI systems to individuals has been suggested to improve engagement and efficacy of interventions in prior research [36]. Our findings suggest that certain age or gender groups may benefit more than others from our system. The lack of engagement or the lack of improvement despite engagement for some groups highlights opportunities for targeting research and design efforts to further understand unobserved barriers to engagement and effective responses. Although it is generally recommended that intervention content be tailored based on individuals, we also recommend that the efficacy and user rating of interventions be closely monitored to ensure that certain groups enjoy the same level of benefits as other groups. In addition, new intervention content could be codesigned with groups that may not be benefiting as much and added to the system on a regular basis to equalize outcomes across subgroups.

Modifiable Individual Factors

Our findings corroborate prior research that personality trait is a strong factor in the engagement of DMHIs [42]. A prior study found that openness to experience is associated with better adherence and lower odds of attrition [84]. In our analysis, we did not find a significant effect on engagement, but we found that *openness* significantly increases the likelihood of choosing *distract* but decreases the likelihood of choosing *calm* interventions. One possible explanation may be the variety in the intervention content, which people with high openness would prefer [85]. *Distract* interventions tend to offer more variety in content with videos of travel destinations and humor as well as opportunities to explore social connections, whereas *calm* interventions were mostly introspective activities such as breathing or focused observations.

Although the effects were only moderately significant (P<.06), we found that an increase by one point in *neuroticism* and *agreeableness* scales increases the likelihood of choosing *address* interventions, which were designed to help users directly address resolve the stress-inducing components of their lives. For participants with high scores in *neuroticism* and *agreeableness* scales, it is possible that interventions that help them directly address their stress were more appealing than others that were designed to distract from stress and refocus on the present. Prior study has also found neuroticism and agreeableness to associate with a stronger interest in the use of stress management apps [86]. We also found that one point increase in *agreeableness* scale was associated with an increased likelihood of improvement by 73%, whereas the same point increase in *neuroticism* scale was associated with a lower likelihood of improvement by 43%, despite both having higher tendency to choose *address* interventions. Because agreeableness is known to be positively associated with the therapeutic alliance in mental health treatments [87], it is possible that the prosocial and cooperative nature of those with high agreeableness [88] allowed them to fully engage in the *address* interventions that were more action-oriented, leading to a greater improvement. In contrast, neuroticism has been known to negatively correlate with adherence to mental health recommendations [89] and to a wide variety of mental health treatment outcomes [87].

Prior work has associated extraversion with lower interest in using web-based mental health services over face-to-face interactions with a provider [83]. Although our interventions were delivered solely through technology, we found that one point increase in *extraversion* scale increases the likelihood of liking the intervention by 46%. We found no significant association between the intervention content (category, modality, or location) and liking them to explain this behavior. Future research should explore how extraversion facilitates momentary engagement in interventions.

Corroborating with prior study-long findings that associate personalities with DMHI engagement, our study found that personality traits also influence momentary engagement and efficacy of interventions. We recommend that JITAI systems carefully monitor potentially unhelpful use behaviors by incorporating personality traits in the system adaptation algorithm because they may impact the choice of interventions that may lead to negative downstream effects on outcomes. For example, it may be beneficial to offer a variety of more effortful interventions for people who report higher scores in openness. For people who report higher scores in neuroticism, the system could suggest less effortful interventions.

Across the board, emotion regulation styles had significant effects. One point increase in the *cognitive reappraisal* scale was associated with a 14% increase in the likelihood of engaging, a 27% increase in the likelihood of choosing *calm* interventions, a 35% increase in the likelihood of liking the intervention, a 33% decrease in the likelihood of choosing *address* interventions, and a 31% decrease in the likelihood of improvement after engagement. It is important to note the differences in the range of point scales. An increase of 14% for a 7-point scale is equivalent to an increase of 20% for a 5-point scale. In contrast, one point increase in the *expressive suppression* scale was associated with an 18% decrease in the likelihood of liking the intervention. We also found that one point increase in the *resilience* scale increases the likelihood of choosing *address* interventions by 42% and increases the likelihood of subsequently engaging in interventions by 63%.

Prior research has explored the role of emotion regulation in stress coping. For example, emotion regulation skills help assess stressful situations and determine the appropriate emotional response [90] or act as buffers against the negative effects of stress [91]. Emotion regulation has also been theorized as a moderator for increased resilience after encountering a stressful situation [92]. Although our analysis cannot claim the causal direction between coping skills and engagement, our findings suggest that emotion regulation and resilience may play a role in not only the stress-coping process but also in choosing different interventions or deciding to engage in an intervention. While most research has argued for increasing coping skills as an outcome measure or a treatment target [93], our findings suggest promising new research directions in understanding how coping skills could impact our decisions to engage in therapeutic interventions in the moment.

The role of emotion regulation strategies in altering our decisions and choices in various contexts has been empirically studied in highly controlled laboratory settings [94]. Prior research has also studied personality traits [82,95] and coping skills [93] as "states" that exhibit intrapersonal variations, and such states can be modified through therapeutic strategies (eg, cognitive behavioral therapy). Although our study has assumed personality traits, emotion regulation skills, and resilience to be stable over the course of the study, it remains to be seen how shifting the perspectives of these characteristics to be more dynamic would inform the JITAI systems. Because of the potential mediating role of coping skills in perceived stress [96] and the role of perceived stress in outcome improvement (ie, one point increase in stress rating before the intervention leads to being >5 times more likely to improve in our findings), coping skills should be measured periodically and incorporated into the decision-making process of JITAI systems. Therefore, we recommend further research to explore how taking a dynamic approach to personality and coping skills would inform the design of JITAI systems.

Contextual Factors

We found that the more the participant tends to engage at the end of the day, the more likely they would engage in interventions and that the system-initiated interventions that were rescheduled to a later time increased the likelihood of engagement. These findings may suggest a tendency to defer interventions to later in the day. Prior work that applied the self-determination theory [97] to JITAI systems suggested that perceived competence and self-regulation abilities may deplete throughout the day, potentially leading to unhealthy choices (eg, unhealthy food and alcohol) toward the end of the day [34]. This has a serious consequence for those who tend to procrastinate or postpone healthy behaviors (eg, exercise, and stress intervention) toward the end of the day. In contrast, allowing people to defer an intervention to a specific time may increase self-efficacy and perception of control, which might lead to an increased chance of behavior change [98]. In fact, Howe et al [19] reported that many participants liked the ability to perform the interventions when they wanted. We also found that the increase in nudge probability decreases the likelihood of liking the intervention but increases the likelihood of improvement, revealing that a JIT intervention might be "a bitter pill to swallow" but a useful pill, nonetheless. However, a relentless reminder could lead to distraction [99] and eventual system abandonment [44].

Our findings show encouragement that intelligent timing based on contextual information could improve the engagement and effectiveness of interventions, but perhaps at the cost of a lowered sense of user agency and control, negative perceptions toward the interventions, or leading to unhealthy choices toward the end of the day. It also highlights that relying solely on what people like may not be helpful for stress reduction and that there needs to be an additional investigation into how momentary factors could influence user ratings. Therefore, future JITAI systems should carefully balance user ratings (eg, likes) with intervention efficacy (eg, stress reduction and engagement) and help users discover what works best for them. Considering that higher user ratings may not always reflect the effectiveness of the interventions, such systems should aim to simultaneously explore the rating and the improvement in determining the timing of interventions.

In evaluating the work context, as expected, we found that a nudge sent at a time when the user is less likely to be in a meeting but more active at the desk improves engagement. Contrary to our hypothesis, we found no significant associations with *no meeting minutes* or *self-event counts*. It is possible that there is high variability in the level of focus and attention needed during times carved out for self. For example, our data sources cannot discern if the times carved out for self were work-related (eg, focus time for reading and writing) or non–work related (eg, running errands, child pick up, and exercise). This study's data sources cannot achieve automatic detection of activities beyond basic work activities, such as meetings, emails, chats, calls, or computer activities.

Although tailoring to the activity context is the defining promise of JITAI systems, automatically detecting the activities performed within a time window is not an easy task. In addition, how much tailoring and the level of system intelligence are really necessary to maximize engagement and outcomes is unknown, especially given the cost and risks associated with the invasion of privacy in passive sensing. Further research is necessary to understand the cost and benefit of accurate activity detection and intelligent timing on the engagement and effectiveness of JITAI systems.

Content Factors

In our analysis, intervention categories, modalities, and locations showed pronounced effects on engagement and improvement, suggesting the importance of the intervention content in the design of JITAI systems. We found that having an intervention that could be done at the desk more than doubled the likelihood of engagement compared with an intervention that could be done indoors, but not at the desk. It is possible that leaving the desk at the moment of the nudge was not appropriate given the situation or there was an unobserved motivational barrier. In these scenarios, suggesting a different activity, rescheduling the activity, or waiting until the next appropriate transition time might have been beneficial. To improve engagement, intervention designers could provide additional desk-based stress reduction techniques to minimize the burden of leaving the desk.

In contrast, we found that having a prompt- or video-based intervention increased the likelihood of engagement by more than 3-fold and improvement by more than 5-fold, compared with a conversation-based intervention. Although prompt- and video-based interventions were typically less effortful than conversation-based interventions that require many turn-taking interactions with the bot, conversation-based interventions were designed to address the sources of the stress with the hope of creating a longer-lasting impact. It is possible that the conversations were not usable for participants to fully engage in the content. This finding suggests that quick, effortless interventions could be useful at the moment, but complex, turn-taking interventions need more thoughtful redesign.

Our findings revealed that the effects of intervention content types were significant and large. To investigate the impact of choices on improving outcomes and user ratings, future digital intervention systems that provide a catalog of interventions would benefit from characterizing each intervention on multiple dimensions, such as the level of effort, location, modality, etc. Such characterization would help understand the interplay among personality traits, contextual cues, and intervention types that users are likely to choose and benefit from. Therefore, we recommend that JITAI systems provide a variety of interventions to fine-tune its recommendations based on contexts but also to identify interventions that may need redesign.

Limitations

Our system design and analysis setup have several limitations. The retrospective analysis of the inferred stress scores indicates that there is room for improvement. Our stress scores were computed from a generalized algorithm, but recent studies have shown that stress can be idiosyncratic and, therefore, needs to be modeled at an individual level [18]. Our user engagement flow cannot differentiate the dismissal of the nudge due to bad timing or low stress or both. Our analysis setup does not allow for determining the causal relationship between the individual, contextual, and content factors with engagement, stress reduction, and intervention ratings. Even though the deployment study collected long-term stress measures via the DASS-21 scale, our analysis was limited to momentary stress ratings because microinterventions are more appropriate for proximal outcomes over distal outcomes [31]. Microrandomized trials are a promising research direction for JITAI systems to quantify the impact of tailored interventions on both short- and long-term outcomes [100] with careful considerations for the appropriate sample size [101]. Our data were also limited by a small sample population that exhibited low-stress levels, focusing on US-based information workers from a large technology organization, skewed toward engineers and those that identified as man. Prior work has suggested that, when evaluating engagement (or attrition), the severity of symptoms should also be considered [41]. Therefore, future research should evaluate the system through microrandomized trials with a sample population exhibiting high severity of stress symptoms while also expanding into other information worker roles from different sectors.

The deployment of the system used in the study is limited by the system requirements. Our system assumes that a single organization may offer a standardized set of software tools to workers for several reasons (eg, security, privacy, productivity, compliance, and scalability) and does not support interoperability across different software tools used by workers within an organization or across organizations. Beyond this challenge, it is important to note that the data required to describe an individual worker's stress necessarily crosses work–non-work boundaries, and such boundaries are increasingly blurred by technologies [102] and remote work [103]. Our contextual data captured at work were noisy and were aggregated without peeking into the content (eg, of email or documents) to preserve privacy; in addition, it is challenging to separate blurred contexts. Such blurred boundaries raise concerns surrounding boundary preferences, data ownership, values and incentives, well-being definitions, and power dynamics that may undermine the successful deployment of personal stress support systems [104]. Perhaps more important than the interoperability of a system or clean data is a human-centered implementation process that allows for flexibility and adaptation of the system based on multistakeholder perspectives and use over time. This paper piloted one aspect of such adaptation. However, much work is needed in realizing the full potential of JITAI systems that take sociotechnical considerations into account and where users can participate in drawing the boundary and providing the right level and fidelity of data that meets their needs.

Conclusions

JITAI systems have the potential to integrate timely support into the workplace. To identify key factors that should be incorporated into the system's adaptable decision-making process, we analyzed data from a 4-week deployment of a JIT workplace stress reduction microintervention system. On the basis of our findings, we recommend that individual, contextual, and content-based factors be incorporated into the system for tailoring as well as for monitoring unhelpful use behaviors across subgroups and contexts. Future work should explore careful balancing of individual preferences, intervention efficacy, and system accuracy to help users discover what works best for them and to continuously improve system recommendations.

Acknowledgments

This research was funded by Microsoft Research. TA was supported in part by the National Science Foundation (NSF) (grant CAREER IIS-2142794; NSF grant IIS-1901386) and the National Institutes of Health (grant R01MH125179). The authors are deeply grateful to the participants for their time and willingness to install our system, engage in our interventions, and provide valuable feedback and data. This study would not have been possible without the guidance and support of Dr Patricia Areán, Dr James Fogarty, Dr Eric Horvitz, and the Microsoft Research Human Understanding and Empathy group. We thank Ann Paradiso for generating the title image for this article.

Data Availability

The data sets generated during or analyzed during this study are not publicly available due to privacy and legal restrictions but are available from the corresponding author on reasonable request with a clear justification and a license agreement. The request will be reviewed and approved case by case by Microsoft Research Release and Compliance team, at which point a license agreement will be drafted and shared.

Conflicts of Interest

The research was conducted when JS, EH, RL, JH, KS, and MC were employed at Microsoft Research. JS, EH, JH, and MC were involved in the design, development, and deployment of the innovations and software evaluated in this study.

Multimedia Appendix 1

Completed Checklist of Guidelines and Checklist for the Reporting on Digital Health Implementations (iCHECK-DH).

[PDF File (Adobe PDF File), 122 KB - mental_v11i1e48974_app1.pdf]

Multimedia Appendix 2

(1) Detailed system architecture and stress inference methods, (2) intervention design and example interaction flows, (3) detailed description of how interaction and context data were extracted, and (4) additional results from ANOVA type II and multicollinearity analyses conducted for each logistic regression model.

[PDF File (Adobe PDF File), 2820 KB - mental_v11i1e48974_app2.pdf]

References

1. Ray J. Americans' stress, worry and anger intensified in 2018. Gallup. 2019. URL: https://news.gallup.com/poll/249098/americans-stress-worry-anger-intensified2018.aspx [accessed 2021-08-20]
2. Boudreaux E, Mandry C, Brantley PJ. Stress, job satisfaction, coping, and psychological distress among emergency medical technicians. Prehosp Disaster Med 2012 Jun 28;12(4):9-16. [doi: 10.1017/s1049023x00037742]
3. Elizabeth B, Kahaema B, Peter JG, Liu C, Prather AA, Thomas K, et al. Stress and health disparities: contexts, mechanisms, and interventions among racial/ethnic minority and low-socioeconomic status populations. American Psychiatric Association. 2017. URL: http://www.apa.org/pi/health-disparities/resources/stress-report.aspx, [accessed 2024-04-29]
4. Colligan TW, Higgins EM. Workplace stress: etiology and consequences. J Workplace Behav Health 2006 Jul 25;21(2):89-97. [doi: 10.1300/j490v21n02_07]
5. Westman M, Etzion D. The impact of vacation and job stress on burnout and absenteeism. Psychol Health 2001 Sep;16(5):595-606. [doi: 10.1080/08870440108405529] [Medline: 22804501]
6. Cooper CL, Cartwright S. An intervention strategy for workplace stress. J Psychosom Res 1997 Jul;43(1):7-16. [doi: 10.1016/s0022-3999(96)00392-3] [Medline: 9263926]
7. Richardson KM, Rothstein HR. Effects of occupational stress management intervention programs: a meta-analysis. J Occup Health Psychol 2008 Jan;13(1):69-93. [doi: 10.1037/1076-8998.13.1.69] [Medline: 18211170]
8. Kinman G, Jones F. Lay representations of workplace stress: what do people really mean when they say they are stressed? Work Stress 2007 Feb 23;19(2):101-120. [doi: 10.1080/02678370500144831]
9. Dollard MF. Psychosocial safety climate: a lead indicator of workplace psychological health and engagement and a precursor to intervention success. In: Biron C, Karanika-Murray M, Cooper C, editors. Improving Organizational Interventions for Stress and Well-Being: Addressing Process and Context. New York, NY: Routledge; 2012:77-110.
10. Herrero SG, Saldaña M, Rodriguez JG, Ritzel DO. Influence of task demands on occupational stress: gender differences. J Safety Res 2012 Dec;43(5-6):365-374. [doi: 10.1016/j.jsr.2012.10.005] [Medline: 23206509]
11. Helbig S, Fehm L. Problems with homework in CBT: rare exception or rather frequent? Behav Cogn Psychother 2004 Jun 14;32(3):291-301. [doi: 10.1017/s1352465804001365]
12. Lee E, Kim Y. Effect of university students' sedentary behavior on stress, anxiety, and depression. Perspect Psychiatr Care 2019 Apr 24;55(2):164-169 [FREE Full text] [doi: 10.1111/ppc.12296] [Medline: 29797324]
13. Yuhan L, Bongshin L, Donghee YW, Amanda LR, David EC, Eun KC. Time for break: understanding information workers' sedentary behavior through a break prompting system. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018 Presented at: CHI '18; April 21-26, 2018; Montreal, QC p. 1-14 URL: https://dl.acm.org/doi/10.1145/3173574.3173701 [doi: 10.1145/3173574.3173701]
14. Thomée S, Härenstam A, Hagberg M. Computer use and stress, sleep disturbances, and symptoms of depression among young adults – a prospective cohort study. BMC Psychiatry 2012 Oct 22;12(1):176. [doi: 10.1186/1471-244x-12-176]
15. Gloria M, Yiran W, Melissa N. Stress and multitasking in everyday college life: an empirical study of online activity. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2014 Presented at: CHI '14; April 26-May1, 2014; Toronto, ON p. 41-50 URL: https://dl.acm.org/doi/10.1145/2556288.2557361 [doi: 10.1145/2556288.2557361]

16. Day A, Paquet S, Scott N, Hambley L. Perceived information and communication technology (ICT) demands on employee outcomes: the moderating effect of organizational ICT support. J Occup Health Psychol 2012 Oct;17(4):473-491. [doi: 10.1037/a0029837] [Medline: 23066697]
17. William CM. Who?s running the machine? A theoretical exploration of work stress and burnout of technologically tethered workers. J individ Employ Rights 2007;12(3):263. [doi: 10.2190/ie.12.3.f]
18. Mehrab BM, Javier H, Daniel M, Jina S, Esther H, Kael R, et al. Advancing the understanding and measurement of workplace stress in remote information workers from passive sensors and behavioral data. In: Proceedings of the 10th International Conference on Affective Computing and Intelligent Interaction. 2022 Presented at: ACII '22; October 18-21, 2022; Nara, Japan p. 1-8 URL: https://ieeexplore.ieee.org/document/9953824 [doi: 10.1109/acii55700.2022.9953824]
19. Howe G, Suh J, Bin Morshed M, McDuff D, Rowan K, Hernandez J, et al. Design of digital workplace stress-reduction intervention systems: effects of intervention type and timing. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022 Presented at: CHI '22; April 29-May 5, 2022; New Orleans, LA p. 1-16 URL: https://dl.acm.org/doi/10.1145/3491102.3502027 [doi: 10.1145/3491102.3502027]
20. Cornet VP, Holden RJ. Systematic review of smartphone-based passive sensing for health and wellbeing. J Biomed Inform 2018 Jan;77:120-132 [FREE Full text] [doi: 10.1016/j.jbi.2017.12.008] [Medline: 29248628]
21. Martinez GJ, Mattingly SM, Robles-Granda P, Saha K, Sirigiri A, Young J, et al. Predicting participant compliance with fitness tracker wearing and ecological momentary assessment protocols in information workers: observational study. JMIR Mhealth Uhealth 2021 Nov 12;9(11):e22218 [FREE Full text] [doi: 10.2196/22218] [Medline: 34766911]
22. Gonzalo JM, Stephen MM, Shayan M, Subigya KN, Andrew TC, Anind KD, et al. On the quality of real-world wearable data in a longitudinal study of information workers. In: Proceedings of the 2020 IEEE International Conference on Pervasive Computing and Communications Workshops. 2020 Presented at: PerComWorkshops '20; March 23-27, 2020; Austin, TX p. 1-6 URL: https://ieeexplore.ieee.org/document/9156113 [doi: 10.1109/percomworkshops48775.2020.9156113]
23. Daniel M, Kael R, Piali C, Jessica W, ThuVan P, Mary C. A multimodal emotion sensing platform for building emotion-aware applications. arXiv Preprint posted online March 28, 2019 [FREE Full text]
24. Poh M, McDuff DJ, Picard RW. Advancements in noncontact, multiparameter physiological measurements using a webcam. IEEE Trans Biomed Eng 2011 Jan;58(1):7-11. [doi: 10.1109/tbme.2010.2086456]
25. Gloria M, Shamsi TI, Mary C, Paul J. Bored Mondays and focused afternoons: the rhythm of attention and online activity in the workplace. In: Proceedings of the 2014 SIGCHI Conference on Human Factors in Computing Systems. 2014 Presented at: CHI '14; April 26-May 1, 2014; Toronto, ON p. 3025-3034 URL: https://dl.acm.org/doi/10.1145/2556288.2557204 [doi: 10.1145/2556288.2557204]
26. Javier H, Pablo P, Asta R, Mary C. Under pressure: sensing stress of computer users. In: Proceedings of the 2014 SIGCHI Conference on Human Factors in Computing Systems. 2014 Presented at: CHI '14; April 26-May 1, 2014; Toronto, ON p. 51-60 URL: https://dl.acm.org/doi/10.1145/2556288.2557165#:~:text=During%20the%20stressful%20conditions%2C%20the,75%25%20of%20the%20participants) [doi: 10.1145/2556288.2557165]
27. Bidargaddi N, Schrader G, Klasnja P, Licinio J, Murphy S. Designing m-Health interventions for precision mental health support. Transl Psychiatry 2020 Jul 07;10(1):222 [FREE Full text] [doi: 10.1038/s41398-020-00895-2] [Medline: 32636358]
28. Huguet A, Rao S, McGrath PJ, Wozney L, Wheaton M, Conrod J, et al. A systematic review of cognitive behavioral therapy and behavioral activation apps for depression. PLoS One 2016 May 2;11(5):e0154248 [FREE Full text] [doi: 10.1371/journal.pone.0154248] [Medline: 27135410]
29. Neary M, Schueller SM. State of the field of mental health apps. Cogn Behav Pract 2018 Nov;25(4):531-537 [FREE Full text] [doi: 10.1016/j.cbpra.2018.01.002] [Medline: 33100810]
30. de Korte EM, Wiezer N, Janssen JH, Vink P, Kraaij W. Evaluating an mHealth app for health and well-being at work: mixed-method qualitative study. JMIR Mhealth Uhealth 2018 Mar 28;6(3):e72 [FREE Full text] [doi: 10.2196/mhealth.6335] [Medline: 29592846]
31. Baumel A, Fleming T, Schueller SM. Digital micro interventions for behavioral and mental health gains: core components and conceptualization of digital micro intervention care. J Med Internet Res 2020 Oct 29;22(10):e20631 [FREE Full text] [doi: 10.2196/20631] [Medline: 33118946]
32. Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, et al. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. Ann Behav Med 2018 May 18;52(6):446-462 [FREE Full text] [doi: 10.1007/s12160-016-9830-8] [Medline: 27663578]
33. Akane S, Paul J, Mary C. Designing opportune stress intervention delivery timing using multi-modal data. In: Proceedings of the 7th International Conference on Affective Computing and Intelligent Interaction. 2017 Presented at: ACII '17; October 23-26, 2017; San Antonio, TX p. 346-353 URL: https://ieeexplore.ieee.org/document/8273623 [doi: 10.1109/acii.2017.8273623]
34. Donna SN. Dynamic models of behavior for just-in-time adaptive interventions. IEEE Pervasive Comput 2014;13(3):17.

35. Juarascio AS, Parker MN, Lagacey MA, Godfrey KM. Just-in-time adaptive interventions: a novel approach for enhancing skill utilization and acquisition in cognitive behavioral therapy for eating disorders. Int J Eat Disord 2018 Aug 26;51(8):826-830 [FREE Full text] [doi: 10.1002/eat.22924] [Medline: 30051495]
36. Wang L, Miller LC. Just-in-the-moment adaptive interventions (JITAI): a meta-analytical review. Health Commun 2020 Nov 05;35(12):1531-1544. [doi: 10.1080/10410236.2019.1652388] [Medline: 31488002]
37. Liao P, Dempsey W, Sarker H, Hossain SM, Al'absi M, Klasnja P, et al. Just-in-time but not too much: determining treatment timing in mobile health. Proc ACM Interact Mob Wearable Ubiquitous Technol 2018 Dec 27;2(4):1-21. [doi: 10.1145/3287057] [Medline: 30801052]
38. Liao P, Greenewald K, Klasnja P, Murphy S. Personalized HeartSteps: a reinforcement learning algorithm for optimizing physical activity. Proc ACM Interact Mob Wearable Ubiquitous Technol 2020 Mar 18;4(1):1-22 [FREE Full text] [doi: 10.1145/3381007] [Medline: 34527853]
39. Mishra V, Künzler F, Kramer J, Fleisch E, Kowatsch T, Kotz D. Detecting receptivity for mHealth interventions in the natural environment. Proc ACM Interact Mob Wearable Ubiquitous Technol 2021 Jun 24;5(2):1-24 [FREE Full text] [doi: 10.1145/3463492] [Medline: 34926979]
40. Künzler F, Mishra V, Kramer J, Kotz D, Fleisch E, Kowatsch T. Exploring the state-of-receptivity for mHealth interventions. Proc ACM Interact Mob Wearable Ubiquitous Technol 2019 Dec 14;3(4):1-27 [FREE Full text] [doi: 10.1145/3369805] [Medline: 36159209]
41. Gavin D, David C, John S. Engagement with online mental health interventions: an exploratory clinical study of a treatment for depression. In: Proceedings of the 2012 SIGCHI Conference on Human Factors in Computing Systems. 2012 Presented at: CHI '12; May 5-10, 2012; Austin, TX p. 1421-1430 URL: https://dl.acm.org/doi/10.1145/2207676.2208602 [doi: 10.1145/2207676.2208602]
42. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
43. Hardeman W, Houghton J, Lane K, Jones A, Naughton F. A systematic review of just-in-time adaptive interventions (JITAIs) to promote physical activity. Int J Behav Nutr Phys Act 2019 Apr 03;16(1):31 [FREE Full text] [doi: 10.1186/s12966-019-0792-7] [Medline: 30943983]
44. Bidargaddi N, Almirall D, Murphy S, Nahum-Shani I, Kovalcik M, Pituch T, et al. To prompt or not to prompt? A microrandomized trial of time-varying push notifications to increase proximal engagement with a mobile health app. JMIR Mhealth Uhealth 2018 Nov 29;6(11):e10123 [FREE Full text] [doi: 10.2196/10123] [Medline: 30497999]
45. Yardley L, Spring BJ, Riper H, Morrison LG, Crane DH, Curtis K, et al. Understanding and promoting effective engagement with digital behavior change interventions. Am J Prev Med 2016 Nov;51(5):833-842. [doi: 10.1016/j.amepre.2016.06.015] [Medline: 27745683]
46. Perrin Franck C, Babington-Ashaye A, Dietrich D, Bediang G, Veltsos P, Gupta PP, et al. iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res 2023 May 10;25:e46694 [FREE Full text] [doi: 10.2196/46694] [Medline: 37163336]
47. Perrin Franck C, Babington-Ashaye A, Dietrich D, Bediang G, Veltsos P, Gupta PP, et al. iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res 2023 May 10;25:e46694 [FREE Full text] [doi: 10.2196/46694] [Medline: 37163336]
48. Susanne H, Sayantani M, Ariane B, Gillian C. Women in the tech industry: gaining ground, but facing new headwinds. Deloitte Insights. 2021 Dec 1. URL: https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2022/statistics-show-women-in-technology-are-facing-new-headwinds.html [accessed 2024-02-18]
49. Lovibond P, Lovibond S. The structure of negative emotional states: comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. Behav Res Ther 1995 Mar;33(3):335-343. [doi: 10.1016/0005-7967(94)00075-u] [Medline: 7726811]
50. Rammstedt B, John OP. Measuring personality in one minute or less: a 10-item short version of the Big Five Inventory in English and German. J Res Pers 2007 Feb;41(1):203-212. [doi: 10.1016/j.jrp.2006.02.001]
51. Gross JJ, John OP. Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. J Pers Soc Psychol 2003 Aug;85(2):348-362. [doi: 10.1037/0022-3514.85.2.348] [Medline: 12916575]
52. Smith BW, Dalen J, Wiggins K, Tooley E, Christopher P, Bernard J. The brief resilience scale: assessing the ability to bounce back. Int J Behav Med 2008 Sep;15(3):194-200. [doi: 10.1080/10705500802222972] [Medline: 18696313]
53. Gloria M, Shamsi TI, Mary C, Paul J, Akane S, Yuliya L. Email duration, batching and self-interruption: patterns of email use on productivity and stress. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. 2016 Presented at: CHI '16; May 7-12, 2016; San Jose, CA p. 1717-1728 URL: https://dl.acm.org/doi/10.1145/2858036.2858262 [doi: 10.1145/2858036.2858262]
54. Gloria M, Shamsi I, Mary C. How blocking distractions affects workplace focus and productivity. In: Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM

International Symposium on Wearable Computers. 2017 Presented at: UbiComp '17; September 11-15, 2017; Maui, HI p. 928-934 URL: https://dl.acm.org/doi/abs/10.1145/3123024.3124558 [doi: 10.1145/3123024.3124558]
55. McDuff D, Jun E, Rowan K, Czerwinski M. Longitudinal observational evidence of the impact of emotion regulation strategies on affective expression. IEEE Trans Affective Comput 2021 Jul 1;12(3):636-647. [doi: 10.1109/taffc.2019.2961912]
56. Rosenberg EL, Ekman P. What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS) (Series in Affective Science). Oxford, UK: Oxford University Press; 1998.
57. Mayo LM, Heilig M. In the face of stress: interpreting individual differences in stress-induced facial expressions. Neurobiol Stress 2019 Feb;10:100166 [FREE Full text] [doi: 10.1016/j.ynstr.2019.100166] [Medline: 31193535]
58. Karim Sadik K. Assessment of emotional experience through facial expression. ProQuest Dissertation & Theses, Harvard University. 2010. URL: https://www.proquest.com/openview/bb4208bb09f6a82b0247e66d69cba33a/ [accessed 2024-06-18]
59. Vrijkotte TG, van Doornen LJ, de Geus EJ. Effects of work stress on ambulatory blood pressure, heart rate, and heart rate variability. Hypertension 2000 Apr;35(4):880-886 [FREE Full text] [doi: 10.1161/01.hyp.35.4.880] [Medline: 10775555]
60. Hjortskov N, Rissén D, Blangsted AK, Fallentin N, Lundberg U, Søgaard K. The effect of mental stress on heart rate variability and blood pressure during computer work. Eur J Appl Physiol 2004 Jun 1;92(1-2):84-89. [doi: 10.1007/s00421-004-1055-z] [Medline: 14991326]
61. Butler AC, Chapman JE, Forman EM, Beck AT. The empirical status of cognitive-behavioral therapy: a review of meta-analyses. Clin Psychol Rev 2006 Jan;26(1):17-31. [doi: 10.1016/j.cpr.2005.07.003] [Medline: 16199119]
62. Anita L, Linehan MM. Dialectical behavior therapy: a comprehensive multi- and transdiagnostic intervention. In: Nezu CM, Nezu AM, editors. The Oxford Handbook of Cognitive and Behavioral Therapies. Oxford, UK: Oxford University Press; 2016:200-214.
63. Linehan MM. DBT Skills Training Manual. 2nd edition. New York, NY: Guilford Publications; 2014.
64. Overview of Microsoft graph. Microsoft Learn. 2023. URL: https://learn.microsoft.com/en-us/graph/overview [accessed 2024-05-25]
65. Advanced insights metrics description. Microsoft Learn. 2023. URL: https://learn.microsoft.com/en-us/previous-versions/viva/insights/use/metric-definitions [accessed 2024-05-25]
66. Liu X, Fromm J, Patel S, McDuff D. Multi-task temporal shift attention networks for on-device contactless vitals measurement. arXiv Preprint posted online February 28, 2021 [FREE Full text]
67. Microsoft bot framework. Microsoft. 2024. URL: https://dev.botframework.com [accessed 2024-05-25]
68. Adaptive cards. Microsoft. 2024. URL: https://adaptivecards.io [accessed 2024-05-25]
69. Smyth JM, Heron KE. Is providing mobile interventions "just-in-time" helpful? An experimental proof of concept study of just-in-time intervention for stress management. In: Proceedings of the 2016 IEEE Wireless Health. 2016 Presented at: WH '16; October 25-28, 2016; Bethesda, MD p. 1-7 URL: https://ieeexplore.ieee.org/document/7764561 [doi: 10.1109/wh.2016.7764561]
70. Klasnja P, Harrison BL, LeGrand L, LaMarca A, Froehlich JE, Hudson S. Using wearable sensors and real time inference to understand human recall of routine activities. In: Proceedings of the 10th international conference on Ubiquitous computing. 2008 Presented at: UbiComp '08; September 21-24, 2008; Seoul, Republic of Korea p. 154-163 URL: https://dl.acm.org/doi/10.1145/1409635.1409656 [doi: 10.1145/1409635.1409656]
71. Horvitz E. Principles of mixed-initiative user interfaces. In: Proceedings of 1999 the SIGCHI conference on Human Factors in Computing Systems. 1999 Presented at: CHI '99; May 15-20, 1999; Pittsburgh, PA p. 159-166 URL: https://dl.acm.org/doi/10.1145/302979.303030 [doi: 10.1145/302979.303030]
72. Schroeder J, Suh J, Wilks C, Czerwinski M, Munson SA, Fogarty J, et al. Data-driven implications for translating evidence-based psychotherapies into technology-delivered interventions. Int Conf Pervasive Comput Technol Healthc 2020 May;2020:274-287 [FREE Full text] [doi: 10.1145/3421937.3421975] [Medline: 33912357]
73. Schroeder J, Suh J, Wilks C, Czerwinski M, Munson SA, Fogarty J, et al. Data-driven implications for translating evidence-based psychotherapies into technology-delivered interventions. Int Conf Pervasive Comput Technol Healthc 2020 May;2020:274-287 [FREE Full text] [doi: 10.1145/3421937.3421975] [Medline: 33912357]
74. Choi W, Park S, Kim D, Lim Y, Lee U. Multi-stage receptivity model for mobile just-in-time health intervention. Proc ACM Interact Mob Wearable Ubiquitous Technol 2019 Jun 21;3(2):1-26. [doi: 10.1145/3328910]
75. McCrae RR, Costa Jr PT. A contemplated revision of the NEO five-factor inventory. Pers Individ Dif 2004 Feb;36(3):587-596. [doi: 10.1016/S0191-8869(03)00118-1]
76. McCrae RR, John OP. An introduction to the five-factor model and its applications. J Pers 1992 Jun 28;60(2):175-215. [doi: 10.1111/j.1467-6494.1992.tb00970.x] [Medline: 1635039]
77. John OP, Srivastava S. The big five trait taxonomy: history, measurement, and theoretical perspectives. In: Pervin LA, John OP, editors. Handbook of Personality: Theory and Research. 2nd edition. New York, NY: Guilford Press; 1999:102-138.
78. Connor-Smith JK, Flachsbart C. Relations between personality and coping: a meta-analysis. J Pers Soc Psychol 2007 Dec;93(6):1080-1107. [doi: 10.1037/0022-3514.93.6.1080] [Medline: 18072856]
79. Bolger N, Zuckerman A. A framework for studying personality in the stress process. J Pers Soc Psychol 1995 Nov;69(5):890-902. [doi: 10.1037//0022-3514.69.5.890] [Medline: 7473036]

80. Langsrud Ø. ANOVA for unbalanced data: Use type ii instead of type iii sums of squares. Stat Comput 2003;13(2):167 [FREE Full text] [doi: 10.1023/A:1023260610025]
81. Ho FK, Celis-Morales CA, Gray SR, Katikireddi SV, Niedzwiedz CL, Hastie C, et al. Modifiable and non-modifiable risk factors for COVID-19, and comparison to risk factors for influenza and pneumonia: results from a UK Biobank prospective cohort study. BMJ Open 2020 Nov 19;10(11):e040402 [FREE Full text] [doi: 10.1136/bmjopen-2020-040402] [Medline: 33444201]
82. Roberts BW, Luo J, Briley DA, Chow PI, Su R, Hill PL. A systematic review of personality trait change through intervention. Psychol Bull 2017 Feb;143(2):117-141. [doi: 10.1037/bul0000088] [Medline: 28054797]
83. March S, Day J, Ritchie G, Rowe A, Gough J, Hall T, et al. Attitudes toward e-Mental health services in a community sample of adults: online survey. J Med Internet Res 2018 Feb 19;20(2):e59 [FREE Full text] [doi: 10.2196/jmir.9109] [Medline: 29459357]
84. Mikolasek M, Witt CM, Barth J. Adherence to a mindfulness and relaxation self-care app for cancer patients: mixed-methods feasibility study. JMIR Mhealth Uhealth 2018 Dec 06;6(12):e11271 [FREE Full text] [doi: 10.2196/11271] [Medline: 30522990]
85. McCrae RR, Costa PT. Conceptions and correlates of openness to experience. In: Hogan R, Johnson JA, Briggs SR, editors. Handbook of Personality Psychology. Cambridge, MA: Academic Press; 1997:825-847.
86. Ervasti M, Kallio J, Määttänen I, Mäntyjärvi J, Jokela M. Influence of personality and differences in stress processing among Finnish students on interest to use a mobile stress management app: survey study. JMIR Ment Health 2019 May 13;6(5):e10039 [FREE Full text] [doi: 10.2196/10039] [Medline: 31094358]
87. Bucher MA, Suzuki T, Samuel DB. A meta-analytic review of personality traits and their associations with mental health treatment outcomes. Clin Psychol Rev 2019 Jun;70:51-63. [doi: 10.1016/j.cpr.2019.04.002] [Medline: 30981042]
88. Graziano WG, Eisenberg N. Agreeableness: a dimension of personality. In: Hogan R, Johnson JA, Briggs SR, editors. Handbook of Personality Psychology. Cambridge, MA: Academic Press; 1997:795-824.
89. Pereira A, Cabacos C, Soares S, Araújo A, Manão A, Amaral A, et al. Relationship between personality, adherence to (mental) health behaviours and psychological distress during the COVID-19 pandemic. Eur Psychiatr 2022 Sep 01;65(S1):S499-S500. [doi: 10.1192/j.eurpsy.2022.1270]
90. Myruski S, Denefrio S, Dennis-Tiwary TA. Stress and emotion regulation: the dynamic fit model. In: Harkness KL, Hayden EP, editors. The Oxford Handbook of Stress and Mental Health. Oxford, UK: Oxford University Press; 2019:415-434.
91. Wang M, Saudino KJ. Emotion regulation and stress. J Adult Dev 2011 Jan 11;18(2):95-103. [doi: 10.1007/s10804-010-9114-7]
92. Troy AS, Mauss IB. Resilience in the face of stress: emotion regulation as a protective factor. In: Southwick SM, Litz BT, Charney D, Friedman MJ, editors. Resilience and Mental Health: Challenges Across the Lifespan. Cambridge, MA: Cambridge University Press; 2011:30-44.
93. Berking M, Wupperman P, Reichardt A, Pejic T, Dippel A, Znoj H. Emotion-regulation skills as a treatment target in psychotherapy. Behav Res Ther 2008 Nov;46(11):1230-1237. [doi: 10.1016/j.brat.2008.08.005] [Medline: 18835479]
94. Grecucci A, Sanfey AG. Emotion regulation and decision making. In: Gross JJ, editor. Handbook of Emotion Regulation. 2nd edition. New York, NY: The Guilford Press; 2014:140-153.
95. Glinski K, Page AC. Modifiability of neuroticism, extraversion, and agreeableness by group cognitive behaviour therapy for social anxiety disorder. Behav Change 2012 Feb 22;27(1):42-52. [doi: 10.1375/bech.27.1.42]
96. Boyle CC, Stanton AL, Ganz PA, Crespi CM, Bower JE. Improvements in emotion regulation following mindfulness meditation: effects on depressive symptoms and perceived stress in younger breast cancer survivors. J Consult Clin Psychol 2017 Apr;85(4):397-402 [FREE Full text] [doi: 10.1037/ccp0000186] [Medline: 28230391]
97. Deci EL, Ryan RM. Intrinsic Motivation and Self-Determination in Human Behavior. Berlin, Germany: Springer Science & Business Media; 2013.
98. Ajzen I. The theory of planned behavior. Organ Behav Hum Decis Process 1991 Dec;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-t]
99. Kostadin K, Jason P, Elizabeth WD. "Silence your phones": smartphone notifications increase inattention and hyperactivity symptoms. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. 2016 Presented at: CHI '16; May 7-12, 2016; San Jose, CA p. 1011-1020 URL: https://dl.acm.org/doi/10.1145/2858036.2858359 [doi: 10.1145/2858036.2858359]
100. Klasnja P, Smith S, Seewald NJ, Lee A, Hall K, Luers B, et al. Efficacy of contextually tailored suggestions for physical activity: a micro-randomized optimization trial of HeartSteps. Ann Behav Med 2019 May 03;53(6):573-582 [FREE Full text] [doi: 10.1093/abm/kay067] [Medline: 30192907]
101. Liao P, Klasnja P, Tewari A, Murphy SA. Sample size calculations for micro-randomized trials in mHealth. Stat Med 2016 May 30;35(12):1944-1971 [FREE Full text] [doi: 10.1002/sim.6847] [Medline: 26707831]
102. Olson-Buchanan JB, Boswell WR. Blurring boundaries: correlates of integration and segmentation between work and nonwork. J Vocat Behav 2006 Jun;68(3):432-445. [doi: 10.1016/j.jvb.2005.10.006]
103. Allen TD, Merlo K, Lawrence RC, Slutsky J, Gray CE. Boundary management and work - nonwork balance while working from home. Applied Psychology 2021 Jan 20;70(1):60-84. [doi: 10.1111/apps.12300]

104. Jina S, Javier H, Koustuv S, Kathy D, Mehrab BM, Esther H, et al. Towards successful deployment of wellbeing sensing technologies: identifying misalignments across contextual boundaries. In: Proceedings of the 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos. 2023 Presented at: ACIIW '23; September 10-13, 2023; Cambridge, MA p. 1-8 URL: https://ieeexplore.ieee.org/document/10388088 [doi: 10.1109/aciiw59127.2023.10388088]

Abbreviations

DASS-21: Depression, Anxiety, and Stress Scale-21
DMHI: digital mental health intervention
EMA: ecological momentary assessment
iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations
JIT: just-in-time
JITAI: just-in-time adaptive intervention
OR: odds ratio

Edited by I Lefter; submitted 17.05.23; peer-reviewed by M Rabbi, H Wen; comments to author 15.01.24; revised version received 05.06.24; accepted 17.07.24; published 12.09.24.

Please cite as:
Suh J, Howe E, Lewis R, Hernandez J, Saha K, Althoff T, Czerwinski M
Toward Tailoring Just-in-Time Adaptive Intervention Systems for Workplace Stress Reduction: Exploratory Analysis of Intervention Implementation
JMIR Ment Health 2024;11:e48974
URL: https://mental.jmir.org/2024/1/e48974
doi:10.2196/48974
PMID:

Original Paper

Developing and Implementing a Web-Based Branching Logic Survey to Support Psychiatric Crisis Evaluations of Individuals With Developmental Disabilities: Qualitative Study and Evaluation of Validity

Deborah A Bilder[1], MD; Mariah Mthembu[2], MD; Whitney Worsham[2], MD; Patricia Aguayo[1], MPH, MD; Jacob R Knight[2], BS; Steven W Deng[2], HBS; Tejinder P Singh[2], BDS, MDS, MPH; John Davis[3], PhD

[1]University of Utah Huntsman Mental Health Institute, Salt Lake City, UT, United States

[2]University of Utah School of Medicine, Salt Lake City, UT, United States

[3]University of Utah, Salt Lake City, UT, United States

Corresponding Author:
Deborah A Bilder, MD
University of Utah Huntsman Mental Health Institute
383 Colorow Drive, Room 360
Salt Lake City, UT, 84108
United States
Phone: 1 801 410 0852
Email: deborah.bilder@hsc.utah.edu

Abstract

Background: Individuals with developmental disabilities (DD) experience increased rates of emotional and behavioral crises that necessitate assessment and intervention. Psychiatric disorders can contribute to crises; however, screening measures developed for the general population are inadequate for those with DD. Medical conditions can exacerbate crises and merit evaluation. Screening tools using checklist formats, even when designed for DD, are too limited in depth and scope for crisis assessments. The Sources of Distress survey implements a web-based branching logic format to screen for common psychiatric and medical conditions experienced by individuals with DD by querying caregiver knowledge and observations.

Objective: This paper aims to (1) describe the initial survey development, (2) report on focus group and expert review processes and findings, and (3) present results from the survey's clinical implementation and evaluation of validity.

Methods: Sources of Distress was reviewed by focus groups and clinical experts; this feedback informed survey revisions. The survey was subsequently implemented in clinical settings to augment providers' psychiatric and medical history taking. Informal and formal consults followed the completion of Sources of Distress for a subset of individuals. A records review was performed to identify working diagnoses established during these consults.

Results: Focus group members (n=17) expressed positive feedback overall about the survey's content and provided specific recommendations to add categories and items. The survey was completed for 231 individuals with DD in the clinical setting (n=161, 69.7% men and boys; mean age 17.7, SD 10.3; range 2-65 years). Consults were performed for 149 individuals (n=102, 68.5% men and boys; mean age 18.9, SD 10.9 years), generating working diagnoses to compare survey screening results. Sources of Distress accuracy rates were 91% (95% CI 85%-95%) for posttraumatic stress disorder, 87% (95% CI 81%-92%) for anxiety, 87% (95% CI 81%-92%) for episodic expansive mood and bipolar disorder, 82% (95% CI 75%-87%) for psychotic disorder, 79% (95% CI 71%-85%) for unipolar depression, and 76% (95% CI 69%-82%) for attention-deficit/hyperactivity disorder. While no specific survey items or screening algorithm existed for unspecified mood disorder and disruptive mood dysregulation disorder, these conditions were caregiver-reported and working diagnoses for 11.7% (27/231) and 16.8% (25/149) of individuals, respectively.

Conclusions: Caregivers described Sources of Distress as an acceptable tool for sharing their knowledge and insights about individuals with DD who present in crisis. As a screening tool, this survey demonstrates good accuracy. However, better differentiation among mood disorders is needed, including the addition of items and screening algorithm for unspecified mood disorder and disruptive mood dysregulation disorder. Additional validation efforts are necessary to include a more geographically diverse population and reevaluate mood disorder differentiation. Future study is merited to investigate the survey's impact on the psychiatric and medical management of distress in individuals with DD.

(JMIR Ment Health 2024;11:e50907) doi:10.2196/50907

KEYWORDS

developmental disabilities; disruptive behavior; psychiatric comorbidity; web-based; psychiatric crisis; disability; mental health; behavioral crises; intervention; general population; screening; accuracy; mood disorder; sources of distress; autism; intellectual disability

Introduction

Background

Individuals with developmental disabilities (DD) such as autism and intellectual disability (ID) experience mental health crises more frequently than the general population [1,2]. A broad range of psychiatric and medical conditions can contribute to the agitation, aggression, and self-injury that often characterize these crises [3-10]. Rates of anxiety (20%-77%), depression (10%-20%), expansive mood and bipolar disorder (5%-11%), and psychosis (5%-10%) among individuals with autism exceed those in neurotypical individuals [11-19]. Elevated rates of psychiatric disorders have also been identified in individuals with ID, notably for unspecified psychosis (4.8%), schizophrenia (3.9%), and bipolar disorder (8%) [20-22]. A history of trauma or abuse should also be considered in individuals with DD presenting in crisis [23].

When psychiatric and medical conditions are recognized as factors contributing to a person's mental health crisis, clear long-term treatment targets emerge. Nevertheless, for those with DD, co-occurring medical and psychiatric conditions are often unrecognized, leaving them vulnerable to experiencing diagnostic overshadowing. Diagnostic overshadowing occurs when disruptive behaviors in individuals with DD are attributed to their disability without consideration of other potential medical or psychiatric conditions that could contribute to their behavioral presentation [24].

Self-, parent-, and caregiver-report mental health questionnaires provide an efficient means of screening for common psychiatric conditions in the neurotypical population. However, for those with DD, self-report questionnaires may be impeded by communication deficits or a limited capacity to reflect on internal experiences. Parent- and caregiver-report questionnaires normed in typically developing children may also provide inadequate mental health screening for those with ID because they often include items that are inapplicable to children with minimal language ability, exclude severe conditions that disproportionately affect children with DD (eg, mania and psychosis), and overlook the individualized manner in which psychiatric symptoms manifest in this population [20,25-27].

The American Psychiatric Association and the National Association for the Dually Diagnosed published the *Diagnostic Manual–Intellectual Disability* in 2007, and subsequently, in 2016, the second edition (*Diagnostic Manual–Intellectual Disability–Second Edition*; *DM-ID-2*) [28,29]. These texts adapt the *Diagnostic and Statistical Manual of Mental Disorders* criteria to reflect their presentation in individuals with ID. The Psychopathology Instrument for Mentally Retarded Adults and the Psychiatric Assessment Schedule for Adults with Developmental Disabilities (PAS-ADD) operationalize adapted diagnostic criteria into structured interviews to provide a framework through which to identify psychiatric conditions in this population [30,31]. These interviews are quite lengthy and require training to administer. Even as an abbreviated semistructured interview, the Mini PAS-ADD Clinical Interview takes approximately 45 minutes to complete [32]. Existing parent- and caregiver-report psychiatric screening tools for individuals with ID create a more efficient and practical means of collecting information [33-36]; yet, the checklist format of parent- and caregiver-report questionnaires limits depth and scope, both of which are necessary when evaluating crises in a population with complex medical and mental health needs. In addition, there is a great need for the inclusion of items that query symptoms of common medical conditions (eg, epilepsy, gastrointestinal disorders, and poor dentition) that manifest with agitation and aggression and occur more frequently in individuals with DD [3,37,38].

Sources of Distress is a survey developed for parents and caregivers (hereinafter collectively referred to as *caregivers*) that uses a web-based branching logic format to screen for mental health and medical conditions among individuals with DD who present in crisis. This tool informs the care of individuals experiencing distress and is intended for use when the severity or persistence of disruptive behavior prompts the consideration of medication intervention. Screening information endorsed by caregivers is organized into relevant psychiatric and medical categories within a report. This report (Multimedia Appendix 1 [39]) is developed for the caregiver and can subsequently facilitate their shared decision-making process with health care providers as specific underlying conditions are evaluated. Sources of Distress aims to minimize diagnostic overshadowing and optimize the ability of the caregiver and the provider to recognize the presence of psychiatric and medical conditions that merit targeted intervention. The web-based branching logic format is adaptive in nature—optimizing caregiver and health care provider convenience and efficiency and minimizing caregiver burden for survey completion [40].

Objectives

This paper aims to (1) describe the initial development of Sources of Distress; (2) report on the findings from focus group evaluations and expert reviews and indicate how this feedback shaped the subsequent version of the survey; and (3) present the results from the evaluation of validity for Sources of Distress after its implementation in the clinical setting. The *Methods* and *Results* sections are divided into 3 subsections (apart from the

Ethical Considerations section in *Methods*) corresponding to the development, initial evaluation, and clinical implementation phases of Sources of Distress.

Methods

Ethical Considerations

The University of Utah Institutional Review Board approved focus group activities for Sources of Distress content validation (IRB_00111975). Focus group participants provided informed consent and received compensation for their time in the form of an Amazon gift card worth US $50. The University of Utah Institutional Review Board approved with a waiver of consent for the retrospective records review, data collection, and subsequent deidentified data analysis for individuals for whom Sources of Distress was completed as part of their clinical care (IRB_00170868).

Early Survey Development

Funding for the development of Sources of Distress was provided by the Autism Council of Utah based in Murray, Utah, United States [41]. The development team comprised a triple board physician (pediatrics, general psychiatry, and child and adolescent psychiatry), an educational psychologist, a medical student, and a business consultant grandparent of a child with autism and ID. In the initial development phase, Sources of Distress was built in Qualtrics (Qualtrics International Inc) using a branching logic format to approximate the history-taking component of a DD psychiatric evaluation. This evaluation queries psychiatric symptom clusters, physical complaints, and psychiatric medical history to support the development of a diagnostic impression for which treatment recommendations could be made.

Multiple expert opinion sources were reviewed to identify pertinent screening categories and corresponding items to include in Sources of Distress. The expert sources included published literature, the *DM-ID-2*, the Mini PAS-ADD Clinical Interview, and the screening interview for the Kiddie Schedule for Affective Disorders and Schizophrenia–Present and Lifetime (a semistructured psychiatric diagnostic interview for children and adolescents) [28,32,42]. As Sources of Distress is intended for use in the context of distress, the presence of at least 1 manifestation of a behavioral or emotional crisis must be endorsed to initiate survey questions.

Initial Survey Evaluation

Focus Group Evaluation

In 2018 and early 2019, focus group participants were recruited from (1) a university-based outpatient program that provides medical and psychiatric care for individuals with DD across the lifespan and (2) the Autism Council of Utah (a community stakeholder organization for individuals and families affected by autism). Six focus groups were conducted that consisted collectively of parents (6/17, 35%), professional caregivers (6/17, 35%), and adults with both DD and the ability to provide verbal feedback (5/17, 29%). Participants completed Sources of Distress before attending the focus group and reported on specific items, missing items, item wording, and attribution of items to corresponding conditions. Interviews and discussions were transcribed and analyzed following the framework analysis of Ritchie and Spencer [43]. Inductive reasoning and the constant comparative method put forth by Strauss and Corbin [44] were used to compare statements by parents, professional caregivers, and individuals with disability within and across focus groups.

Expert Review Evaluation

Revisions were made to Sources of Distress based on focus group feedback. Experts reviewed the revised survey version, and additional changes were made. The experts included a pediatrician and 2 child psychiatrists, all with national recognition for their clinical and research work in DD.

Clinical Implementation

Overview

Sources of Distress was implemented in various clinical settings to augment the clinical history-taking process—outpatient (primary care, neurology, developmental pediatrics, and psychiatry), emergency department, psychiatric inpatient, and residential care. Caregivers were given a link to the survey when their health care provider identified the need for expert support in managing severe agitation and aggression. All caregivers (231/231, 100%) completed the survey outside of the clinical setting. An informal or formal consult followed survey completion for a subset of individuals. In August 2020, the survey was transitioned from the Qualtrics platform to the REDCap (Research Electronic Data Capture; Vanderbilt University) platform to automate the Sources of Distress report generation using the custom template engine [45]. This external REDCap module was developed and has been maintained by the Integrated Research Informatics Services of British Columbia Children’s Hospital Research Institute [46].

Survey Data Collection

Sources of Distress responses were collected from its first use in a clinical setting from February 2019 through June 2022. The following information was obtained: respondent type, individual characteristics, caregiver-reported diagnoses, current medications, distress manifestations, psychiatric symptoms, and medical symptoms, conditions, or concerns. When multiple caregivers reported on the same individual, responses were used from the caregiver closest to where the individual lived (eg, parent for a child living at home and professional caregiver for an individual living in a residential setting). Psychotropic medications were organized within the following mutually exclusive categories: antipsychotics, antidepressants, non-antidepressant anxiolytics, anticonvulsants, lithium, alpha-2 agonists, stimulants, and atomoxetine.

Consults

A medical decision-making support consultation took place after survey completion as either an informal or a formal consult for a subset of individuals. This consult was conducted by a clinical team led by the triple board physician member of the survey’s development team. The consult team used *DM-ID-2* criteria as the basis for establishing psychiatric diagnoses. At a minimum (as an informal consult), the consult involved a discussion between a DD clinical expert and the referring provider. This discussion resulted in a collective determination of working diagnoses and treatment plan. A formal consult included the additional components of medical records review, caregiver interview, and direct participant evaluation. Psychiatric diagnoses that were not reported in the survey but discussed by the provider or documented in the medical record were included among preexisting diagnoses.

Working diagnoses were abstracted from formal and informal consult documentation and served as the standard to define true case status.

Mood Disorder Classification

The presence of a mood disorder among preexisting and working diagnoses was classified into mutually exclusive categories such that there was no overlap among individuals across mood disorder categories to allow for direct comparisons across preexisting diagnoses, survey screening status results, and working diagnoses. The following mood disorder classification hierarchy was used from highest to lowest: (1) episodic expansive mood, hypomania, mania, and bipolar disorder, hereafter collectively referred to as *bipolar disorder*, (2) disruptive mood dysregulation disorder (DMDD) and unspecified mood disorder, and (3) unipolar depression. If an individual had a diagnosis of bipolar disorder, regardless of what other mood disorder diagnosis was reported or identified, their mood disorder classification would be bipolar disorder. An individual was only classified with unipolar depression if (1) they had a depression diagnosis and (2) they had no other mood disorder diagnosis.

Statistical Analyses and Evaluation of Validity

Descriptive statistics and chi-square tests were conducted in SPSS (version 28.0; IBM Corp) with an α of .05 selected to assess statistical significance. Differences between surveys with an accompanying consult and those without were measured. Positive predictive value (PPV), negative predictive value (NPV), and accuracy rates were calculated for (1) preexisting diagnoses and (2) survey screening results with working diagnoses used as the determinant of true case status. We calculated 95% CIs for the binomial distribution of accuracy rates.

Results

Early Survey Development

Table 1 lists the modules and corresponding items initially selected as the categories, characteristics, and symptoms to be queried by Sources of Distress. The initial version of the survey included scoring algorithms to determine positive screen status for the following conditions: anxiety, unipolar depression, bipolar disorder, psychosis, and attention-deficit/hyperactivity disorder (ADHD).

Table 1. Description of Sources of Distress and additions in response to focus group feedback.

Module	Original items	Added in response to feedback
Introduction and demographics	• Respondent's relationship to the individual who is affected • Distress symptoms • Language ability • Age • Known diagnoses • Current medications	• For professional caregivers: how long have you known the affected individual? • Added "increased fixation on certain things" and "changes in behavior such as increased isolation, social withdrawal" to distress symptoms • Is there a difference in language ability at the physician's office? If so, is there something the provider can do to improve the individual's ability to speak for themselves?
Behavior patterns and triggers	• Circumstances of disruptive behavior (recognized triggers, patterns, motivation and reinforcement, and location)	• Query perceived function to behavior surrounding distress
Sleep	• Time of sleep onset and awakening • Middle-of-the-night interruptions • Naps • Activities interfering with sleep onset or returning to sleep • Intermittent periods of decreased need for sleep	• Food seeking as an activity interfering with sleep • Sleep apnea diagnosis and symptoms • Discomfort precipitating sleep disturbance
Anxiety	• Leading to significant outbursts or discomfort: transitioning activities, getting stuck on certain topics or things, and minor changes in daily activities	• Panic and nightmares • Sensory sensitivity that leads to discomfort • Repeated checking or rituals, which interferes with daily activities
Depression	• Less energy than usual, increased crying spells, sadness, irritability, isolative, loss of interest in activities typically enjoyed, and excess sleep • Injures self on purpose; if yes: location of injury and whether self-injury is causing discomfort?	• Whether self-injury is concerning to parent or caregiver • Whether self-injury could be perpetuated by attention seeking or avoidance
Mania	• Establish baseline energy • Query discrete periods out of the blue lasting ≥2 days of increased energy compared to baseline, laughing or vocalizing for no clear reason, particularly happy or giddy, risk taking, sexually acting out, increased impulsivity, and decreased need for sleep	• No changes made
Psychosis	• Appearing to be responding to internal auditory or visual stimuli • Yelling angrily in a room where no one else is present as if yelling at someone who is not there	• No changes made
ADHD[a]	• Difficulty following through on instructions, avoiding task demands, easily distractible, fidgety or restless, high activity when expected to remain in 1 place, constantly moving, blurting into other people's conversations, and demanding attention or desired items	• Excessive talking
General medical problems	• Query history of headaches, seizures, injuries that can be causing discomfort, thyroid abnormalities, and tooth pain • Could any of these issues be contributing to distress?	• Are there unusual ways of responding to physical discomfort? • Added joint pain; ear, nose, or throat pain; and seasonal allergies
Trauma[b]	• N/A[c]	• History of trauma • Related to trauma: avoidance, flashbacks, and nightmares • Hypervigilance
Gastrointestinal concerns[b]	• N/A	• Bowel movement frequency • Query history of constipation, stool accidents, frequent stomachaches, food allergies, and acid reflux. Could any of these issues contribute to distress? • Subsequent additions: changes in appetite, nausea, and variable bowel movements

Module	Original items	Added in response to feedback
Menstrual concerns[b] (for female patients only)	• N/A	• Query presence of mood changes during menses, endometriosis, polycystic ovary syndrome, significant menstrual pain, excess bleeding during or between cycles, and anxiety surrounding periods. Could any of these conditions be leading to distress? • Birth control: oral contraceptives, hormonal IUD[d], nonhormonal IUD, and Depo-Provera (a contraceptive injection).
Dental concerns[b]	• N/A	• When was the last dental visit? • Query presence of changes in eating patterns: texture preference, sensitivity to hot or cold food or drink preference for eating on 1 side of the mouth, and reduced oral intake • Grinding teeth

[a]ADHD: attention-deficit/hyperactivity disorder.

[b]Module added in response to focus group feedback.

[c]N/A: not applicable.

[d]IUD: intrauterine device.

Initial Survey Evaluation

Focus Group Feedback

During the focus groups, 3 main themes emerged in this analysis.

- Theme A: respondents gave overall positive feedback regarding existing content and specific feedback regarding areas where there was room to expand content. Table 1 describes the modules and items added in response to this feedback. Notably, a posttraumatic stress disorder (PTSD) module was added along with a PTSD scoring algorithm to determine positive screen status.
- Theme B: most of the respondents (15/17, 88%) agreed that the symptoms queried matched their understanding of the psychiatric and medical conditions to which they are attributed.
- Theme C: all participant groups reported positive acceptability of the branching logic format and time required to complete the measure.

Expert Review

Overall, the expert review supported the Sources of Distress categories and respective items attributed to each condition. One expert recommended adding items that query gender and replacing sex as the basis for pronoun selection within the tool and its report. This expert also suggested that the report include screening results for each psychiatric condition. The former recommendations were implemented when Sources of Distress was transitioned to the REDCap platform. The latter recommendation was deferred until after screening algorithms are validated in a clinical setting.

Clinical Implementation

Sample Characteristics

Surveys (N=264) were completed by parents or guardians (n=200, 75.8%), professional caregivers (n=43, 16.3%), and other caregivers (n=21, 8%) of 231 individuals (n=161, 69.7% men and boys; n=69, 29.9% women and girls; and n=1, 0.4% other; mean age 17.7, SD 10.3; range 2-65 years). Informal (n=62, 41.6%) and formal (n=87, 58.4%) consults were performed for 149 individuals collectively. Table 2 presents sample characteristics, the manifestations of distress, and a comparison between individuals with a consult and those without.

Table 2. Sample characteristics and distress manifestations.

Characteristics	With consult[a] (n=149), n (%)	Without consult (n=82), n (%)	Total (N=231), n (%)	Chi-square (*df*)	*P* value
Gender[b]				2.3 (2)	.34
Man or boy	102 (68.5)	59 (72)	161 (69.7)		
Woman or girl	47 (31.5)	22 (25.5)	69 (29.9)		
Other[b]	0 (0)	1 (1.2)	1 (0.4)		
Caregiver[c]				7.8 (2)	.02
Parent or guardian	108 (72.5)	71 (86.6)	179 (77.5)		
Professional caregiver	32 (21.5)	6 (7.3)	38 (16.5)		
Other	9 (6)	5 (6.1)	14 (6.1)		
Age range (y)				7.4 (2)	.03
<13	46 (30.9)	38 (46.3)	84 (36.4)		
13-22	57 (38.3)	30 (36.6)	87 (37.7)		
>22	46 (30.9)	14 (17.1)	60 (26)		
Language ability				0.0 (2)	.99
Full verbal ability	76 (51)	42 (51.2)	118 (51.1)		
Limited use of words	46 (30.9)	25 (30.5)	71 (30.7)		
Nonverbal	27 (18.1)	15 (18.3)	42 (18.2)		
Manifestation of distress					
Agitation	130 (87.2)	70 (85.4)	200 (86.6)	0.2 (1)	.69
Aggression	97 (65.1)	50 (61)	147 (63.6)	0.4 (1)	.53
Change in sleep	86 (57.7)	38 (46.3)	124 (53.7)	2.8 (1)	.10
Moodiness	122 (81.9)	65 (79.3)	187 (81)	0.2 (1)	.63
Increased fixation	115 (77.2)	57 (69.5)	172 (74.5)	1.6 (1)	.20
Change in eating patterns	52 (34.9)	19 (23.2)	71 (30.7)	3.4 (1)	.07
Change in personality	99 (66.4)	59 (72.0)	158 (68.4)	0.7 (1)	.39
Change in behavior	96 (64.4)	45 (54.9)	141 (61)	2.0 (1)	.15
Self-injurious behavior	73 (49)	39 (47.6)	112 (48.5)	0.0 (1)	.84
Type of disability					
Autism without ID[d]	52 (34.9)	47 (57.3)	99 (42.9)	10.9 (1)	<.001
ID without autism	15 (10.1)	7 (8.5)	22 (9.5)	0.1 (1)	.71
ID and autism	74 (49.7)	18 (22.0)	92 (39.8)	17.0 (1)	<.001
Genetic syndrome[e]	17 (11)	17 (20.7)	34 (14.7)	3.7 (1)	.06

[a]Includes informal and formal consults.

[b]One participant reported *other* as gender: no participants reported *non-binary* as gender.

[c]When multiple caregivers completed Sources of Distress, the report from the caregiver with whom the participant spends the most time was used in this table.

[d]ID: intellectual disability.

[e]Genetic syndrome includes some individuals who also populate the autism or ID categories.

Preexisting Psychiatric Diagnoses

The presence of at least 1 preexisting psychiatric diagnosis was reported in 65.4% (151/231) of the individuals. Individuals who received a consult compared to those without a consult were more likely to have a caregiver-reported history of psychotic disorder (14/149, 9.4% vs 1/82, 1%; *P*=.02; Table 3).

Table 3. Medical conditions, preexisting psychiatric diagnoses, and psychiatric screening results.

Characteristics	With consult[a] (n=149), n (%)	Without consult (n=82), n (%)	Total (N=231), n (%)	Chi-square (*df*)	*P* value
Medical conditions[b]					
Gastrointestinal concerns	82 (55)	37 (45.1)	119 (51.5)	2.0 (1)	.15
Dental concerns	32 (21.5)	25 (30.5)	57 (24.7)	2.3 (1)	.13
Menstrual concerns[c]	16 (53.3)	5 (27.8)	21 (43.8)	3.0 (1)	.13
General					
Headache	26 (17.4)	8 (9.9)	34 (14.7)	2.5 (1)	.11
Ear, nose, and throat concerns	17 (11.4)	8 (9.9)	25 (10.8)	0.2 (1)	.70
Seasonal allergies	34 (22.8)	13 (15.9)	47 (20.3)	1.6 (1)	.21
Injury pain	14 (9.4)	9 (11)	23 (10)	0.2 (1)	.70
Thyroid abnormalities	5 (3.4)	6 (7.3)	11 (4.8)	1.8 (1)	.18
Joint pain	9 (6)	3 (3.7)	12 (5.2)	0.6 (1)	.44
Seizures	29 (19.5)	16 (19.5)	45 (19.5)	0.0 (1)	.99
Seizure History	40 (26.8)	18 (22)	58 (25.1)	0.7 (1)	.41
Sleep disturbance	124 (83.2)	68 (82.9)	192 (83.1)	0.0 (1)	.95
Preexisting psychiatric diagnoses					
Any psychiatric condition	103 (69.1)	48 (58.5)	151 (65.4)	2.6 (1)	.11
Depression[d]	20 (13.4)	14 (17.1)	34 (14.7)	0.6 (1)	.46
Bipolar disorder[d]	21 (14.1)	9 (11)	30 (13)	0.5 (1)	.50
Unspecified mood disorder or DMDD[d,e]	20 (13.4)	7 (8.5)	27 (11.7)	1.2 (1)	.27
Anxiety[f]	62 (41.6)	31 (37.8)	93 (40.3)	0.3 (1)	.57
PTSD[g]	11 (7.4)	4 (4.9)	15 (6.5)	0.6 (1)	.46
Psychotic disorder	14 (9.4)	1 (1.2)	15 (6.5)	5.8 (1)	.02
ADHD[h]	50 (33.6)	26 (31.7)	76 (32.9)	0.1 (1)	.78
Psychiatric screening status					
Any psychiatric condition	146 (98)	80 (97.6)	226 (97.8)	0.1 (1)	.83
Unipolar depression[d]	61 (40.9)	30 (36.6)	91 (39.4)	0.4 (1)	.52
Episodic expansive mood and bipolar disorder[d]	60 (40.3)	28 (34.1)	88 (38.1)	0.8 (1)	.36
Anxiety	130 (87.2)	71 (86.6)	201 (87)	0.0 (1)	.89
PTSD	37 (24.8)	15 (18.3)	52 (22.5)	1.3 (1)	.26
Psychosis	52 (34.9)	15 (18.3)	67 (29)	7.1 (1)	.008
ADHD	102 (68.5)	56 (68.3)	158 (68.4)	0.0 (1)	.98

[a]Includes informal and formal consults.

[b]Medical conditions perceived by the caregiver as contributing to the current presentation of distress.

[c]Analysis for menstrual concerns restricted to female patients aged >12.

[d]Unipolar depression, unspecified mood disorder and disruptive mood dysregulation disorder, and episodic expansive mood and bipolar disorder are mutually exclusive categories.

[e]DMDD: disruptive mood dysregulation disorder.

[f]Preexisting diagnosis of obsessive-compulsive disorder is included within the anxiety disorder category.

[g]PTSD: posttraumatic stress disorder.

[h]ADHD: attention-deficit/hyperactivity disorder.

Caregiver-Reported Medical Conditions

Table 3 describes medical conditions reported by caregivers. Caregivers of 73.2% (169/231) of the individuals identified at least 1 physical concern that they perceived as contributing to distress. The most common conditions were gastrointestinal concerns (119/231, 51.5%), menstrual concerns (21/48, 44% of female patients aged >12 y), seasonal allergies (47/231, 20.3%), and seizures (45/231, 19.5%).

Psychiatric Screening Results

Table 3 lists the frequency of positive psychiatric screening results. All but 2% (5/231) of the individuals screened positive for a psychiatric condition, with a mean of 2.8 (SD 1.1; range 0-5) conditions per individual. Of those who were classified as having bipolar disorder, 89% (78/88) screened positive for a recent depressive episode. Positive screen status for psychiatric conditions were similar between those with a consult and those without, except in the case of psychosis (52/149, 34.9% vs 15/82, 18%; *P*=.008).

Psychotropic Medication Use

Table 4 reports on the frequency of medication use. Most of the individuals (194/231, 84%) were taking psychotropic medication, and the majority were receiving antipsychotics (142/231, 61.5%) and antidepressants (129/231, 55.8%).

Table 4. Medication use reported in Sources of Distress.

Medication	With consult (n=149), n (%)	Without consult (n=82), n (%)	Total (N=231), n (%)	Chi-square (*df*)	*P* value
Any medication	146 (98)	68 (82.9)	214 (92.6)	17.6 (1)	<.001
Any psychotropic medication	136 (91.3)	58 (70.7)	194 (84)	16.6 (1)	<.001
Antipsychotic	102 (68.5)	40 (48.8)	142 (61.5)	8.7 (1)	.003
Antidepressant[a]	87 (58.4)	42 (51.2)	129 (55.8)	1.1 (1)	.29
Anxiolytic[b]	66 (44.3)	23 (28.0)	89 (38.5)	5.9 (1)	.02
Anticonvulsant[c]	45 (30.2)	12 (14.6)	57 (24.7)	6.9 (1)	.009
Lithium	10 (6.7)	7 (8.5)	17 (7.4)	0.3 (1)	.61
Alpha-2 agonist	72 (48.3)	27 (32.9)	100 (43.3)	5.1 (1)	.02
Stimulant and atomoxetine	30 (20.1)	14 (17.1)	44 (19)	0.3 (1)	.57

[a]Selective serotonin reuptake inhibitors, duloxetine, tricyclics, mirtazapine, and trazodone were included exclusively within the antidepressant category.
[b]Benzodiazapines, buspirone, hydroxyzine, beta-blockers, and prazosin were included exclusively within the anxiolytic category.
[c]Anticonvulsant medication use in the absence of a reported seizure history.

Working Psychiatric Diagnoses

Of the 149 individuals who received a consult, 148 (99.3%) were diagnosed with at least 1 psychiatric condition with a mean of 2.7 (SD 1.0; range 0-5) diagnoses per individual. The conditions identified were anxiety (129/149, 86.6%), ADHD (84/149, 56.4%), bipolar disorder (67/149, 45%), unipolar depression (33/149, 22.1%), PTSD (35/149, 23.5%), and psychosis (31/149, 20.8%). Furthermore, 25 (16.8%) of the 149 individuals were diagnosed with either unspecified mood disorder or DMDD. Nearly all individuals identified with psychosis (29/31, 94%) had a co-occurring mood disorder diagnosis: bipolar disorder (22/31, 71%), unipolar depression (5/31, 16%), and unspecified mood disorder or DMDD (2/31, 6%).

Evaluation of Validity

Sources of Distress accuracy rates ranged from 76% (95% CI 69%-82%) for ADHD to 91% (95% CI 85%-95%) for PTSD and exceeded those of preexisting diagnoses, except in the case of psychosis, for which the accuracy rates were equivocal (82%, 95% CI 75%-87%; Table 5). The survey demonstrated higher NPVs (81%-98%) than PPVs (51%-78%) for all conditions, with the exceptions of anxiety (53% and 92%, respectively) and episodic expansive mood bipolar disorder (85% and 90%, respectively). Low PPVs were notable for depression (51%) and psychosis (54%).

Table 5. Association between consult diagnoses after completing Sources of Distress with preexisting psychiatric diagnoses and Sources of Distress screening status (n=149).

Working diagnosis	Preexisting psychiatric diagnosis[a]					Sources of Distress screening status				
	Case negative, n	Case positive, n	PPV[b] (%)	NPV[c] (%)	Accuracy rate[d], % (95% CI)	Screen negative, n	Screen positive, n	PPV (%)[b]	NPV (%)[c]	Accuracy rate[d], % (95% CI)
Unipolar depression[e,f]			50	82	78 (71-84)			51	98	79 (71-85)
Case negative	106	10				86	30			
Case positive	23	10				2	31			
Episodic expansive mood and bipolar disorder[e,f]			76	60	62 (55-70)			90	85	87 (81-92)
Case negative	77	5				76	6			
Case positive	51	16				13	54			
DMDD[g] and unspecified mood disorder[e]			45	88	82 (75-88)			N/A[h]	N/A	N/A
Case negative	113	11				N/A	N/A			
Case positive	16	9				N/A	N/A			
Anxiety disorder[i]			97	20	52 (44-60)			92	53	87 (81-92)
Case negative	18	2				10	10			
Case positive	69	60				9	120			
Posttraumatic stress disorder			100	83	84 (77-89)			78	95	91 (85-95)
Case negative	114	0				106	8			
Case positive	24	11				6	29			
Psychotic disorder[j]			64	84	82 (75-87)			54	97	82 (75-87)
Case negative	113	5				94	24			
Case positive	22	9				3	28			
Attention-deficit/hyperactivity disorder			84	58	66 (59-74)			74	81	76 (69-82)
Case negative	57	8				38	27			
Case positive	42	42				9	75			

[a]Preexisting diagnoses included caregiver-reported diagnoses in Sources of Distress and diagnoses in the medical record before survey completion.

[b]PPV: positive predictive value =

[c]NPV: negative predictive value =

[d]Accuracy rate =

[e]Depression, episodic expansive mood and bipolar disorder, and disruptive mood dysregulation disorder and unspecified mood disorder are mutually exclusive categories. There is no Sources of Distress screening algorithm for disruptive mood dysregulation disorder or unspecified mood disorder.

[f]Preexisting and working diagnoses included schizoaffective disorder when hypomanic, manic, or mixed episode was specified.

[g]DMDD: disruptive mood dysregulation disorder.

[h]N/A: not applicable.

[i]Preexisting and working diagnoses of anxiety disorder and obsessive-compulsive disorder are combined to coincide with anxiety disorder screening status.

[j]Preexisting and working diagnoses were schizophrenia, schizoaffective disorder, unspecified psychotic disorder, and psychotic features associated with a mood disorder.

Exploration of Mood Disorder Categories

Figure 1 demonstrates the distribution of mood disorder diagnoses among individuals based on (1) preexisting mood disorder diagnosis and (2) Sources of Distress mood disorder screening status. The majority of the individuals (18/25, 72%) who received a working diagnosis of unspecified mood disorder and DMDD screened positive for either unipolar depression or bipolar disorder.

Figure 1. Comparison of mood disorder categorization between working diagnosis established during consultation and (A) preexisting diagnosis and (B) Sources of Distress positive screen. DMDD: disruptive mood dysregulation disorder.

Discussion

Principal Findings

The focus group feedback indicates that Sources of Distress provides an acceptable means for caregivers to share their knowledge and insights about individuals with DD who present in crisis. As a screening tool, this survey demonstrates good accuracy, although additional work is needed to differentiate among mood disorders. The purpose of this survey is to screen individuals with DD for mental health and common medical concerns in health care settings when they present in crisis. By querying what symptom clusters and physical conditions coincide with their patient's crisis, providers can direct their evaluation toward specific psychiatric and medical conditions that have established treatment protocols in the general population. This approach aims to reduce diagnostic overshadowing and improve medical decision-making surrounding the management of agitation and aggression in individuals with DD. Focus group participants validated the survey content and provided recommendations that prompted the inclusion of additional modules and items. Despite its length (ie, 15-20 min), participants reported positive acceptability of the survey's format and duration. This feedback may reflect the convenience of completing a web-based survey at home versus in the medical setting and highlights caregivers' motivation toward understanding potential factors contributing to the person's distress. After incorporating caregiver recommendations, Sources of Distress content was also reviewed and supported by clinical and research experts.

Caregivers of most of the individuals (200/231, 86.6%) identified agitation as a presenting concern. The Food and Drug Administration has approved short-term antipsychotic medication for treating irritability in individuals with autism [47]; 61.5% (142/231) of the individuals were taking antipsychotics at the time of presenting in crisis. This frequency exceeds previously reported estimates of antipsychotic use in the population with DD (ie, 10%-48%) and reflects the high acuity and potentially treatment-resistant nature of individuals for whom the survey was completed [48,49]. This study group's acuity is further supported by the high frequency in which severe mental health conditions were diagnosed in those receiving a consultation (eg, bipolar disorder and psychosis).

Anxiety was the most common condition to screen positive (201/231, 87%) and be established as a working diagnosis (129/149, 86.6%). These rates exceeded measured anxiety prevalence rates in the population with DD (ie, 20%-77%), indicating a higher propensity toward experiencing anxiety among those presenting in crisis [12,15,16]. As a precipitant of distress, prior studies have identified aggression, disruptive behavior, sleep disturbance, and self-injurious behavior as symptoms of anxiety among individuals with DD [4,8,50]. To reduce overclassification among individuals whose autism core features overlap with some anxiety symptoms [51], the Sources of Distress anxiety scoring algorithm was set at a higher threshold than the generalized anxiety disorder criteria described in *DM-ID-2*. The survey's low NPV (53%) and high PPV (92%) for anxiety likely reflect this adaption.

Sources of Distress captured well the presence of a mood disturbance; however, the type of mood disorder was not. Study results report a diagnosis frequency of 16.8% (25/149) for unspecified mood disorder and DMDD and indicate the need to add items and a screening algorithm for this condition. The low PPV (51%) for depression primarily resulted from individuals screening positive for depression who were subsequently diagnosed with unspecified mood disorder and DMDD. The *DM-ID-2*, survey data, and records review will inform new items and algorithm development as well as revisions for the depression screening algorithm. In the interim, the Sources of Distress report will replace the "depression" category label with "depression and unspecified mood disorder" to broaden the range of conditions which it currently captures.

Caregivers of the majority of the individuals (169/231, 73.2%) identified at least 1 physical concern that they perceived as contributing to distress. As agitation may be one of the few visible indicators of pain in an individual with limited expressive language ability and DD, sources of pain should be considered when unexplained agitation is present [3,9,52]. Limited access to medical care by the population with DD further reduces the likelihood that pain and other underlying physical causes of agitation are recognized [53]. Through Sources of Distress, caregivers demonstrated their ability to provide meaningful insight into the potential presence of physical discomfort. This attention was directed most frequently to gastrointestinal, menstrual, dental, and seizure concerns.

Limitations

The generalizability of study results is limited to the geographic, racial, and ethnic diversity of Utah. While survey access requires internet or smartphone access, it has been completed by parents without this access through the assistance of state-sponsored support coordinators and medical assistants. Sources of Distress has a Spanish translation available (*Causas de Aflicción*); however, these data were not included because its content has not yet been validated by Spanish-speaking caregivers and individuals who are affected. The expert leading the consult team was a member of the survey's development team, which introduces the inherent bias of evaluating for the presence of mental health conditions through the lens of *DM-ID-2* criteria on which survey components were also based. While the *DM-ID-2* is well recognized and accepted in the ID provider community, few autism specialty providers are familiar with its use.

Future Directions

Edits and additions to mood disorder items and scoring algorithms are being made to improve differentiation across mood disorders. Branching logic that incorporates the individual's language ability has recently been added to the psychosis module to improve question clarity and scoring algorithm accuracy. The most updated version of the Sources of Distress can be accessed through the Utah Department of Health and Human Services Autism Systems Development Program webpage [39]. Reevaluation of the survey's PPVs, NPVs, and accuracy will follow the completion of these changes. Additional studies of this survey are needed to measure its acceptability and validity in clinical settings outside of Utah and by other DD specialty providers. REDCap has also demonstrated capacity to integrate digital mental health screening results into electronic medical records, significantly improving provider adoption of the screening tools [54]. The integration of Sources of Distress into electronic medical records could further enhance its impact on provider efficiency. This survey has already been used during medical evaluations to facilitate the consideration of potential discordance between medications prescribed and conditions present [55]. Prospective studies are merited to determine the survey's impact on treatment approaches, hospital and emergency department use, and outcomes for individuals with DD who experience crisis.

Conclusions

Individuals with DD presenting in crisis experience high rates of psychiatric disorders and medical concerns that may contribute to, or manifest as, distress. Sources of Distress is a valuable screening tool for psychiatric and medical conditions that commonly accompany treatment-resistant agitation in individuals with DD. When systematically queried, caregivers' knowledge provides essential information to minimize diagnostic overshadowing and support an evaluation focused on the individual rather than their disability when persistent agitation is assessed in the population with DD.

Acknowledgments

The authors thank the individuals with developmental disabilities and their parents and caregivers whose support and participation shaped the development, validation, and piloting of the Sources of Distress. In particular, the authors appreciate the insight and assistance provided by Jaye Olafson. The Autism Council of Utah funded the development of this survey, and access to REDCap (Research Electronic Data Capture) is supported through UM1TR004409 NCATS/NIH.

Conflicts of Interest

The Sources of Distress survey is copyrighted by the University of Utah; DAB, JD, and WW are listed as inventors. DAB consults for BioMarin Pharmaceutical Inc, Encoded Therapeutics, Taysha Gene Therapies, and Synlogic Therapeutics and attended an advisory board meeting for Sanofi. All other authors declare no other conflicts of interest.

Multimedia Appendix 1
A sample of the Sources of Distress report.
[PDF File (Adobe PDF File), 80 KB - mental_v11i1e50907_app1.pdf]

References

1. Durbin A, Balogh R, Lin E, Wilton AS, Selick A, Dobranowski KM, et al. Repeat emergency department visits for individuals with intellectual and developmental disabilities and psychiatric disorders. Am J Intellect Dev Disabil 2019 May;124(3):206-219. [doi: 10.1352/1944-7558-124.3.206] [Medline: 31026200]
2. Lunsky Y, Lin E, Balogh R, Klein-Geltink J, Wilton AS, Kurdyak P. Emergency department visits and use of outpatient physician services by adults with developmental disability and psychiatric disorder. Can J Psychiatry 2012 Oct 01;57(10):601-607. [doi: 10.1177/070674371205701004] [Medline: 23072951]
3. Antonacci DJ, Manuel C, Davis E. Diagnosis and treatment of aggression in individuals with developmental disabilities. Psychiatr Q 2008 Sep 23;79(3):225-247. [doi: 10.1007/s11126-008-9080-4] [Medline: 18726157]
4. Kim JA, Szatmari P, Bryson SE, Streiner DL, Wilson FJ. The prevalence of anxiety and mood problems among children with Autism and Asperger syndrome. Autism 2016 Jun 30;4(2):117-132. [doi: 10.1177/1362361300004002002]
5. Kishore MT, Nizamie SH, Nizamie A. The behavioural profile of psychiatric disorders in persons with intellectual disability. J Intellect Disabil Res 2005 Nov 06;49(Pt 11):852-857. [doi: 10.1111/j.1365-2788.2005.00763.x] [Medline: 16207283]
6. Lowe K, Allen D, Jones E, Brophy S, Moore K, James W. Challenging behaviours: prevalence and topographies. J Intellect Disabil Res 2007 Aug 09;51(Pt 8):625-636. [doi: 10.1111/j.1365-2788.2006.00948.x] [Medline: 17598876]
7. McGuire K, Fung LK, Hagopian L, Vasa RA, Mahajan R, Bernal P, et al. Irritability and problem behavior in autism spectrum disorder: a practice pathway for pediatric primary care. Pediatrics 2016 Feb;137 Suppl 2:S136-S148. [doi: 10.1542/peds.2015-2851L] [Medline: 26908469]
8. Oliver C, Licence L, Richards C. Self-injurious behaviour in people with intellectual disability and autism spectrum disorder. Curr Opin Psychiatry 2017 Mar;30(2):97-101 [FREE Full text] [doi: 10.1097/YCO.0000000000000313] [Medline: 28030375]
9. Regnard C, Reynolds J, Watson B, Matthews D, Gibson L, Clarke C. Understanding distress in people with severe communication difficulties: developing and assessing the Disability Distress Assessment Tool (DisDAT). J Intellect Disabil Res 2007 Apr 29;51(Pt 4):277-292. [doi: 10.1111/j.1365-2788.2006.00875.x] [Medline: 17326809]
10. Rojahn J, Matson JL, Naglieri JA, Mayville E. Relationships between psychiatric conditions and behavior problems among adults with mental retardation. Am J Ment Retard 2004 Jan;109(1):21-33. [doi: 10.1352/0895-8017(2004)109<21:RBPCAB>2.0.CO;2] [Medline: 14651452]
11. Brookman-Frazee L, Stadnick N, Chlebowski C, Baker-Ericzén M, Ganger W. Characterizing psychiatric comorbidity in children with autism spectrum disorder receiving publicly funded mental health services. Autism 2018 Nov 15;22(8):938-952 [FREE Full text] [doi: 10.1177/1362361317712650] [Medline: 28914082]
12. Buck TR, Viskochil J, Farley M, Coon H, McMahon WM, Morgan J, et al. Psychiatric comorbidity and medication use in adults with autism spectrum disorder. J Autism Dev Disord 2014 Dec 24;44(12):3063-3071 [FREE Full text] [doi: 10.1007/s10803-014-2170-2] [Medline: 24958436]
13. Eaves LC, Ho HH. Young adult outcome of autism spectrum disorders. J Autism Dev Disord 2008 Apr 1;38(4):739-747. [doi: 10.1007/s10803-007-0441-x] [Medline: 17764027]
14. Leyfer OT, Folstein SE, Bacalman S, Davis NO, Dinh E, Morgan J, et al. Comorbid psychiatric disorders in children with autism: interview development and rates of disorders. J Autism Dev Disord 2006 Oct 15;36(7):849-861. [doi: 10.1007/s10803-006-0123-0] [Medline: 16845581]
15. Lai MC, Kassee C, Besney R, Bonato S, Hull L, Mandy W, et al. Prevalence of co-occurring mental health diagnoses in the autism population: a systematic review and meta-analysis. Lancet Psychiatry 2019 Oct;6(10):819-829. [doi: 10.1016/S2215-0366(19)30289-5] [Medline: 31447415]

16. Lecavalier L, McCracken CE, Aman MG, McDougle CJ, McCracken JT, Tierney E, et al. An exploration of concomitant psychiatric disorders in children with autism spectrum disorder. Compr Psychiatry 2019 Jan;88:57-64 [FREE Full text] [doi: 10.1016/j.comppsych.2018.10.012] [Medline: 30504071]
17. Morgan CN, Roy M, Chance P. Psychiatric comorbidity and medication use in autism: a community survey. Psychiatr Bull 2018 Jan 02;27(10):378-381. [doi: 10.1192/pb.27.10.378]
18. Simonoff E, Pickles A, Charman T, Chandler S, Loucas T, Baird G. Psychiatric disorders in children with autism spectrum disorders: prevalence, comorbidity, and associated factors in a population-derived sample. J Am Acad Child Adolesc Psychiatry 2008 Aug;47(8):921-929. [doi: 10.1097/CHI.0b013e318179964f] [Medline: 18645422]
19. Varcin KJ, Herniman SE, Lin A, Chen Y, Perry Y, Pugh C, et al. Occurrence of psychosis and bipolar disorder in adults with autism: a systematic review and meta-analysis. Neurosci Biobehav Rev 2022 Mar;134:104543. [doi: 10.1016/j.neubiorev.2022.104543] [Medline: 35063494]
20. Mazza MG, Rossetti A, Crespi G, Clerici M. Prevalence of co-occurring psychiatric disorders in adults and adolescents with intellectual disability: a systematic review and meta-analysis. J Appl Res Intellect Disabil 2020 Mar;33(2):126-138. [doi: 10.1111/jar.12654] [Medline: 31430018]
21. Platt JM, Keyes KM, McLaughlin KA, Kaufman AS. Intellectual disability and mental disorders in a US population representative sample of adolescents. Psychol Med 2019 Apr;49(6):952-961 [FREE Full text] [doi: 10.1017/S0033291718001605] [Medline: 29996960]
22. Munir KM. The co-occurrence of mental disorders in children and adolescents with intellectual disability/intellectual developmental disorder. Curr Opin Psychiatry 2016 Mar;29(2):95-102 [FREE Full text] [doi: 10.1097/YCO.0000000000000236] [Medline: 26779862]
23. Smit MJ, Scheffers M, Emck C, van Busschbach JT, Beek PJ. Clinical characteristics of individuals with intellectual disability who have experienced sexual abuse. An overview of the literature. Res Dev Disabil 2019 Dec;95:103513. [doi: 10.1016/j.ridd.2019.103513] [Medline: 31654947]
24. Jamieson D, Mason J. Investigating the existence of the diagnostic overshadowing bias in Australia. J Ment Health Res Intellect Disabil 2019 Apr 22;12(1-2):58-70. [doi: 10.1080/19315864.2019.1595231]
25. Joshi G, Biederman J, Petty C, Goldin RL, Furtak SL, Wozniak J. Examining the comorbidity of bipolar disorder and autism spectrum disorders: a large controlled analysis of phenotypic and familial correlates in a referred population of youth with bipolar I disorder with and without autism spectrum disorders. J Clin Psychiatry 2013 Jun;74(6):578-586. [doi: 10.4088/JCP.12m07392] [Medline: 23842009]
26. Rapoport J, Chavez A, Greenstein D, Addington A, Gogtay N. Autism spectrum disorders and childhood-onset schizophrenia: clinical and biological contributions to a relation revisited. J Am Acad Child Adolesc Psychiatry 2009 Jan;48(1):10-18 [FREE Full text] [doi: 10.1097/CHI.0b013e31818b1c63] [Medline: 19218893]
27. Bertelli M, Scuticchio D, Ferrandi A, Lassi S, Mango F, Ciavatta C, et al. Reliability and validity of the SPAID-G checklist for detecting psychiatric disorders in adults with intellectual disability. Res Dev Disabil 2012;33(2):382-390. [doi: 10.1016/j.ridd.2011.08.020] [Medline: 22119685]
28. Fletcher RJ, Barnhill J, Cooper SA. Diagnostic Manual—Intellectual Disability 2 (DM-ID): A Textbook of Diagnosis of Mental Disorders in Persons with Intellectual Disability. Washington, DC: NADD Press; 2016.
29. Fletcher R, Loschen E, Stavrakiki C, Kingston MF. Diagnostic Manual - Intellectual Disability: A Textbook of Diagnosis of Mental Disorders in Persons With Intellectual Disability. New York, NY: NADD Press; 2007.
30. Moss S, Patel P, Prosser H, Goldberg D, Simpson N, Rowe S, et al. Psychiatric morbidity in older people with moderate and severe learning disability. I: development and reliability of the patient interview (PAS-ADD). Br J Psychiatry 1993 Oct;163:471-480. [doi: 10.1192/bjp.163.4.471] [Medline: 8252286]
31. Matson JL, Kazdin AE, Senatore V. Psychometric properties of the psychopathology instrument for mentally retarded adults. Appl Res Ment Retard 1984;5(1):81-89. [doi: 10.1016/s0270-3092(84)80021-1] [Medline: 6721483]
32. Devine M, Taggart L, McLornian P. Screening for mental health problems in adults with learning disabilities using the Mini PAS‐ADD Interview. Br J Learn Disabil 2010 Nov 04;38(4):252-258. [doi: 10.1111/j.1468-3156.2009.00597.x]
33. Moss S, Prosser H, Costello H, Simpson N, Patel P, Rowe S, et al. Reliability and validity of the PAS-ADD Checklist for detecting psychiatric disorders in adults with intellectual disability. J Intellect Disabil Res 1998 Apr;42 (Pt 2):173-183. [doi: 10.1046/j.1365-2788.1998.00116.x] [Medline: 9617701]
34. Sevin JA, Matson JL, Williams D, Kirkpatrick-Sanchez S. Reliability of emotional problems with the diagnostic assessment for the Severely Handicapped (DASH). Br J Clin Psychol 1995 Feb;34(1):93-94. [doi: 10.1111/j.2044-8260.1995.tb01442.x] [Medline: 7757045]
35. Myrbakk E, von Tetzchner S. Screening individuals with intellectual disability for psychiatric disorders: comparison of four measures. Am J Ment Retard 2008 Jan;113(1):54-70. [doi: 10.1352/0895-8017(2008)113[54:SIWIDF]2.0.CO;2] [Medline: 18173300]
36. Taylor JL, Hatton C, Dixon L, Douglas C. Screening for psychiatric symptoms: PAS-ADD Checklist norms for adults with intellectual disabilities. J Intellect Disabil Res 2004 Jan;48(1):37-41. [doi: 10.1111/j.1365-2788.2004.00585.x] [Medline: 14675229]

37. Marcangelo MJ, Ovsiew F. Psychiatric aspects of epilepsy. Psychiatr Clin North Am 2007 Dec;30(4):781-802. [doi: 10.1016/j.psc.2007.07.005] [Medline: 17938045]
38. Périsse D, Amiet C, Consoli A, Thorel MV, Gourfinkel-An I, Bodeau N, et al. Risk factors of acute behavioral regression in psychiatrically hospitalized adolescents with autism. J Can Acad Child Adolesc Psychiatry 2010 May;19(2):100-108 [FREE Full text] [Medline: 20467546]
39. Bilder DA. Sources of distress - Autism Systems Development. Utah Department of Health and Human Services. URL: https://familyhealth.utah.gov/cshcn/asd/# [accessed 2024-03-11]
40. Martin-Key NA, Spadaro B, Funnell E, Barker EJ, Schei TS, Tomasik J, et al. The current state and validity of digital assessment tools for psychiatry: systematic review. JMIR Ment Health 2022 Mar 30;9(3):e32824 [FREE Full text] [doi: 10.2196/32824] [Medline: 35353053]
41. Home page. Autism Council of Utah. URL: https://www.autismcouncilofutah.org/ [accessed 2024-02-11]
42. Kaufman J, Birmaher B, Brent D, Rao U, Flynn C, Moreci P, et al. Schedule for affective disorders and schizophrenia for school-age children-present and lifetime version (K-SADS-PL): initial reliability and validity data. J Am Acad Child Adolesc Psychiatry 1997 Jul;36(7):980-988 [FREE Full text] [doi: 10.1097/00004583-199707000-00021] [Medline: 9204677]
43. Ritchie J, Spencer L. Qualitative data analysis for applied policy research. In: Bryman A, Burgess B, editors. Analyzing Qualitative Data. London, UK: Routledge; 1994.
44. Strauss A, Corbin J. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. 2nd edition. Thousand Oaks, CA: Sage Publications; 1998.
45. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, REDCap Consortium. The REDCap consortium: building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]
46. Custom template engine. GitHub. URL: https://github.com/BCCHR-IT/custom-template-engine, [accessed 2024-02-18]
47. LeClerc S, Easley D. Pharmacological therapies for autism spectrum disorder: a review. P T 2015 Jun;40(6):389-397 [FREE Full text] [Medline: 26045648]
48. Downs J, Hotopf M, Ford T, Simonoff E, Jackson RG, Shetty H, et al. Clinical predictors of antipsychotic use in children and adolescents with autism spectrum disorders: a historical open cohort study using electronic health records. Eur Child Adolesc Psychiatry 2016 Jun 15;25(6):649-658 [FREE Full text] [doi: 10.1007/s00787-015-0780-7] [Medline: 26472118]
49. Esbensen AJ, Greenberg JS, Seltzer MM, Aman MG. A longitudinal investigation of psychotropic and non-psychotropic medication use among adolescents and adults with autism spectrum disorders. J Autism Dev Disord 2009 Sep 12;39(9):1339-1349 [FREE Full text] [doi: 10.1007/s10803-009-0750-3] [Medline: 19434487]
50. Shui AM, Katz T, Malow BA, Mazurek MO. Predicting sleep problems in children with autism spectrum disorders. Res Dev Disabil 2018 Dec;83:270-279. [doi: 10.1016/j.ridd.2018.10.002] [Medline: 30393065]
51. Wood JJ, Gadow KD. Exploring the nature and function of anxiety in youth with autism spectrum disorders. Clin Psychol Sci Pract 2010 Dec;17(4):281-292. [doi: 10.1111/j.1468-2850.2010.01220.x]
52. Gössler A, Schalamon J, Huber-Zeyringer A, Höllwarth ME. Gastroesophageal reflux and behavior in neurologically impaired children. J Pediatr Surg 2007 Sep;42(9):1486-1490. [doi: 10.1016/j.jpedsurg.2007.04.009] [Medline: 17848235]
53. Matson JL, Rivet TT. Characteristics of challenging behaviours in adults with autistic disorder, PDD-NOS, and intellectual disability. J Intellect Dev Disabil 2008 Dec;33(4):323-329 [FREE Full text] [doi: 10.1080/13668250802492600] [Medline: 19039692]
54. Hawley S, Yu J, Bogetic N, Potapova N, Wakefield C, Thompson M, et al. Digitization of measurement-based care pathways in mental health through REDCap and electronic health record integration: development and usability study. J Med Internet Res 2021 May 20;23(5):e25656 [FREE Full text] [doi: 10.2196/25656] [Medline: 34014169]
55. Bartel RL, Knight JR, Worsham W, Bilder D. Discordance between psychiatric diagnoses and medication use in children and adults with autism presenting in crisis. Focus (forthcoming) 2024;22. [doi: 10.1176/appi.focus.20230027]

Abbreviations

ADHD: attention-deficit/hyperactivity disorder
DD: developmental disabilities
DMDD: disruptive mood dysregulation disorder
DM-ID-2: Diagnostic Manual–Intellectual Disability–Second Edition
ID: intellectual disability
NPV: negative predictive value
PAS-ADD: Psychiatric Assessment Schedule for Adults with Developmental Disabilities
PPV: positive predictive value
PTSD: posttraumatic stress disorder
REDCap: Research Electronic Data Capture

Edited by J Torous; submitted 17.07.23; peer-reviewed by H Onyeaka, A AL-Asadi; comments to author 04.02.24; revised version received 18.02.24; accepted 20.02.24; published 29.03.24.

Please cite as:
Bilder DA, Mthembu M, Worsham W, Aguayo P, Knight JR, Deng SW, Singh TP, Davis J
Developing and Implementing a Web-Based Branching Logic Survey to Support Psychiatric Crisis Evaluations of Individuals With Developmental Disabilities: Qualitative Study and Evaluation of Validity
JMIR Ment Health 2024;11:e50907
URL: https://mental.jmir.org/2024/1/e50907
doi:10.2196/50907
PMID:38551644

Original Paper

Tablet-Based Cognitive and Eye Movement Measures as Accessible Tools for Schizophrenia Assessment: Multisite Usability Study

Kentaro Morita[1,2], MD, PhD; Kenichiro Miura[3], PhD; Atsuhito Toyomaki[4], PhD; Manabu Makinodan[5], MD, PhD; Kazutaka Ohi[6], MD, PhD; Naoki Hashimoto[4], MD, PhD; Yuka Yasuda[7], MD, PhD; Takako Mitsudo[8], PhD; Fumihiro Higuchi[9], MD, PhD; Shusuke Numata[10], MD, PhD; Akiko Yamada[11], MD; Yohei Aoki[12], BA; Hiromitsu Honda[12], BA; Ryo Mizui[5], MD; Masato Honda[5], MD; Daisuke Fujikane[6], MD; Junya Matsumoto[3], MD, PhD; Naomi Hasegawa[3], PhD; Satsuki Ito[3], PhD; Hisashi Akiyama[4], MD; Toshiaki Onitsuka[13], MD, PhD; Yoshihiro Satomura[2,14], MD, PhD; Kiyoto Kasai[2], MD, PhD; Ryota Hashimoto[3], MD, PhD

[1]Department of Rehabilitation, The University of Tokyo Hospital, Bunkyo-ku Tokyo, Japan
[2]Department of Neuropsychiatry, Graduate School of Medicine, The University of Tokyo, Bunkyo-ku Tokyo, Japan
[3]Department of Pathology of Mental Diseases, National Institute of Mental Health, National Center of Neurology and Psychiatry, Kodaira, Japan
[4]Department of Psychiatry, Graduate School of Medicine, Hokkaido University, Sapporo, Japan
[5]Department of Psychiatry, Nara Medical University, Kashihara, Japan
[6]Department of Psychiatry, Graduate School of Medicine, Gifu University, Gifu, Japan
[7]Life Grow Brilliant Mental Clinic, Medical Corporation Foster, Kita-ku Osaka, Japan
[8]Division of Clinical Research, National Hospital Organization Hizen Psychiatric Center, Kanzaki-gun, Japan
[9]Department of Neuroscience, Division of Neuropsychiatry, Yamaguchi University School of Medicine, Ube City, Japan
[10]Department of Psychiatry, Graduate School of Biomedical Science, Tokushima University, Tokushima, Japan
[11]Department of Neuropsychiatry, Graduate School of Medicine, University of Kyoto, Sakyo-ku Kyoto, Japan
[12]Healthcare Innovation Group, Future Corporation, Shinagawa-ku Tokyo, Japan
[13]National Hospital Organization Sakakibara Hospital, Tsu, Japan
[14]Center for Diversity in Medical Education and Research, Graduate School of Medicine and Faculty of Medicine, The University of Tokyo, Bunkyo-ku Tokyo, Japan

Corresponding Author:
Ryota Hashimoto, MD, PhD
Department of Pathology of Mental Diseases
National Institute of Mental Health
National Center of Neurology and Psychiatry
4-1-1 Ogawa-Higashi
Kodaira, 187 8553
Japan
Phone: 81 42 346 2046
Email: ryotahashimoto55@ncnp.go.jp

Abstract

Background: Schizophrenia is a complex mental disorder characterized by significant cognitive and neurobiological alterations. Impairments in cognitive function and eye movement have been known to be promising biomarkers for schizophrenia. However, cognitive assessment methods require specialized expertise. To date, data on simplified measurement tools for assessing both cognitive function and eye movement in patients with schizophrenia are lacking.

Objective: This study aims to assess the efficacy of a novel tablet-based platform combining cognitive and eye movement measures for classifying schizophrenia.

Methods: Forty-four patients with schizophrenia, 67 healthy controls, and 41 patients with other psychiatric diagnoses participated in this study from 10 sites across Japan. A free-viewing eye movement task and 2 cognitive assessment tools (Codebreaker task from the THINC-integrated tool and the CognitiveFunctionTest app) were used for conducting assessments in a 12.9-inch iPad

Pro. We performed comparative group and logistic regression analyses for evaluating the diagnostic efficacy of the 3 measures of interest.

Results: Cognitive and eye movement measures differed significantly between patients with schizophrenia and healthy controls (all 3 measures; P<.001). The Codebreaker task showed the highest classification effectiveness in distinguishing schizophrenia with an area under the receiver operating characteristic curve of 0.90. Combining cognitive and eye movement measures further improved accuracy with a maximum area under the receiver operating characteristic curve of 0.94. Cognitive measures were more effective in differentiating patients with schizophrenia from healthy controls, whereas eye movement measures better differentiated schizophrenia from other psychiatric conditions.

Conclusions: This multisite study demonstrates the feasibility and effectiveness of a tablet-based app for assessing cognitive functioning and eye movements in patients with schizophrenia. Our results suggest the potential of tablet-based assessments of cognitive function and eye movement as simple and accessible evaluation tools, which may be useful for future clinical implementation.

(JMIR Ment Health 2024;11:e56668) doi:10.2196/56668

KEYWORDS

schizophrenia; cognitive function; eye movement; diagnostic biomarkers; digital health tools

Introduction

Schizophrenia is a severe mental illness that affects neurobiological processes, leading to difficulties in social and occupational functioning [1]. Years of research have revealed multiple candidate diagnostic markers for schizophrenia. Although some of these tests are promising, to date, none of them have been approved for use in diagnostic testing for schizophrenia [2]. This lack of approved diagnostic tests is caused by multiple factors such as the heterogeneity of mental illness, insufficient knowledge about the brain mechanisms and functions underlying mental states, and effects of known (and unknown) confounding factors [2]. Large-sample research can help overcome these issues [3]. This approach enables the observation of diverse symptoms and characteristics across a broad spectrum of individuals. However, as the sample size increases, maintaining good quality and feasibly performing assessments become difficult. Among the candidate markers for schizophrenia, cognitive function impairment and eye movement characteristics are key areas of interest. These features are not only easy to measure but also offer insights into the underlying pathophysiology of the psychiatric disorder [4].

Cognitive function impairments are observed in multiple domains among patients with schizophrenia. For the evaluation of specific cognitive dysfunctions in patients with schizophrenia, standardized test batteries have been developed by the Food and Drug Administration–National Institute of Mental Health–Measurement and Treatment Research to Improve Cognition in Schizophrenia (MATRICS), such as the MATRICS Consensus Cognitive Battery and Brief Assessment of Cognition in Schizophrenia [5,6]. Furthermore, efforts have been made to shorten and simplify these assessments [7,8]. Impairments have been replicated in aspects such as processing speed, verbal memory, working memory, attention, and executive functioning [9]. The relationships between these aspects of cognitive functioning impairment are complex. However, several network analyses have shown that processing speed has high centrality within neurocognitive functioning networks [10] and is associated with everyday life skills among community-dwelling individuals with schizophrenia [11]. The Digit Symbol Substitution Test is used to measure processing speed and is commonly included in cognitive functioning assessment batteries of patients with schizophrenia [12,13]. Due to the complexities and the multifaceted nature of cognitive impairments in patients with schizophrenia, establishing objective and reliable assessment methods is necessary.

Although cognitive function impairments are not specific to schizophrenia, there may be a greater need for objective measurements of cognitive function among patients with schizophrenia than there is among patients with other mental illnesses. In patients with schizophrenia, subjective perception of cognitive dysfunction is known; some reports suggest its association with quality of life and depression [14,15]. Although some studies indicate that subjective assessments and objective evaluations of cognitive dysfunction generally align [16,17], many papers point out discrepancies between the two [15,18-20]. Despite the emphasis on the importance of objective cognitive function assessments, these assessments have not been introduced into routine care for schizophrenia [20]. Tools such as the Digit Symbol Substitution Task are valuable assets and offer a standardized approach to quantifying cognitive deficits. Therefore, simple and objective means to measure cognitive function would be useful for the treatment of schizophrenia.

Eye movement characteristics in patients with schizophrenia have been widely studied [21]; these characteristics are among the most common behavioral deficits of this disorder. Various aspects of eye movement can be observed, such as smooth pursuit, saccade control, and visual search [21]. These analyses have been replicated in numerous samples [22-25]. In addition, these characteristics have been shown to be present from prodromal stages in patients with schizophrenia [26]. However, findings regarding the specificity of these characteristics are still inconsistent [23,27].

One example measure of eye movement is scanpath length, which is a measure of the total distance covered by the eye during visual exploration. In patients with schizophrenia, the scanpath length is generally shorter than it is in healthy controls, indicating more restricted visual exploration [28]. However, the reasons for shorter scanpath length in patients with schizophrenia are still being explored [29,30]. The simplicity of this assessment method and the need for few instructions during free-viewing tasks are beneficial for searching clinically feasible biomarkers.

Okazaki et al [31] explored the potential utility of combinations of cognitive function measures and eye movement measures in discriminating between patients with schizophrenia and healthy controls. They found that 7 pairs of cognitive functioning tests and eye movement measures, particularly pairs including digit-symbol coding or symbol search, demonstrated high discrimination performance. These pairs were acquirable in 10-15 minutes, suggesting that the combination of cognitive functioning measures and eye movement measures could be a simple and less time-consuming option for studying clinical patient groups such as patients with schizophrenia.

Based on these insights, our study aims to integrate the assessment of cognitive functioning with the analysis of eye movement characteristics. We utilized a novel tablet-based app to evaluate cognitive functioning and eye movement measurements among patients with schizophrenia. We aimed to explore whether a tablet-based app could replicate previous findings from desktop-based applications in terms of Digit Symbol Substitution Test scores and free-viewing scanpath lengths among patients with schizophrenia. Furthermore, we aimed to explore the diagnostic effectiveness of these measures in distinguishing patients with schizophrenia from other patient groups.

Methods

Participants

Forty-eight patients with schizophrenia, 69 healthy controls, and 49 patients with other psychiatric diagnoses were recruited from 10 clinical study sites across Japan. All study sites used the same study protocol and machines. The data were collected from distinct samples among which there were no overlapping participants. All participants were of Japanese descent and had normal or corrected-to-normal vision. Healthy controls were recruited independently at each study site through local advertisements or websites; they participated in structured interviews and were excluded from the study if a current mental illness diagnosis was suspected after the interview. Patients diagnosed with psychiatric disorders, including schizophrenia, were recruited from the outpatient and inpatient departments of each study site and included in the study after diagnosis by board-certified psychiatrists based on the Diagnostic and Statistical Manual of Mental Disorders, fifth edition criteria [32]. Distinguishing between schizophrenia and other specific psychiatric disorders is beyond the scope of this study. Therefore, we created a single group of patients with psychiatric disorders other than schizophrenia. All participants older than 12 years and who provided written consent were allowed to participate in this study (if the participant was younger than18 years, the participant provided assent, and a legal guardian provided consent before participation). The data were obtained between October and December 2022.

Two healthy controls, 4 patients with schizophrenia, and 8 patients with other psychiatric diagnoses were excluded from this study based on the predefined exclusion criteria, including deviations during measurement procedures (eg, distracting noises during the test procedure or wearing masks for infection prevention, which could disrupt the calibration process), equipment failure, or a medical history of ocular disease that may affect visual acuity (eg, glaucoma). The use of vision correction such as glasses or soft contact lenses was not a criterion for exclusion. A total of 44 patients with schizophrenia, 67 healthy controls, and 41 patients with other psychiatric diagnoses were ultimately included in the analysis. The dropout rates were 8% (4/48) among patients with schizophrenia, 3% (2/69) among healthy controls, and 16% (8/49) among patients with other psychiatric diagnoses.

The patients with other psychiatric diagnoses included 3 individuals with bipolar type 1 disorder, 5 with bipolar type 2 disorder, 10 with major depressive disorder, 19 with autism spectrum disorder, 1 with epilepsy, and 3 with comorbid diagnoses (1 individual with comorbid autism spectrum disorder and adjustment disorder, 1 with comorbid major depression and social anxiety disorder, and 1 with comorbid major depressive disorder and autism spectrum disorder).

Ethics Approval

Ethics approval was obtained after a central review process at the National Center of Neurology and Psychiatry (B2021-120) and after reviews by the ethics committees of all 10 study sites. This study was conducted in accordance with the Declaration of Helsinki. Basic demographic information, including age, sex, and years of education, was obtained from each participant in addition to the cognitive and eye movement measurements listed below.

Tablet Machine and Tasks

Novel tablet-based cognitive and eye movement assessments were conducted using a 12.9-inch iPad Pro (fifth generation) tablet. The display has a 2732×2048 pixel resolution at 264 ppi and a refresh rate of 120 Hz. Instructions were provided via voice or text within the app, and these instructions were given by a tester (a mental health professional) whenever necessary to ensure adequate understanding. A brief explanation about the 3 tasks used in this study is provided in Table 1.

Table 1. Brief descriptions about the tasks conducted.

Task name	Feature	Brief explanation	Approximate duration (min)	Measure
Codebreaker	Cognitive functioning impairment	In the Codebreaker task, participants match symbols to numbers from 1 to 6 using a key. They select matching symbols for a number sequence within 2 min to score points.	2	Codebreaker score: number of correct responses
Digit Symbol Substitution Test	Cognitive functioning impairment	In the Digit Symbol Substitution Test, participants match numbers with patterns on screen, aiming for the maximum correct pairs within 2 min.	2	Digit Symbol Substitution Test score: total score (1 point per correct)
Free-viewing test	Eye movement characteristic	In the free-viewing test, participants looked at 20 photos of buildings, items, and foods for 8 seconds each.	5	Scanpath length: mean scanpath length per image

Cognitive Measurements

Based on the results of a previous study [31], we focused on performing cognitive assessments homologous to the Digit Symbol Substitution Test. Regarding the assessment, we used the Codebreaker task from the Japanese version of the THINC-integrated tool (THINC-it) iPad/iOS version 1.261 (THINC Task Force), which is a computerized version of the Digit Symbol Substitution Test paradigm. The reliability and stability of the THINC-it have been validated [33,34]; it has also been used in patients with schizophrenia in previous studies [35].

During the Codebreaker task, the participants were given a legend that pairs numbers ranging between 1 and 6 with specific symbols at the top of the screen. The participants were then asked to correctly associate a series of symbols with their respective numbers based on this key. During the task, the participants were shown a sequence of numbers, and the corresponding symbols must be selected from a set at the bottom of the screen. The time limit was 2 minutes; correct matches related to higher scores. There were on-screen instructions before starting the task. The correct number of responses within 2 minutes was used as the Codebreaker score in the analysis.

We also used the Digit Symbol Substitution Test app CognitiveFunctionTest (version 1.0.3; Future Corporation). The participants were instructed to pair numbers with black and white patterns. The numbers were presented in the middle of the screen, and the participants selected the corresponding patterns from the bottom of the screen. The participants were instructed to correctly select as many pairs as possible within a time limit (2 minutes). During this task, instructions were given through audio-recorded guidance and a demo session. A score of 1 was assigned for every correct answer, and a score of 0 was assigned for incorrect answers. The total Digit Symbol Substitution Test score was then calculated and used for analyses. Considering both tasks, the participants were instructed to use only 1 hand and press the display with their fingers.

Eye Movement Measurements (iPad Pro)

The acquisition of eye movements and the calculation of eye movement measures were conducted using EyeMovementTest (version 1.0.2; Future Corporation). Eye movement measurements were performed using a front-facing camera (1200 megapixel resolution). The participants faced a tablet that was placed approximately 40 cm from their face. Eye movements were collected at a frequency of 60 Hz. A 9-point calibration and validation method was used to ensure accurate measurements of the data. The free-viewing test was conducted using 20 original photographic images of buildings, everyday items, and foods. The participants were instructed to freely view each image for 8 seconds. The images were randomly presented. A gray screen with a fixation point at the center was presented before the images, and a blank gray screen was displayed after each image. Based on previous studies [22,31], we calculated the average scanpath length across the 20 images and used this value for further analyses (scanpath length).

Statistical Analyses

The effects of demographic characteristics on Digit Symbol Substitution Test scores, Codebreaker scores, and scanpath length were first examined in healthy controls. Correlation analysis was used to examine age and years of education, and a 2-sided Mann-Whitney *U* test was used for sex. We calculated adjusted scores by using a linear model for measures that had confounding effects among participant demographics; these were used for the analysis.

Comparisons between 2 groups were performed using the Mann-Whitney *U* test. Comparisons between 3 groups were performed using the Kruskal-Wallis test with Dunn test for pairwise comparisons. *P* values for the post hoc tests were adjusted using the Benjamini-Hochberg false discovery rate procedure to adjust for multiple comparisons. Effect sizes for group comparisons were calculated using Cliff's δ and epsilon squared (denoted as ε^2).

The classification performance of the outcomes of interest was tested using a simple logistic regression model with leave-one-out cross validation. Receiver operating characteristic (ROC) analyses and confusion matrices were used to assess the performance of the classifier. The threshold value for each model was determined from the ROC curve, as this value could provide the best balance between sensitivity and specificity. Bootstrap resampling with 1000 iterations was used to calculate the 95% CIs for accuracy, specificity, and sensitivity. We also performed an exploratory analysis of the classification performance of the same model to distinguish patients with schizophrenia from those with other psychiatric diagnoses. The patients with schizophrenia were considered as patients in both classification analyses.

Data analyses were performed using R version 4.3.1 (R Foundation for Statistical Computing) implemented in R studio. The libraries *rstatix* (version 0.7.2) and *effsize* (version 0.8.1) were used for statistical analysis; the libraries *pROC* (version 1.18.5), *caret* (version 6.0.94), and *plyr* (version 1.8.9) were used for classification and model evaluation; and the library *ggplot2* (version 3.4.4) was used for visualization.

Results

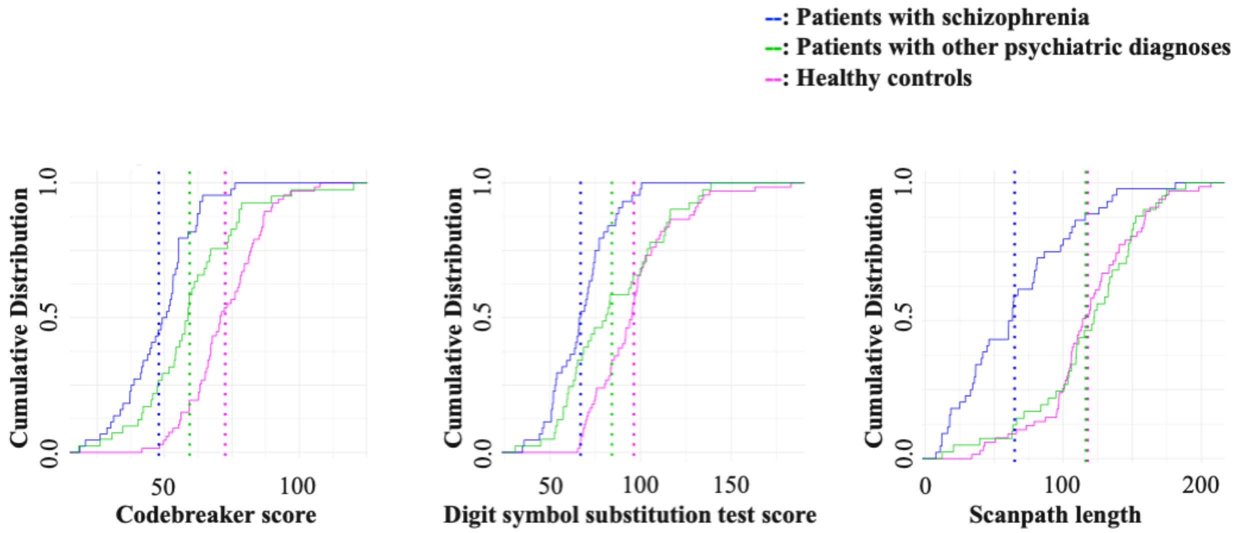
Exploratory Analysis of the Potential Confounders

We first explored the effects of demographic characteristics (age, years of education, and sex) on the 3 outcomes measured by the tablet app among healthy controls. The demographic characteristics of each group are provided in Table S1 in Multimedia Appendix 1. Correlation analysis revealed that Codebreaker scores and Digit Symbol Substitution Test scores were negatively correlated with age (ρ=–0.50; *P*<.001 and ρ=–0.34; *P*<.001, respectively) and positively correlated with years of education (ρ=0.32; *P*<.001 and ρ=0.24; *P*=.049, respectively). Free viewing was not significantly correlated with age (ρ=0.18; *P*=.15) or education years (ρ=0.17; *P*=.17). The Mann-Whitney *U* test revealed that there were no significant sex differences in Codebreaker scores (*U*=467; *P*=.35), Digit Symbol Substitution Test scores (*U*=565; *P*=.75), or scanpath length (*U*=652; *P*=.16) between the groups. Based on these results, we calculated age- and years of education–adjusted values for the Codebreaker scores and Digit Symbol Substitution Test scores. Thereafter, we used these values for the following analyses (unless otherwise specified as raw scores).

Distributions of the Measurement Values

We first examined the distribution of each measure in patients with schizophrenia, healthy controls, and patients with other psychiatric disorders. To visualize the distributions of each measure, we developed plots for the cumulative distribution function of each measure (Figure 1). The cumulative distribution function accumulated probabilities to indicate the likelihood of a value being at or below a certain point. Healthy controls and patients with schizophrenia differed significantly in all measures, with the Codebreaker scores having the largest difference (*P*<.001; Cliff's δ value –0.82, 95% CI –0.90 to –0.69). Notably, the distribution of scanpath length showed overlapping patterns between healthy controls and patients with other psychiatric disorders and did not differ significantly (*P*=.92; Cliff's δ value=–0.03, 95% CI –0.20 to 0.25). Group comparisons of each measure are provided in Table S2 in Multimedia Appendix 1.

Figure 1. Distribution of the measured values. The cumulative distribution function for the 3 measures of interest (ie, Codebreaker score, Digit Symbol Substitution Test score, and scanpath length) are provided across the 3 groups. The dotted lines in each plot represent the mean values in each group.



Classification Performance in Patients With Schizophrenia and in Healthy Controls

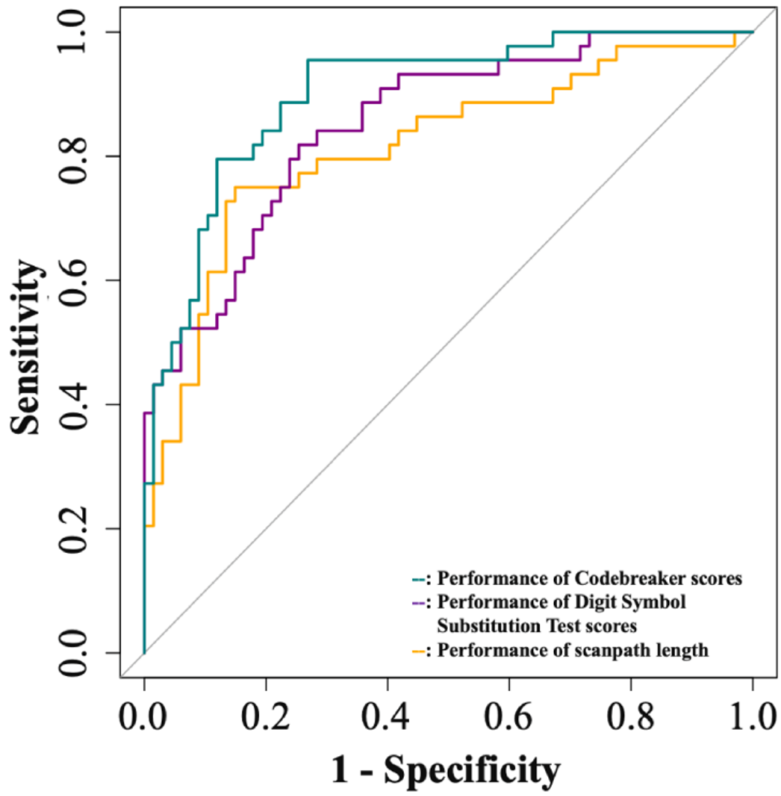
To test the diagnostic utility of the 3 measures, we classified the 2 diagnostic groups as dependent variables by using a logistic regression model and each of the 3 outcomes of interest as independent variables (Table 2). The area under the curve (AUC) was the largest for the Codebreaker score, with a value of 0.90 (95% CI 0.84-0.96). The accuracy (0.82, 95% CI 0.74-0.89) and specificity (0.95, 95% CI 0.89-1.00) were also the highest for the Codebreaker score, but the sensitivity was the highest for scanpath length (0.85, 95% CI 0.76-0.93). The ROC curves are shown in Figure 2. We also tested the classification performance of combining measures to identify the model with the best performance. The AUC was the largest for the model combining all 3 measures (AUC 0.94, 95% CI 0.90-0.98).

Table 2. Summary of the classification performance metrics in patients with schizophrenia versus healthy control participants.[a]

	Area under the curve (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Codebreaker score	0.90 (0.84-0.96)	0.82 (0.74-0.89)	0.73 (0.62-0.83)	0.95 (0.89-1.00)
Digit Symbol Substitution Test score	0.85 (0.78-0.92)	0.77 (0.69-0.85)	0.75 (0.64-0.85)	0.82 (0.70-0.92)
Scanpath length	0.81 (0.73-0.90)	0.81 (0.73-0.88)	0.85 (0.76-0.93)	0.75 (0.63-0.87)
Codebreaker score + Digit Symbol Substitution Test score	0.91 (0.86-0.96)	0.85 (0.77-0.91)	0.87 (0.78-0.94)	0.82 (0.70-0.93)
Codebreaker score + scan-path length	0.93 (0.88-0.97)	0.84 (0.76-0.90)	0.78 (0.68-0.87)	0.93 (0.85-1.00)
Digit Symbol Substitution Test score + scanpath length	0.90 (0.85-0.96)	0.85 (0.77-0.91)	0.94 (0.88-0.99)	0.70 (0.55-0.83)
Codebreaker score + Digit Symbol Substitution Test score + scanpath length	0.94 (0.90-0.98)	0.85 (0.77-0.91)	0.84 (0.75-0.92)	0.86 (0.75-0.95)

[a]Logistic regression models were used to classify patients with schizophrenia and healthy controls according to each measure of interest. We implemented leave-one-out cross-validation to estimate the generalizability of the trained model. We also conducted classifications by using pairs of eye movement and cognitive functioning measures or a combination of all 3 measures.

Figure 2. Receiver operating characteristic curves for each of the 3 outcomes of interest.



Classification Performance in Patients With Schizophrenia and in Patients With Other Psychiatric Diagnoses

The model made to classify patients with schizophrenia and healthy controls was also tested for classification performance in patients with schizophrenia and those with other psychiatric diagnoses (Table 3). This analysis revealed that scanpath length had the largest AUC (0.82, 95% CI 0.73-0.91), accompanied by notable accuracy (0.78, 95% CI 0.67-0.86) and sensitivity (0.83, 95% CI 0.71-0.93). Conversely, the Digit Symbol Substitution Test score showed the highest specificity (0.93, 95% CI 0.85-1.00), albeit with the lowest sensitivity (0.41, 95% CI 0.27-0.57). Table 3 also assesses the combined efficacy of eye movement and cognitive functioning measures in classification, which did not increase in this analysis.

Table 3. Summary of the classification performance metrics in patients with schizophrenia versus those with other psychiatric diagnoses. The models, which include individual measures and their combinations, were initially trained using data of healthy controls and patients with schizophrenia.

	Area under the curve (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Codebreaker score	0.70 (0.59-0.81)	0.71 (0.60-0.80)	0.61 (0.46-0.76)	0.80 (0.68-0.91)
Digit Symbol Substitution Test score	0.68 (0.57-0.80)	0.68 (0.57-0.78)	0.41 (0.27-0.57)	0.93 (0.85-1.00)
Scanpath length	0.82 (0.73-0.91)	0.78 (0.67-0.86)	0.83 (0.71-0.93)	0.73 (0.59-0.85)
Codebreaker score + Digit Symbol Substitution Test score	0.72 (0.61-0.83)	0.68 (0.57-0.78)	0.37 (0.23-0.52)	0.98 (0.92-1.00)
Codebreaker score + scanpath length	0.80 (0.70-0.89)	0.75 (0.65-0.84)	0.71 (0.57-0.84)	0.80 (0.67-0.90)
Digit Symbol Substitution Test score + scanpath length	0.81 (0.72-0.90)	0.76 (0.66-0.85)	0.88 (0.77-0.97)	0.66 (0.52-0.80)
Codebreaker score + Digit Symbol Substitution Test score + scanpath length	0.80 (0.71-0.89)	0.75 (0.65-0.84)	0.76 (0.62-0.88)	0.75 (0.62-0.87)

Due to the small sample size, the same patients with schizophrenia were used in both the training and testing phases for the previous analysis. Therefore, we also conducted leave-one-out cross validation for a model trained for all participants, classifying patients with schizophrenia versus participants without schizophrenia (healthy controls and patients with other diagnoses). Similar performance was seen for codebreaker score (AUC 0.82, 95% CI 0.75-0.89) and scanpath length (AUC 0.81, 95% CI 0.73-0.89). Combining all measures highly increased the performance (AUC 0.88, 95% CI 0.83-0.94). The complete results are presented in Table S3 in Multimedia Appendix 1.

Discussion

In this multisite study conducted in Japan, we assessed the results of tablet-based measurements of cognitive function and eye movement characteristics in patients with schizophrenia, healthy controls, and in patients with other psychiatric diagnoses. All 3 measures significantly differed among patients with schizophrenia, healthy controls, and those with other psychiatric diagnoses. We also demonstrated that the individual measures had high classification performances. The Codebreaker score was the most effective single measure in distinguishing patients with schizophrenia from healthy controls. Additionally, scanpath length was the most effective single measure in distinguishing patients with schizophrenia from those with other psychiatric diagnoses. Combining all 3 measures further improved the diagnostic performance. These findings suggest that cognitive and eye movement assessments performed using tablet-based platforms could have potential utility as a novel diagnostic approach.

Using a novel tablet-based measurement, we were able to replicate the findings of Digit Symbol Substitution Task deficits and shorter scanpath lengths in participants with schizophrenia compared with healthy controls. Cliff's δ effect sizes of the measures obtained using the tablet app were greater for the Codebreaker score (–0.82) and Digit Symbol Substitution Test score (–0.73) than that for the scanpath length (–0.66). A meta-analysis on processing speed deficits in patients with schizophrenia for Digit Symbol Substitution Test scores revealed effect sizes (Hedge *g*) ranging between –1.57 and –1.34 [36]. Regarding scanpath length, in our previous study [22], the effect size (Cohen *d*) was –1.5, and in a literature review by Beedie et al [28], the pooled effect size (Hedge *g*) calculated across 28 studies was –0.77 [28]. Therefore, the results of this study seemed to replicate the trend observed in previous studies.

A key finding of our study was the extent to which simple tablet-based assessments were able to match desktop-based tests in terms of distinguishing patients with schizophrenia from healthy controls. The AUC for scanpath length was 0.81, while the AUCs for cognitive function measures ranged between 0.85 and 0.90. These results were similar to the findings of previous studies using desktop-based eye movement measurements, which reported AUCs ranging between 0.77 and 0.89 [22,23,26], and previous studies using paper-based cognitive functioning measurements, which reported AUCs ranging between 0.88 and 0.90 [37,38]. Okazaki et al [31] found that pairing certain cognitive function measures with eye movement measures led to improved performance and robustness in distinguishing patients with schizophrenia from healthy individuals, with an average increase in the AUC of 0.10 compared with eye movement measures alone and 0.05 compared with cognitive function measures alone. Our study revealed similar results—combinations of the 3 measures of interest yielded AUC values ranging between 0.90 and 0.94. These results are promising, given the ease and simplicity of performing cognitive function and eye movement assessments using a tablet.

We further explored the differences between patients with schizophrenia and those with other psychiatric diagnoses. We found differences in both cognitive function measures and eye movement measures, with larger differences in eye movement measures. This finding is in line with that in previous studies that established diagnostic markers for differentiating between patients with schizophrenia and those without schizophrenia by using eye movement characteristics [39,40]. However, previous studies indicated mixed results of both cognitive function [41-43] and eye movement [23,40,44]. Conducting future studies on the transdiagnostic characteristics of multiple modalities is increasingly important. Moreover, mobile measurement will help this process by making studies possible by using less expensive and more accessible formats.

The usage of mobile formats for measuring cognitive function has already been reported in the field of psychiatry and elsewhere. Several studies have assessed the validity of mobile (tablet-based or smartphone-based) apps compared with paper-based studies and have shown equivalent performance [8,45-48]. One study utilized the simplicity of mobile apps to conduct a longitudinal study in which cognitive assessments were conducted outside the experimental setting and had higher completion rates in patients with schizophrenia than in the control group [49]. However, only a few case-control comparisons have been performed [49,50]. Additionally, previous studies included small sample sizes, and there is a need for studying the specific effects of using technological tools, such as accessibility and digital literacy [51].

Eye movement characteristics have not been studied using mobile devices in patients with schizophrenia. Studies [52-55] have used portable eye trackers. An old but innovative study focused on the reliability of a portable head-mounted display for tracking eye movements in patients with schizophrenia and found that it was comparable to traditional fixed-display setups [52]. Dowiasch et al [53] examined the eye movements of patients with schizophrenia in a natural, everyday setting, contrasting them with healthy controls. They identified distinct eye movement behaviors such as more frequent but shorter fixations in patients with schizophrenia. Other studies used portable devices but also required participants to fix their head during the tasks [24,54,55]. These tools are still expensive, and mobile tools such as ours are yet to be utilized in the field of psychiatry.

To the best of our knowledge, there are no studies of tablet-based or simplified measurement tools that focus on both cognitive function and eye movement assessments in patients with schizophrenia. With our tablet-based measurement tool, we were able to observe different classification features of cognitive functioning measures (Codebreaker scores and Digit Symbol Substitution Test scores) and for the eye movement measure (scanpath length). Cognitive functioning measures were more useful for differentiating patients with schizophrenia from healthy controls, while scanpath length differed more between patients with schizophrenia and those with other psychiatric diagnoses. By combining the 2 measures, we were also able to increase diagnostic performance. This is interesting, given that both cognitive function and eye movement have long been known to be affected in patients with schizophrenia [56,57]. One possibility is that differences in neurobiological substrates lead to digit symbol substitution impairments and abnormal visual searches [58,59]. Psychiatric disorders exhibit both within-disorder heterogeneity and transdiagnostic features, and recent large-scale neuroimaging studies have shown disease-specific and shared neuroanatomical alterations [60,61]. Combining different features of schizophrenia, such as combining cognitive function and eye movement measures, may be beneficial not only for disease classification but also for biotyping, which may lead to developing new treatment options in the future.

The use of tablet-based assessment tools such as the one used in this study may facilitate development and implementation of auxiliary diagnostic tools in the field of psychiatry. Conventional cognitive assessment batteries are time-consuming and require special training for implementation, and conventional eye movement measurement requires the use of expensive and complex equipment, necessitating expert setup and skilled operation for measurements. Some studies have found ways to conduct these conventional assessments in a less time-consuming way [31,37]. However, our study is novel in that it combines simplified measurements of cognitive function and eye movement, which do not require such expertise. The tablet used in our study might allow for an inexpensive, more convenient, and compact form of evaluation, combining psychological and neurophysiological findings of schizophrenia. Tablet-based tools will make such assessments more accessible and convenient for patients and clinicians alike and may lead to having more objective information to aid decision-making in the clinical setting.

This study has several limitations. One limitation is the small sample size of participants included in this study, and our research would require replication considering larger samples. Another is that 2 of the measures used in this study (the Digit Symbol Substitution Test score and scanpath length) have still to be tested for reliability and validity like the Codebreaker task [33,62]. The effects of possible confounders were also not assessed in detail. Although we employed age- and years of education–adjusted scores to account for demographic differences between the groups, this was a linear correction due to the small sample size. Further analyses of age- and years of education–matched samples may be needed. There also remains the possibility of other confounding factors such as medication usage that has not been studied or fully controlled. Several studies have shown that shorter scanpath lengths [28] and slower processing speeds [63] are present before medication usage. Additionally, the results of cognitive functioning measures, including the Codebreaker test, have been known to change longitudinally [35,64,65]. In our study, we did not assess the longitudinal robustness of our findings. Studies with larger sample sizes in various age groups and reliability and validity testing of the measures as well as addressing the limitations above would enhance the robustness and generalizability of our findings.

In this study, we successfully demonstrated the feasibility of using novel tablet-based apps to perform cognitive function and eye movement assessments among patients with schizophrenia. Our findings align with those of previous research indicating significant differences in cognitive and eye movement measures between patients with schizophrenia and healthy controls, with the key feature of being able to accomplish this using tablet-based assessments. In particular, the measurement of eye movement by using a tablet device is unique to this study. These findings suggest that assessments using such digital tools may hold the potential for use in various clinical settings, such as an auxiliary diagnostic tool to aid the decision-making process in psychiatric treatment in the future.

Acknowledgments

We would like to thank the study participants for their important contributions. This research was supported by Japan Agency for Medical Research and Development grants JP21uk1024002 (K Morita, K Miura, MM, KO, N Hashimoto, YY, TO, SN, HH, JM, and RH), JP21dk0307103 (RH), JP18dm0307002 (RH), and JP21wm0425012 (K Miura, JM, and RH); Japan Society for the Promotion of Science KAKENHI grants JP20K16617 (K Morita), JP20H03611 (RH, K Miura, JM, and N Hasegawa), JP20K06920 (K Miura, JM, and N Hasegawa), and JP19H05467 (RH); and Intramural Research grant (3-1, 4-6) for Neurological and Psychiatric Disorders of National Center of Neurology and Psychiatry (RH and K Miura).

Authors' Contributions

K Morita, K Miura, and RH designed this study. K Morita, K Miura, AT, MM, KO, N Hashimoto, YY, TO, FH, SN, AY, YA, HH, RM, MH, DF, JM, N Hasegawa, SI, HA, TM, YS, and RH contributed to data collection. K Morita, K Miura, YA, HH, and RH arranged and analyzed the data. K Morita, K Miura, AT, MM, KO, N Hashimoto, YY, TO, FH, SN, AY, YA, HH, RM, MH, DF, JM, HA, KK, and RH interpreted the results. K Morita, K Miura, and RH wrote the manuscript. K Miura and RH supervised the entire project. All the authors have reviewed and approved the final manuscript.

Conflicts of Interest

YA and HH are employees of Future Corporation. All other authors declare no financial or nonfinancial competing interests.

Multimedia Appendix 1
Supplementary tables.
[DOCX File , 54 KB - mental_v11i1e56668_app1.docx]

References

1. Jauhar S, Johnstone M, McKenna PJ. Schizophrenia. Lancet 2022 Jan 29;399(10323):473-486. [doi: 10.1016/S0140-6736(21)01730-X] [Medline: 35093231]
2. Abi-Dargham A, Moeller SJ, Ali F, DeLorenzo C, Domschke K, Horga G, et al. Candidate biomarkers in psychiatric disorders: state of the field. World Psychiatry 2023 Jun;22(2):236-262 [FREE Full text] [doi: 10.1002/wps.21078] [Medline: 37159365]
3. Onitsuka T, Hirano Y, Nemoto K, Hashimoto N, Kushima I, Koshiyama D, et al. Trends in big data analyses by multicenter collaborative translational research in psychiatry. Psychiatry Clin Neurosci 2022 Jan;76(1):1-14 [FREE Full text] [doi: 10.1111/pcn.13311] [Medline: 34716732]
4. Onitsuka T, Hirano Y, Nakazawa T, Ichihashi K, Miura K, Inada K, et al. Toward recovery in schizophrenia: Current concepts, findings, and future research directions. Psychiatry Clin Neurosci 2022 Jul;76(7):282-291 [FREE Full text] [doi: 10.1111/pcn.13342] [Medline: 35235256]
5. Kern RS, Green MF, Nuechterlein KH, Deng B. NIMH-MATRICS survey on assessment of neurocognition in schizophrenia. Schizophr Res 2004 Dec 15;72(1):11-19. [doi: 10.1016/j.schres.2004.09.004] [Medline: 15531403]
6. Keefe RSE, Harvey PD, Goldberg TE, Gold JM, Walker TM, Kennel C, et al. Norms and standardization of the Brief Assessment of Cognition in Schizophrenia (BACS). Schizophr Res 2008 Jul;102(1-3):108-115. [doi: 10.1016/j.schres.2008.03.024] [Medline: 18495435]
7. Lam M, Wang M, Huang W, Eng GK, Rapisarda A, Kraus M, et al. Establishing the brief assessment of cognition - short form. J Psychiatr Res 2017 Oct;93:1-11 [FREE Full text] [doi: 10.1016/j.jpsychires.2017.05.006] [Medline: 28549241]
8. Atkins AS, Tseng T, Vaughan A, Twamley EW, Harvey P, Patterson T, et al. Validation of the tablet-administered Brief Assessment of Cognition (BAC App). Schizophr Res 2017 Mar;181:100-106 [FREE Full text] [doi: 10.1016/j.schres.2016.10.010] [Medline: 27771201]
9. Gebreegziabhere Y, Habatmu K, Mihretu A, Cella M, Alem A. Cognitive impairment in people with schizophrenia: an umbrella review. Eur Arch Psychiatry Clin Neurosci 2022 Oct;272(7):1139-1155 [FREE Full text] [doi: 10.1007/s00406-022-01416-6] [Medline: 35633394]
10. Karyakina M, Shmukler A. Network analysis of cognitive deficit in patients with schizophrenia spectrum disorders. Schizophr Res Cogn 2021 Dec;26:100213 [FREE Full text] [doi: 10.1016/j.scog.2021.100213] [Medline: 34466392]
11. Galderisi S, Rucci P, Mucci A, Rossi A, Rocca P, Bertolino A, Italian Network for Research on Psychoses. The interplay among psychopathology, personal resources, context-related factors and real-life functioning in schizophrenia: stability in relationships after 4 years and differences in network structure between recovered and non-recovered patients. World Psychiatry 2020 Feb;19(1):81-91 [FREE Full text] [doi: 10.1002/wps.20700] [Medline: 31922687]
12. Keefe RSE, Goldberg TE, Harvey PD, Gold JM, Poe MP, Coughenour L. The brief assessment of cognition in schizophrenia: reliability, sensitivity, and comparison with a standard neurocognitive battery. Schizophr Res 2004 Jun 01;68(2-3):283-297. [doi: 10.1016/j.schres.2003.09.011] [Medline: 15099610]

13. Nuechterlein KH, Green MF, Kern RS, Baade LE, Barch DM, Cohen JD, et al. The MATRICS Consensus Cognitive Battery, part 1: test selection, reliability, and validity. Am J Psychiatry 2008 Feb;165(2):203-213. [doi: 10.1176/appi.ajp.2007.07010042] [Medline: 18172019]
14. Paudel S, Coman D, Freudenreich O. Subjective experience of cognitive difficulties as an important attribute of quality of life among individuals with schizophrenia spectrum disorders. Schizophr Res 2020 Jan;215:476-478. [doi: 10.1016/j.schres.2019.09.008] [Medline: 31672382]
15. Sellwood W, Morrison AP, Beck R, Heffernan S, Law H, Bentall RP. Subjective cognitive complaints in schizophrenia: relation to antipsychotic medication dose, actual cognitive performance, insight and symptoms. PLoS One 2013;8(12):e83774 [FREE Full text] [doi: 10.1371/journal.pone.0083774] [Medline: 24376745]
16. Lecardeur L, Briand C, Prouteau A, Lalonde P, Nicole L, Lesage A, et al. Preserved awareness of their cognitive deficits in patients with schizophrenia: convergent validity of the SSTICS. Schizophr Res 2009 Feb;107(2-3):303-306. [doi: 10.1016/j.schres.2008.09.003] [Medline: 18835134]
17. Gilleen J, Greenwood K, David AS. Domains of awareness in schizophrenia. Schizophr Bull 2011 Jan;37(1):61-72 [FREE Full text] [doi: 10.1093/schbul/sbq100] [Medline: 20851850]
18. Medalia A, Thysen J. A comparison of insight into clinical symptoms versus insight into neuro-cognitive symptoms in schizophrenia. Schizophr Res 2010 May;118(1-3):134-139. [doi: 10.1016/j.schres.2009.09.027] [Medline: 19840898]
19. Burton CZ, Harvey PD, Patterson TL, Twamley EW. Neurocognitive insight and objective cognitive functioning in schizophrenia. Schizophr Res 2016 Mar;171(1-3):131-136 [FREE Full text] [doi: 10.1016/j.schres.2016.01.021] [Medline: 26811232]
20. McCutcheon RA, Keefe RSE, McGuire PK. Cognitive impairment in schizophrenia: aetiology, pathophysiology, and treatment. Mol Psychiatry 2023 May;28(5):1902-1918 [FREE Full text] [doi: 10.1038/s41380-023-01949-9] [Medline: 36690793]
21. Morita K, Miura K, Kasai K, Hashimoto R. Eye movement characteristics in schizophrenia: A recent update with clinical implications. Neuropsychopharmacol Rep 2020 Mar;40(1):2-9 [FREE Full text] [doi: 10.1002/npr2.12087] [Medline: 31774633]
22. Morita K, Miura K, Fujimoto M, Yamamori H, Yasuda Y, Iwase M, et al. Eye movement as a biomarker of schizophrenia: Using an integrated eye movement score. Psychiatry Clin Neurosci 2017 Feb;71(2):104-114 [FREE Full text] [doi: 10.1111/pcn.12460] [Medline: 27673731]
23. St Clair D, MacLennan G, Beedie SA, Nouzová E, Lemmon H, Rujescu D, et al. Eye movement patterns can distinguish schizophrenia from the major affective disorders and healthy control subjects. Schizophr Bull Open 2022 Jan;3(1):sgac032 [FREE Full text] [doi: 10.1093/schizbullopen/sgac032] [Medline: 35669867]
24. Lencer R, Sprenger A, Reilly JL, McDowell JE, Rubin LH, Badner JA, et al. Pursuit eye movements as an intermediate phenotype across psychotic disorders: Evidence from the B-SNIP study. Schizophr Res 2015 Dec;169(1-3):326-333 [FREE Full text] [doi: 10.1016/j.schres.2015.09.032] [Medline: 26481615]
25. Radant AD, Millard SP, Braff DL, Calkins ME, Dobie DJ, Freedman R, et al. Robust differences in antisaccade performance exist between COGS schizophrenia cases and controls regardless of recruitment strategies. Schizophr Res 2015 Apr;163(1-3):47-52 [FREE Full text] [doi: 10.1016/j.schres.2014.12.016] [Medline: 25553977]
26. Lyu H, St CD, Wu R, Benson P, Guo W, Wang G, et al. Eye movement abnormalities can distinguish first-episode schizophrenia, chronic schizophrenia, and prodromal patients from healthy controls. Schizophr Bull Open Oxford Academic 2023 Jan 3;4(1):sgac076. [doi: 10.1093/schizbullopen/sgac076]
27. Takahashi J, Hirano Y, Miura K, Morita K, Fujimoto M, Yamamori H, et al. Eye movement abnormalities in major depressive disorder. Front Psychiatry 2021;12:673443 [FREE Full text] [doi: 10.3389/fpsyt.2021.673443] [Medline: 34447321]
28. Beedie SA, Benson PJ, St Clair DM. Atypical scanpaths in schizophrenia: evidence of a trait- or state-dependent phenomenon? J Psychiatry Neurosci 2011 May;36(3):150-164 [FREE Full text] [doi: 10.1503/jpn.090169] [Medline: 21223647]
29. Okada K, Miura K, Fujimoto M, Morita K, Yoshida M, Yamamori H, et al. Impaired inhibition of return during free-viewing behaviour in patients with schizophrenia. Sci Rep 2021 Feb 05;11(1):3237 [FREE Full text] [doi: 10.1038/s41598-021-82253-w] [Medline: 33547381]
30. Yoshida M, Miura K, Fujimoto M, Yamamori H, Yasuda Y, Iwase M, et al. Visual salience is affected in participants with schizophrenia during free-viewing. Sci Rep 2024 Feb 26;14(1):4606 [FREE Full text] [doi: 10.1038/s41598-024-55359-0] [Medline: 38409435]
31. Okazaki K, Miura K, Matsumoto J, Hasegawa N, Fujimoto M, Yamamori H, et al. Discrimination in the clinical diagnosis between patients with schizophrenia and healthy controls using eye movement and cognitive functions. Psychiatry Clin Neurosci 2023 Jul;77(7):393-400. [doi: 10.1111/pcn.13553] [Medline: 37029644]
32. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition. Washington, DC: American Psychiatric Publishing; 2013.
33. Harrison JE, Barry H, Baune BT, Best MW, Bowie CR, Cha DS, et al. Stability, reliability, and validity of the THINC-it screening tool for cognitive impairment in depression: A psychometric exploration in healthy volunteers. Int J Methods Psychiatr Res 2018 Sep;27(3):e1736 [FREE Full text] [doi: 10.1002/mpr.1736] [Medline: 30088298]

34. McIntyre RS, Best MW, Bowie CR, Carmona NE, Cha DS, Lee Y, et al. The THINC-Integrated Tool (THINC-it) screening assessment for cognitive dysfunction: validation in patients with major depressive disorder. J Clin Psychiatry 2017 Jul;78(7):873-881. [doi: 10.4088/JCP.16m11329] [Medline: 28858441]
35. Szmyd JK, Lewczuk K, Teopiz KM, McIntyre RS, Wichniak A. THINC-Integrated Tool (THINC-it): a brief measurement of changes in cognitive functioning and its correlation with the life quality of patients with schizophrenia and related disorders-a pilot study. Brain Sci 2023 Feb 24;13(3):389 [FREE Full text] [doi: 10.3390/brainsci13030389] [Medline: 36979199]
36. Knowles EEM, David AS, Reichenberg A. Processing speed deficits in schizophrenia: reexamining the evidence. Am J Psychiatry 2010 Jul;167(7):828-835. [doi: 10.1176/appi.ajp.2010.09070937] [Medline: 20439390]
37. Fujino H, Sumiyoshi C, Yasuda Y, Yamamori H, Fujimoto M, Fukunaga M, et al. Estimated cognitive decline in patients with schizophrenia: A multicenter study. Psychiatry Clin Neurosci 2017 May;71(5):294-300 [FREE Full text] [doi: 10.1111/pcn.12474] [Medline: 27804186]
38. Bezdicek O, Michalec J, Kališová L, Kufa T, Děchtěrenko F, Chlebovcová M, et al. Profile of cognitive deficits in schizophrenia and factor structure of the Czech MATRICS Consensus Cognitive Battery. Schizophr Res 2020 Apr;218:85-92. [doi: 10.1016/j.schres.2020.02.004] [Medline: 32089475]
39. Kojima T, Matsushima E, Ohta K, Toru M, Han YH, Shen YC, et al. Stability of exploratory eye movements as a marker of schizophrenia--a WHO multi-center study. World Health Organization. Schizophr Res 2001 Dec 01;52(3):203-213. [doi: 10.1016/s0920-9964(00)00181-x] [Medline: 11705714]
40. Suzuki M, Takahashi S, Matsushima E, Tsunoda M, Kurachi M, Okada T, et al. Exploratory eye movement dysfunction as a discriminator for schizophrenia : a large sample study using a newly developed digital computerized system. Eur Arch Psychiatry Clin Neurosci 2009 Apr;259(3):186-194. [doi: 10.1007/s00406-008-0850-7] [Medline: 19165524]
41. Goldberg TE, Gold JM, Greenberg R, Griffin S, Schulz SC, Pickar D, et al. Contrasts between patients with affective disorders and patients with schizophrenia on a neuropsychological test battery. Am J Psychiatry 1993 Sep;150(9):1355-1362. [doi: 10.1176/ajp.150.9.1355] [Medline: 8352346]
42. Lewandowski KE, Sperry SH, Cohen BM, Ongür D. Cognitive variability in psychotic disorders: a cross-diagnostic cluster analysis. Psychol Med 2014 Nov;44(15):3239-3248 [FREE Full text] [doi: 10.1017/S0033291714000774] [Medline: 25066202]
43. Sumiyoshi C, Ohi K, Fujino H, Yamamori H, Fujimoto M, Yasuda Y, et al. Transdiagnostic comparisons of intellectual abilities and work outcome in patients with mental disorders: multicentre study. BJPsych Open 2022 Jun 03;8(4):e98 [FREE Full text] [doi: 10.1192/bjo.2022.50] [Medline: 35656577]
44. Bestelmeyer PEG, Tatler BW, Phillips LH, Fraser G, Benson PJ, St Clair D. Global visual scanning abnormalities in schizophrenia and bipolar disorder. Schizophr Res 2006 Oct;87(1-3):212-222. [doi: 10.1016/j.schres.2006.06.015] [Medline: 16860975]
45. Pratt DN, Luther L, Kinney KS, Osborne KJ, Corlett PR, Powers AR, et al. Comparing a computerized digit symbol test to a pen-and-paper classic. Schizophr Bull Open 2023 Jan;4(1):sgad027 [FREE Full text] [doi: 10.1093/schizbullopen/sgad027] [Medline: 37868160]
46. Tang S, Chen I, Chiang H, Wu C, Hsueh I, Yu W, et al. A comparison between the original and Tablet-based Symbol Digit Modalities Test in patients with schizophrenia: Test-retest agreement, random measurement error, practice effect, and ecological validity. Psychiatry Res 2018 Feb;260:199-206. [doi: 10.1016/j.psychres.2017.11.066] [Medline: 29202384]
47. Cubillos C, Rienzo A. Digital cognitive assessment tests for older adults: systematic literature review. JMIR Ment Health 2023 Dec 08;10:e47487 [FREE Full text] [doi: 10.2196/47487] [Medline: 38064247]
48. McIntyre RS, Lipsitz O, Rodrigues NB, Subramaniapillai M, Nasri F, Lee Y, et al. An app-based digit symbol substitution test for assessment of cognitive deficits in adults with major depressive disorder: evaluation study. JMIR Ment Health 2022 Oct 27;9(10):e33871 [FREE Full text] [doi: 10.2196/33871] [Medline: 36301615]
49. Liu G, Henson P, Keshavan M, Pekka-Onnela J, Torous J. Assessing the potential of longitudinal smartphone based cognitive assessment in schizophrenia: A naturalistic pilot study. Schizophr Res Cogn 2019 Sep;17:100144 [FREE Full text] [doi: 10.1016/j.scog.2019.100144] [Medline: 31024801]
50. Cornelis C, De Picker LJ, Hulstijn W, Dumont G, Timmers M, Janssens L, et al. Preserved learning during the symbol-digit substitution test in patients with schizophrenia, age-matched controls, and elderly. Front Psychiatry 2014;5:189 [FREE Full text] [doi: 10.3389/fpsyt.2014.00189] [Medline: 25610403]
51. Lavigne KM, Sauvé G, Raucher-Chéné D, Guimond S, Lecomte T, Bowie CR, et al. Remote cognitive assessment in severe mental illness: a scoping review. Schizophrenia (Heidelb) 2022 Mar 05;8(1):14 [FREE Full text] [doi: 10.1038/s41537-022-00219-x] [Medline: 35249112]
52. Hong LE, Avila MT, Wonodi I, McMahon RP, Thaker GK. Reliability of a portable head-mounted eye tracking instrument for schizophrenia research. Behav Res Methods 2005 Feb;37(1):133-138. [doi: 10.3758/bf03206407] [Medline: 16097353]
53. Dowiasch S, Backasch B, Einhäuser W, Leube D, Kircher T, Bremmer F. Eye movements of patients with schizophrenia in a natural environment. Eur Arch Psychiatry Clin Neurosci 2016 Feb;266(1):43-54 [FREE Full text] [doi: 10.1007/s00406-014-0567-8] [Medline: 25472882]

54. Delerue C, Boucart M. Visual exploration and action processing in schizophrenia. Cogn Neuropsychiatry 2013 May;18(3):153-167. [doi: 10.1080/13546805.2012.683245] [Medline: 22838400]
55. Obyedkov I, Skuhareuskaya M, Skugarevsky O, Obyedkov V, Buslauski P, Skuhareuskaya T, et al. Saccadic eye movements in different dimensions of schizophrenia and in clinical high-risk state for psychosis. BMC Psychiatry 2019 Apr 08;19(1):110 [FREE Full text] [doi: 10.1186/s12888-019-2093-8] [Medline: 30961571]
56. Dickinson D, Ramsey ME, Gold JM. Overlooking the obvious: a meta-analytic comparison of digit symbol coding tasks and other cognitive measures in schizophrenia. Arch Gen Psychiatry 2007 May;64(5):532-542. [doi: 10.1001/archpsyc.64.5.532] [Medline: 17485605]
57. Phillips ML, David AS. Abnormal visual scan paths: a psychophysiological marker of delusions in schizophrenia. Schizophr Res 1998 Feb 09;29(3):235-245. [doi: 10.1016/s0920-9964(97)00097-2] [Medline: 9516664]
58. Koshiyama D, Fukunaga M, Okada N, Yamashita F, Yamamori H, Yasuda Y, et al. Role of subcortical structures on cognitive and social function in schizophrenia. Sci Rep 2018 Jan 19;8(1):1183 [FREE Full text] [doi: 10.1038/s41598-017-18950-2] [Medline: 29352126]
59. Morita K, Miura K, Fujimoto M, Yamamori H, Yasuda Y, Kudo N, et al. Eye-movement characteristics of schizophrenia and their association with cortical thickness. Psychiatry Clin Neurosci 2019 Aug;73(8):508-509 [FREE Full text] [doi: 10.1111/pcn.12865] [Medline: 31102322]
60. Matsumoto J, Fukunaga M, Miura K, Nemoto K, Okada N, Hashimoto N, et al. Cerebral cortical structural alteration patterns across four major psychiatric disorders in 5549 individuals. Mol Psychiatry 2023 Nov;28(11):4915-4923 [FREE Full text] [doi: 10.1038/s41380-023-02224-7] [Medline: 37596354]
61. Okada N, Fukunaga M, Miura K, Nemoto K, Matsumoto J, Hashimoto N, et al. Subcortical volumetric alterations in four major psychiatric disorders: a mega-analysis study of 5604 subjects and a volumetric data-driven approach for classification. Mol Psychiatry 2023 Dec;28(12):5206-5216. [doi: 10.1038/s41380-023-02141-9] [Medline: 37537281]
62. Dalby M, Annas P, 23andMe Research Team, Harrison JE. Further validation of the THINC-it tool and extension of the normative data set in a study of n = 10.019 typical controls. Int J Methods Psychiatr Res 2022 Dec;31(4):e1922 [FREE Full text] [doi: 10.1002/mpr.1922] [Medline: 35748111]
63. Peng X, Hei G, Yang Y, Liu C, Xiao J, Long Y, et al. The association between cognitive deficits and clinical characteristic in first-episode drug naïve patients with schizophrenia. Front Psychiatry 2021;12:638773 [FREE Full text] [doi: 10.3389/fpsyt.2021.638773] [Medline: 33716832]
64. Lennertz L, An der Heiden W, Kronacher R, Schulze-Rauschenbach S, Maier W, Häfner H, et al. Smaller than expected cognitive deficits in schizophrenia patients from the population-representative ABC catchment cohort. Eur Arch Psychiatry Clin Neurosci 2016 Aug;266(5):423-431. [doi: 10.1007/s00406-015-0625-x] [Medline: 26233432]
65. McIntyre RS, Subramaniapillai M, Park C, Zuckerman H, Cao B, Lee Y, et al. The THINC-it Tool for cognitive assessment and measurement in major depressive disorder: sensitivity to change. Front Psychiatry 2020;11:546 [FREE Full text] [doi: 10.3389/fpsyt.2020.00546] [Medline: 32670103]

Abbreviations

AUC: area under the curve
MATRICS: Measurement and Treatment Research to Improve Cognition in Schizophrenia
ROC: receiver operating characteristic
THINC-it: THINC-integrated tool

Edited by J Torous; submitted 25.01.24; peer-reviewed by A Lynham, S Eack; comments to author 05.03.24; revised version received 10.04.24; accepted 01.05.24; published 30.05.24.

Please cite as:
Morita K, Miura K, Toyomaki A, Makinodan M, Ohi K, Hashimoto N, Yasuda Y, Mitsudo T, Higuchi F, Numata S, Yamada A, Aoki Y, Honda H, Mizui R, Honda M, Fujikane D, Matsumoto J, Hasegawa N, Ito S, Akiyama H, Onitsuka T, Satomura Y, Kasai K, Hashimoto R
Tablet-Based Cognitive and Eye Movement Measures as Accessible Tools for Schizophrenia Assessment: Multisite Usability Study
JMIR Ment Health 2024;11:e56668
URL: https://mental.jmir.org/2024/1/e56668
doi:10.2196/56668
PMID:38815257

The Efficacy of Web-Based Cognitive Behavioral Therapy With a Shame-Specific Intervention for Social Anxiety Disorder: Randomized Controlled Trial

Xu Wen[1,2], MS; Mengke Gou[2], PhD; Huijing Chen[3], PhD; Tomoko Kishimoto[4], PhD; Mingyi Qian[2], Prof Dr; Jürgen Margraf[1], Prof Dr; Thomas Berger[5], Prof Dr

1

2

3

4

5

Corresponding Author:
Mingyi Qian, Prof Dr

Abstract

Background: Social anxiety disorder (SAD) is one of the most prevalent psychological disorders and generally co-occurs with elevated shame levels. Previous shame-specific interventions could significantly improve outcomes in social anxiety treatments. Recent review suggests that integrating a more direct shame intervention could potentially increase the effectiveness of cognitive behavioral therapy. Web-based cognitive behavioral therapy (WCBT) has proven efficacy, sustaining benefits for 6 months to 4 years. Previous evidence indicated that shame predicted the reduction of social anxiety and mediated between engagements in exposure and changes in social anxiety during WCBT.

Objective: This study aimed to design a shame intervention component through a longitudinal study and conduct a randomized controlled trial to investigate the effectiveness of a shame intervention component in reducing social anxiety symptoms and shame experience in a clinical sample of people with SAD.

Methods: The development of a shame intervention component was informed by cognitive behavioral principles and insights from longitudinal data that measured the Experience of Shame Scale (ESS), the Coping Styles Questionnaire, and the Social Interaction Anxiety Scale (SIAS) in 153 participants. The psychoeducation, cognitive construct, and exposure sections were tailored to focus more on shame-related problem-solving and self-blame. A total of 1220 participants were recruited to complete questionnaires, including the ESS, the SIAS, the Social Phobia Scale (SPS), and diagnostic interviews. Following a 2-round screening process, 201 participants with SAD were randomly assigned into a shame WCBT group, a normal WCBT group, and a waiting group. After the 8-week WCBT intervention, the participants were asked to complete posttest evaluations, including the ESS, SIAS and SPS.

Results: Participants in the shame WCBT group experienced significant reductions in shame levels after the intervention (ESS: P<.001; ηp^2=0.22), and the reduction was greater in the shame intervention group compared to normal WCBT (P<.001; mean deviation −12.50). Participants in both the shame WCBT and normal WCBT groups experienced significant reductions in social anxiety symptoms (SIAS: P<.001; ηp^2=0.32; SPS: P<.001; ηp^2=0.19) compared to the waiting group after intervention. Furthermore, in the experience of social interaction anxiety (SIAS), the shame WCBT group showed a higher reduction compared to the normal WCBT group (P<.001; mean deviation −9.58). Problem-solving (SE 0.049, 95% CI 0.025-0.217) and self-blame (SE 0.082, 95% CI 0.024-0.339) mediated the effect between ESS and SIAS.

Conclusions: This is the first study to design and incorporate a shame intervention component in WCBT and to validate its efficacy via a randomized controlled trial. The shame WCBT group showed a significant reduction in both shame and social anxiety after treatment compared to the normal WCBT and waiting groups. Problem-solving and self-blame mediated the effect of shame on social anxiety. In conclusion, this study supports previous findings that a direct shame-specific intervention component could enhance the efficacy of WCBT.

Trial Registration: Chinese Clinical Trial Registry ChiCTR2300072184; https://www.chictr.org.cn/showproj.html?proj=152757

(JMIR Ment Health 2024;11:e50535) doi:10.2196/50535

KEYWORDS

social anxiety disorder; web-based cognitive behavioral therapy; shame intervention; mediating effects; shame experience

Introduction

Social anxiety disorder (SAD), also known as social phobia, is characterized by an overwhelming fear of negative interpersonal evaluations and avoidance of embarrassing or shameful social situations [1], and it is one of the most common psychological disorders, with a 2% global prevalence [2]. The pathological model of SAD suggests that cognitive and behavioral patterns play an important role in the development and maintenance of social anxiety symptoms [3].

Previous SAD intervention research confirmed that socially anxious individuals generally have higher levels of shame [4-6], with more than half of patients with SAD reporting having experienced shame in situations of social anxiety [7,8]. A recent systematic review defined shame as a complicated experience including self-critical cognition, as well as safety and avoidance behavior [9], which shares similar manifestations with SAD [10,11]

Cognitive behavioral therapy (CBT) is considered the gold standard psychotherapy for SAD [12]. Several CBT framework group interventions with shame-specific interventions have showed favorable outcomes in enhancing the efficacy of SAD treatment in nonclinical samples. For example, Li et al [13] conducted cognitive behavioral group therapy (CBGT) among socially anxious individuals and found that CBGT with a shame intervention was significantly more effective than traditional CBT interventions; Golden [14] developed a shame CBGT program for social anxiety, adding the components of "acceptance of shame" and "exposure to shame" to the exposure component of the traditional CBT intervention, and successfully applied it to a group of university students. Therefore, given the consistent empirical evidence of a strong and robust positive relationship between shame and social anxiety [9,15,16], incorporating direct interventions for shame may also enhance the efficacy of CBT for treating SAD.

Web-based CBT (WCBT) has many significant advantages over face-to-face CBT, such as being less time-consuming, less costly, and easier to implement [17]. The effectiveness of WCBT is significant and can be maintained for 4 years after the end of the intervention [18-20]. In China, successive studies have confirmed that scores for social anxiety in WCBT intervention groups are significantly lower in posttests than pretests and that the efficacy persists for 6 months after the end of the intervention [6,21,22]. Furthermore, shame fully mediated the relationship between engagement in exposure and changes in social anxiety during the WCBT intervention [23]. However, no WCBT study has yet designed a component specifically for shame intervention, nor has any study investigated the efficacy of WCBT with a direct shame intervention in clinical patients with social anxiety.

Although shame-related thoughts (eg, "I'm annoying") could benefit from WCBT interventions like restructuring cognitive biases to adaptive and flexible thoughts [6,21], direct evidence supporting the efficacy of shame WCBT interventions remains scarce. Interventions for social anxiety in Eastern cultural contexts should particularly address issues of shame, as individuals might "normalize" their socially anxious behaviors and emotions as experiences of shame [24,25]. Moreover, shame is noted to influence the severity of social anxiety and further sustain the symptoms of social anxiety development in Eastern cultures [13]. Other randomized controlled trials on shame also found that improvements in shame experience could predict the reduction of social anxiety and positive treatment results [26,27]. Therefore, incorporating shame interventions into WCBT holds promise for helping individuals with social anxiety effectively cope with shame and social anxiety.

Overall, the role of shame interventions in WCBT for SAD has been significantly underexplored. Recent evidence indicates that integrating a more direct shame intervention could potentially increase the effectiveness of CBT [9]. Despite WCBT for social anxiety possibly benefiting shame experiences, a WCBT intervention with a shame-specific intervention might be more effective. The objective of this study was to design a shame intervention component within WCBT and investigate its effectiveness. Initially, the relationship between shame, coping styles, and social anxiety was explored through a longitudinal study, forming the basis for designing a shame-specific intervention in WCBT. Subsequently, after 2 rounds of measurements and screenings, patients with social anxiety were randomized into a shame WCBT, a normal WCBT, and a waiting group. Following the 8-week WCBT interventions, the study assessed differences in social anxiety scores among the 3 groups to evaluate the efficacy of the shame WCBT. Given the high association of social anxiety and depression [25], changes in depression scores before and after treatment were also measured in this study.

Based on the above overview, we hypothesized the following: first, there would be a significant difference between the pre- and posttests of social anxiety and shame between the shame WCBT group and the normal WCBT group, and the efficacy in the shame WCBT group would be significantly higher than that in the normal group; that is, the Social Interaction Anxiety Scale (SIAS), the Social Phobia Scale (SPS), and the Experience of Shame Scale (ESS) scores of the participants in the shame WCBT group would be significantly lower than in the normal WCBT group. Second, there would be a significant difference between the pre- and posttests for social anxiety, shame, and depression between both intervention groups and the waiting group. That is, after the treatment, the shame WCBT group and the normal WCBT group would show a greater reduction in SIAS, SPS, and ESS scores compared to the waiting group. Third, the direct effect of shame experience on social anxiety in the longitudinal study would be significant, and several coping styles would significantly mediate the effect between shame and social anxiety.

Methods

The Process of Designing the Shame Intervention Component for WCBT

The WCBT used in this study was adapted from a WCBT program developed in Switzerland. The main intervention sessions were divided into 5 components: motivation and psychoeducation, cognitive reconstruction, attention training, exposure exercises, and problem-solving (details of the specific translation process and session content are described in previous publications [21]). Based on this WCBT program, the shame intervention components were designed and added to a longitudinal study that investigated the relationship between shame and social anxiety. First, to explore shame, coping behaviors, and their relationship to social anxiety, 153 participants were recruited on the web and provided informed consent. They first completed the ESS and the Coping Styles Questionnaire (CSQ). After 1 month, they were assessed with the SIAS. Demographic variables and the results of this part of the study are shown under the subheading Demographic Information and Analysis of Shame Intervention Design. This longitudinal study attempted to provide guidance for the shame intervention in line with the CBT framework, which focuses on self-critical thoughts, social avoidance, and anxious experience [28,29]. The adaptation of the shame WCBT program targeted the modification with 3 treatment sections and homework assignments. First, for the psychoeducation section, additional psychoeducation on shame was incorporated into the section on understanding the anxiety experience and anxiety disorders. Participants could learn and recognize the concept of shame and understand the relationship between shame and social anxiety. It was emphasized that inappropriate coping with shame can even exacerbate symptoms of social anxiety. The assigned homework comprised an analysis of the role of shame in the circle of social anxiety. Second, for the cognitive construct section, interventions targeted on negative perceptions of shame and attribution (self-blame coping) were added to the section on rational thinking. For homework, participants were asked to identify shame-related irrational thoughts and provide evidence. Third, for the exposure section, participants were encouraged to recognize their shame experiences, confront the shame-inducing situation, and solve problems (problem-solving coping), including not avoiding scenarios that may trigger shame. In the section on authenticity testing, shame-related exposure exercises were added as homework. After the content of the revised program was approved and updated on the website, 7 undergraduate students who had never been exposed to WCBT and had not studied CBT were recruited as pretest participants. They followed a strict process of accessing the website, studying the revised version of the program, providing comments and suggestions, and engaging in discussions.

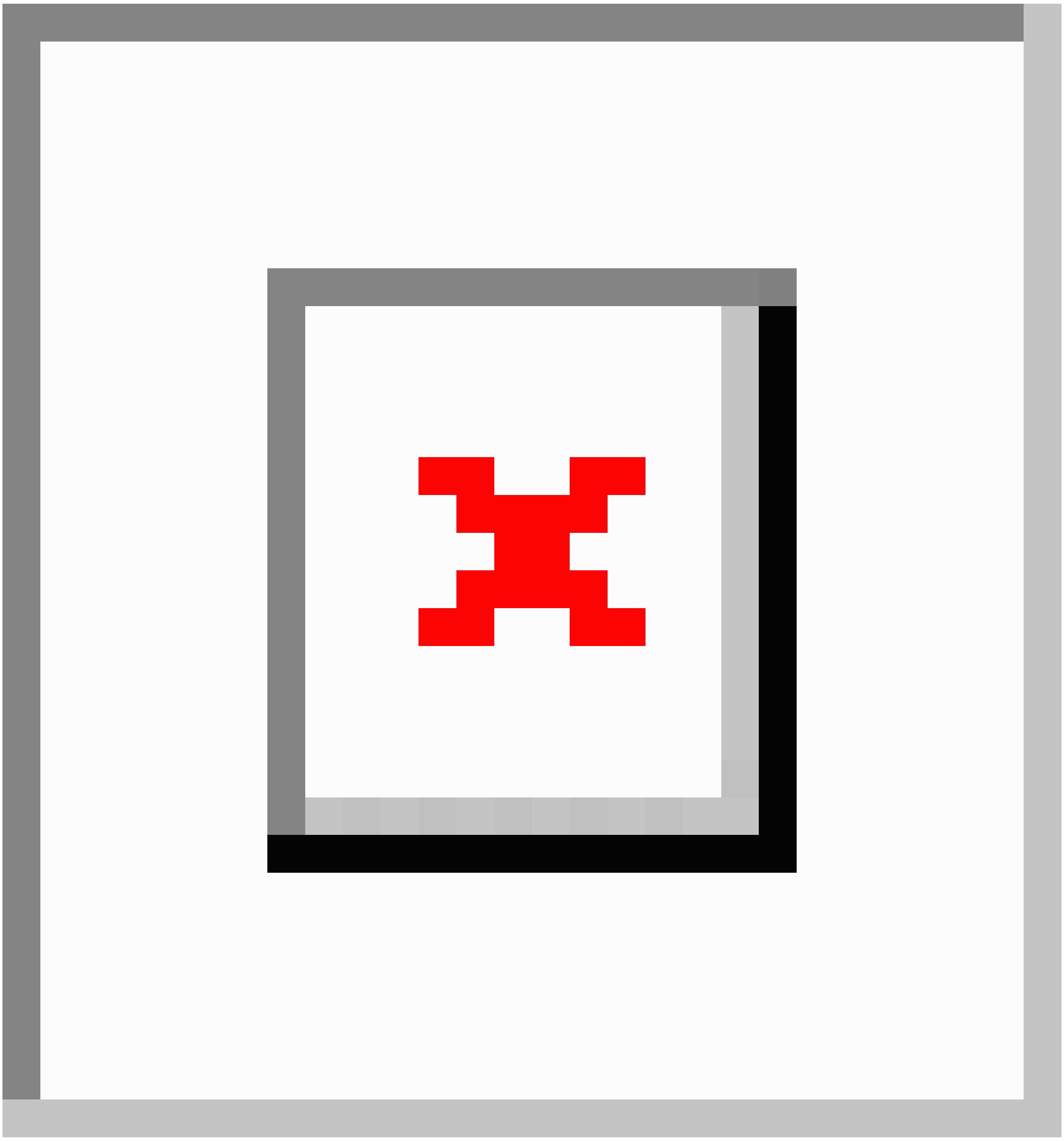
WCBT Intervention Recruitment and Screening Process

All the participants in the WCBT intervention were recruited through various online recruitment channels, including the official website, Weibo, WeChat, and other platforms. The first round of screening was based on the list described by Berger et al [30] of people who are suitable for project intervention: (1) a total SIAS score higher than 32 or a total SPS score higher than 22, (2) age 18 years or older, (3) no previous diagnosis of any mental illness or disorder, (4) no psychotropic medication within the last year, and (5) not having received any form of psychotherapy or psychological counseling. A total of 515 participants passed the first round of screening.

The second round of screening included diagnosed SAD patients and excluded participants who met the diagnostic criteria for certain types of psychiatric or psychological disorders. All participants who passed the first screening were invited to a one-on-one diagnostic interview. All interviewers were postgraduate or doctoral students in clinical psychology who were competent in the Mini-International Neuropsychiatric Interview (MINI) and were supervised by a licensed psychiatrist. At the end of the interview, a total of 201 participants met the following criteria and were able to enter the intervention period: (1) met the criteria in the MINI for SAD, (2) did not have a moderate or high risk of suicide, and (3) did not meet other diagnostic criteria, except major depression disorder.

After 2 rounds of screening, a total of 201 participants were enrolled in the project and received the WCBT intervention for 8 consecutive weeks. Participants were asked to complete pre-and posttest questionnaires before and after the WCBT. Upon logging on to the website for the first time, the website automatically randomly assigned them to the shame group, normal group, or waiting group at a 2:2:1 ratio. The shame and normal groups started the intervention immediately, while the waiting group waited 8 weeks before starting it. The screening, grouping process, and completion status of all participants are shown in Figure 1.

Figure 1. Flowchart of WCBT intervention screening process. MINI: Mini-International Neuropsychiatric Interview; SIAS: Social Interaction Anxiety Scale; SPS: Social Phobia Scale; WCBT: web-based cognitive behavioral therapy.



Materials and Measurement Tools

The Chinese versions of the SIAS and SPS were revised from the English versions developed by Mattick and Clarke [31,32]. The SIAS measures individuals' levels of anxiety, fear, and worry in general social interaction situations, while the SPS assesses the level of anxiety, fear, and worry in situations where individuals are observed by others. Each item in both scales is scored on a 5-point scale from 0 (not at all) to 4 (completely). The scale has good reliability and validity, with an internal consistency coefficient of 0.862 for SIAS and 0.904 for SPS.

The MINI is recognized as a simple and valid definitive interview tool for screening and assessing the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition* (*DSM-IV*) and the *International Statistical Classification of Mental Disorders* (*ICD-10*) [33,34]. The MINI has good reliability and validity.

The ESS examines individual shame in 3 areas: personality, behavior, and body. It consists of 25 questions [35]. Each item is rated on a scale from 1 (not at all) to 4 (often), with higher totals representing higher levels of shame. The scale has good reliability indicators, with internal consistency reliability of 0.825 and retest reliability of 0.88.

The CSQ classifies how individuals cope with difficulties in everyday situations in 6 dimensions: problem-solving, self-blame, help-seeking, fantasy, avoidance, and rationalization, with the number of items under each dimension varying from 4 to 9 [36]. All dimensions of the scale have good reliability and validity, with retest reliability ranging from 0.62 to 0.72. The results are relatively stable and reliable.

The Chinese version of the Beck Depression Inventory (BDI) was used to assess depression [37,38]. The internal consistency reliability of the BDI is 0.879, which indicates good structural validity.

WCBT Intervention Procedure

Participants with SAD were recruited and screened to compare the efficacy of the 2 different WCBT interventions (the shame WCBT and normal WCBT) for SAD and their effects on shame and anxiety levels over a period of 8 weeks. During the 8-week intervention period, participants were able to repeat or practice the material provided on the web, and they could also revise their submitted assignments [6]. All course content was presented in text format with the possibility to attach pictures, except for the relaxation training section, which had audio guidance. The recommended pace of study was 1 session per week plus time to complete and submit assignments, with approximately 2 to 3 hours spent per week [18].

Study Blinding, Data Processing, and Analysis

This study was conducted with a double-blind design. First, both participants and interviewers were unaware of the specifics of the shame intervention and group assignments. Blind assignments were achieved through the use of a computerized randomization process on the website, which automatically allocated participants to intervention groups. Finally, the data analysis phase was conducted on anonymized data sets, ensuring that researchers remained blinded to group allocations until the conclusion of the statistical analysis.

SPSS (version 24.0; IBM Corp) was used for statistical analysis of the data in this study. There was only 1 set of missing data, for the gender and age of 1 participant. We opted not to use imputation and proceeded with complete cases. Demographic information statistics and basic descriptive statistics were analyzed with an ANOVA in the first step. Comparative data were followed by an analysis of the differences in scores for SIAS, SPS, ESS, and BDI measured before and after the study period in the waiting group, normal WCBT group, and shame WCBT group with a repeated measures ANOVA. Longitudinal data from the intervention design used the process self-sampling procedure developed by Preacher and Hayes [39] to analyze the relationship between shame experiences, coping styles, and levels of social anxiety. The shame experience score was used as the independent variable, the 6 subscale scores for coping style as the mediating variables, the social interaction anxiety (from SIAS) score as the dependent variable, and gender and age as control variables; the sampling number was then set to 5000.

Sample Size Estimation

The sample size calculation for this study was conducted using G*Power (version 3.1; Universität Düsseldorf) [40]. This study aimed to detect at least a medium effect size (Cohen f=0.25) in between-group and within-group effects across the 3 intervention groups. Based on previous WCBT studies, which reported a correlation of approximately 0.5 between pre- and posttest scores, the sample size calculation for the repeated measures ANOVA indicated that each intervention group would require a minimum of 50 participants to achieve a statistical power of at least 0.8 at a significance level of .05. Considering previous studies have consistently shown medium to large effect sizes (d>0.5) for WCBT interventions compared to waiting groups, the sample size of the waiting group was set at half of the intervention groups.

Ethical Considerations

All participants provided informed consent before accessing the recruitment questionnaire. The informed consent form provided a detailed description of the purpose, content, screening process, possible benefits and risks, costs, and start and exit of the project. This study was approved by the Committee for Protecting Human and Animal Subjects, Department of Psychology, Peking University (20180504) and registered at Peking University. The trial registration number (Chinese Clinical Trial Registry) is ChiCTR2300072184. All data were anonymized during the data processing and analysis. No monetary compensation or fees were provided. Participation in the WCBT intervention itself was considered compensation for participants.

Results

Pre- and Posttest Scores Among Patients With SAD

A total of 201 participants successfully entered the WCBT intervention program after 2 rounds of screening, with an average age of 27.8 (SD 6.82) years; 67.7% (n=136) were female. According to automatic random assignment, 78 participants (mean age 26.6, SD 5.38 years; n=57, 73% female) were assigned to the shame WCBT group, 80 participants (mean age 28.1, SD 7.04 years; n=51, 65% female) were assigned to the normal WCBT group, and the 43 remaining participants (mean age 29.6, SD 8.32 years; n=28, 65% female) were assigned to the waiting group. The 1-way ANOVA revealed no significant differences in age across the 3 groups ($F_{2, 198}$=2.91; P=.06; ηp^2=0.03). Similarly, no significant differences were found in the baseline SIAS scores ($F_{2, 198}$=0.39; P=.68; ηp^2<0.01), the baseline SPS scores ($F_{2, 198}$=0.14; P=.87; ηp^2<0.01), the baseline ESS scores ($F_{2, 198}$=0.58; P=.56; ηp^2<0.01), or the baseline BDI scores ($F_{2, 198}$=0.90; P=.41; ηp^2<0.01). The χ^2 test also indicated no significant differences in gender among the 3 groups (χ^2_2=1.519; P=.47). After the 8-week intervention or waiting, 98 participants successfully completed at least 6 weeks of the program and submitted posttest measurements. The details of the pre- and posttest measurement scores are shown in Table 1.

Table . Pre- and posttest scores for the 3 groups.

	Shame WCBT[a], mean (SD) score		Normal WCBT, mean (SD) score		Waiting group, mean (SD) score	
	Pre (n=78)	Post (n=37)	Pre (n=80)	Post (n=35)	Pre (n=43)	Post (n=26)
Experience of Shame Scale	50.00 (12.57)	34.32 (11.42)	51.21 (13.43)	45.89 (11.85)	52.60 (12.59)	53.77 (10.21)
Social Interaction Anxiety Scale	49.39 (9.73)	30.16 (9.14)	50.20 (11.69)	38.43 (10.03)	51.11 (8.61)	49.85 (9.50)
Social Phobia Scale	37.67 (14.32)	20.97 (10.79)	37.11 (14.11)	26.25 (12.86)	38.44 (10.78)	39.54 (10.98)
Beck Depression Inventory	19.41 (10.33)	11.78 (10.00)	18.06 (8.62)	13.06 (6.85)	20.35 (9.43)	22.77 (10.83)

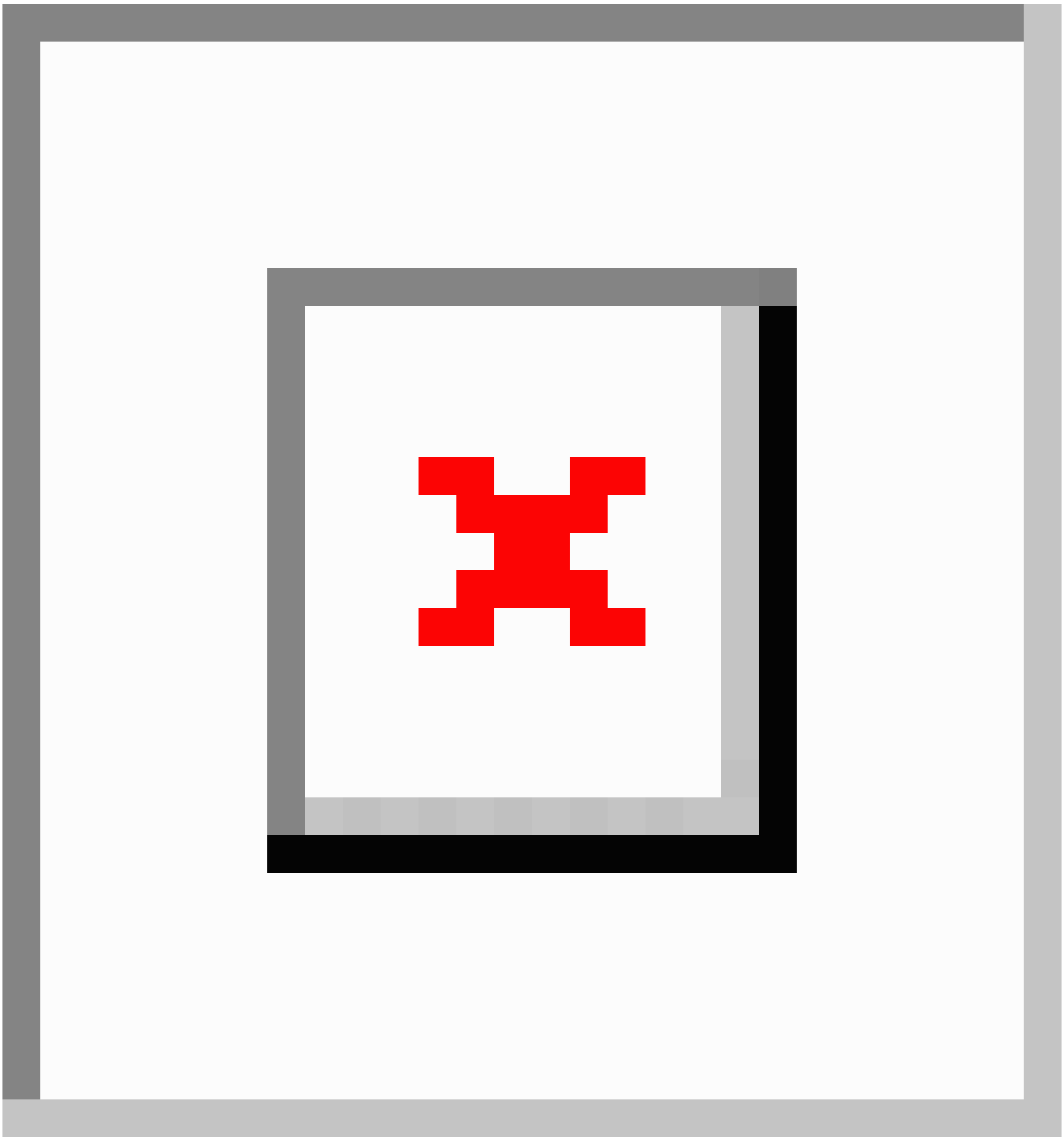
[a]WCBT: web-based cognitive behavioral therapy.

Repeated Measures ANOVA for Pre- and Posttest Scores in the 3 WCBT Groups

A repeated measures ANOVA was conducted to examine the differences in the scores on the SIAS, SPS, ESS, and BDI between the 3 groups before and after WCBT treatment. The analysis used group as the between-group variable, time point as the within-group variable, and all 97 participants' scores on the 4 scales (SIAS, SPS, ESS, and BDI) simultaneously as the dependent variables. Age was used as a covariate because there was a significant difference in age between the 3 groups before the test.

For the ESS, the results indicated a significant interaction between time and group on the dependent variable BDI ($F_{2, 93}$=13.40; P<.001; ηp^2=0.22), and the time×age interaction was not significant ($F_{2, 93}$=0.94; P=.34; ηp^2=0.01). The interaction is shown in Figure 2. The main effect of time was significant ($F_{1, 93}$=6.78; P=.01; ηp^2=0.07), the main effect of group was significant ($F_{2, 93}$=11.54; P<.001; ηp^2=0.20), and the main effect of age was not significant ($F_{1, 93}$=0.60; P=.44; ηp^2=<0.01). Further simple effects analysis was conducted, and the difference in scores between the 2 comparisons of the 3 groups in the pretest scores was not significant (all P values were >.31). At the posttest, the difference in scores between the 3 groups was significant. The shame WCBT group scored significantly lower on the posttest than the waiting group (mean deviation −19.56; P<.001), the normal WCBT group scored marginally lower on the posttest than the waiting group (mean deviation −7.06; P=.06), and the shame WCBT group scored significantly lower than the normal WCBT group (mean deviation −12.50; P<.001).

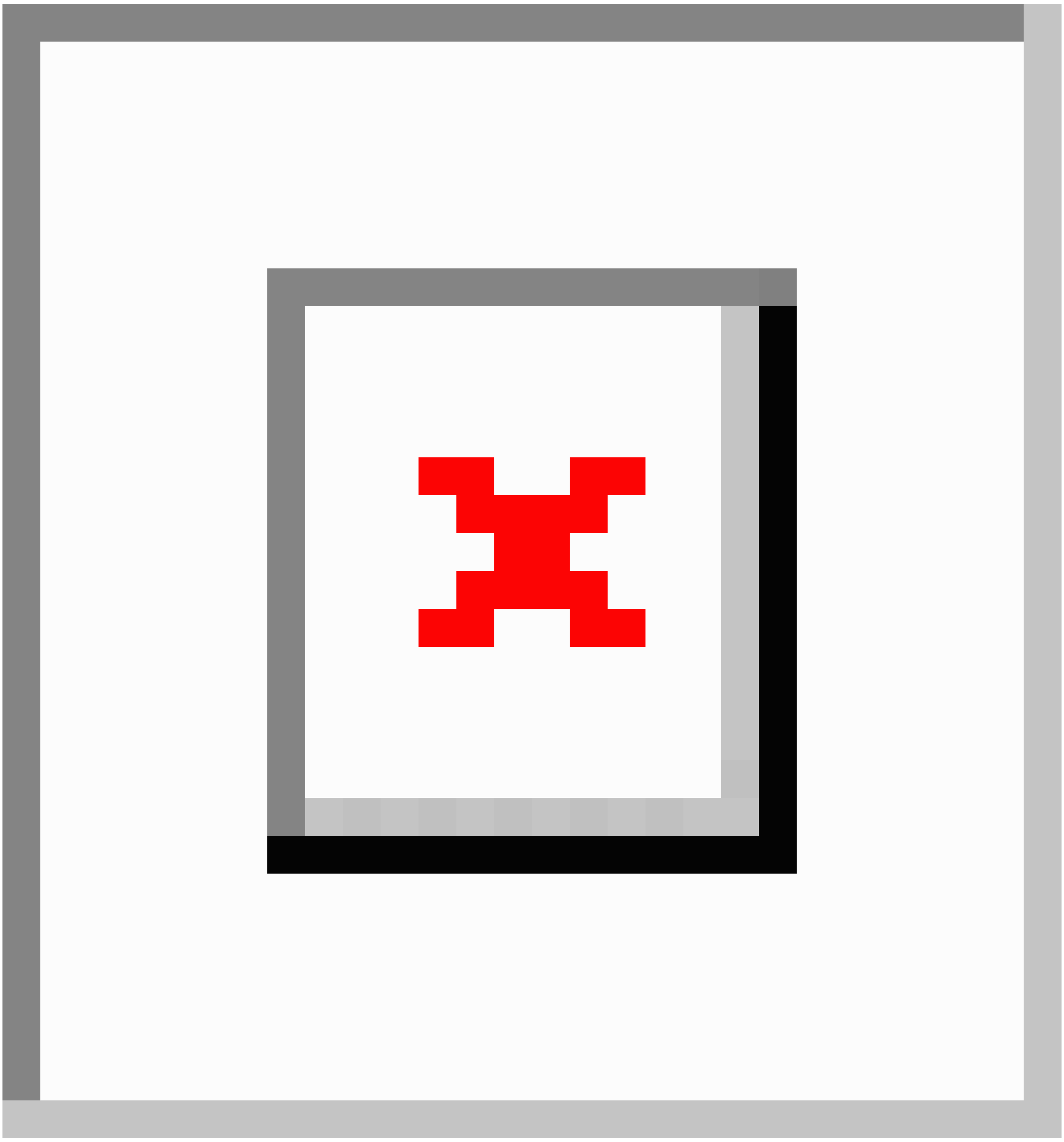
Figure 2. Experience of Shame Scale scores for the 3 groups. WCBT: web-based cognitive behavioral therapy.



The comparison of pre- and posttests in the 3 groups showed that the difference between the pre- and posttest scores in the shame WCBT group was significant (mean deviation −16.01; *P*<.001), and the difference between the pre- and posttest scores in the normal WCBT group was significant (mean deviation −6.06; *P*=.003). There was no significant difference in the waiting group (mean deviation −1.38; *P*=.54). These results indicate that patients with SAD in the shame intervention groups experienced significant reductions in shame levels (ESS) at the end of the intervention, and that the reduction was greater in the shame intervention group.

For the SIAS, the results indicated a significant time×group interaction on the dependent variable SIAS ($F_{2, 93}$=21.69; *P*<.001; ηp^2=0.32), but the time×age interaction was not significant ($F_{2, 93}$=0.11; *P*=.74; ηp^2<0.01). The interaction is shown in Figure 3. The main effect of time was significant ($F_{1, 93}$=6.99; *P*=.01; ηp^2=0.07) the main effect of group was significant ($F_{2, 93}$=16.81; *P*<.001; ηp^2=0.27), and the main effect of age was significant ($F_{1, 93}$=5.46; *P*=.02; ηp^2=0.06). Further simple effects analysis was conducted, showing that the difference in scores between the 2 comparisons of the 3 groups in the pretest was not significant (all *P* values were >.40). At posttest, the difference in scores between the 3 groups was significant; the shame WCBT group scored significantly lower on the posttest than the waiting group (mean deviation −20.70; *P*<.001), the normal WCBT group scored significantly lower on the posttest than the waiting group (mean deviation −11.12; *P*<.001), and the shame WCBT group scored significantly lower than the normal WCBT group (mean deviation −9.58; *P*<.001).

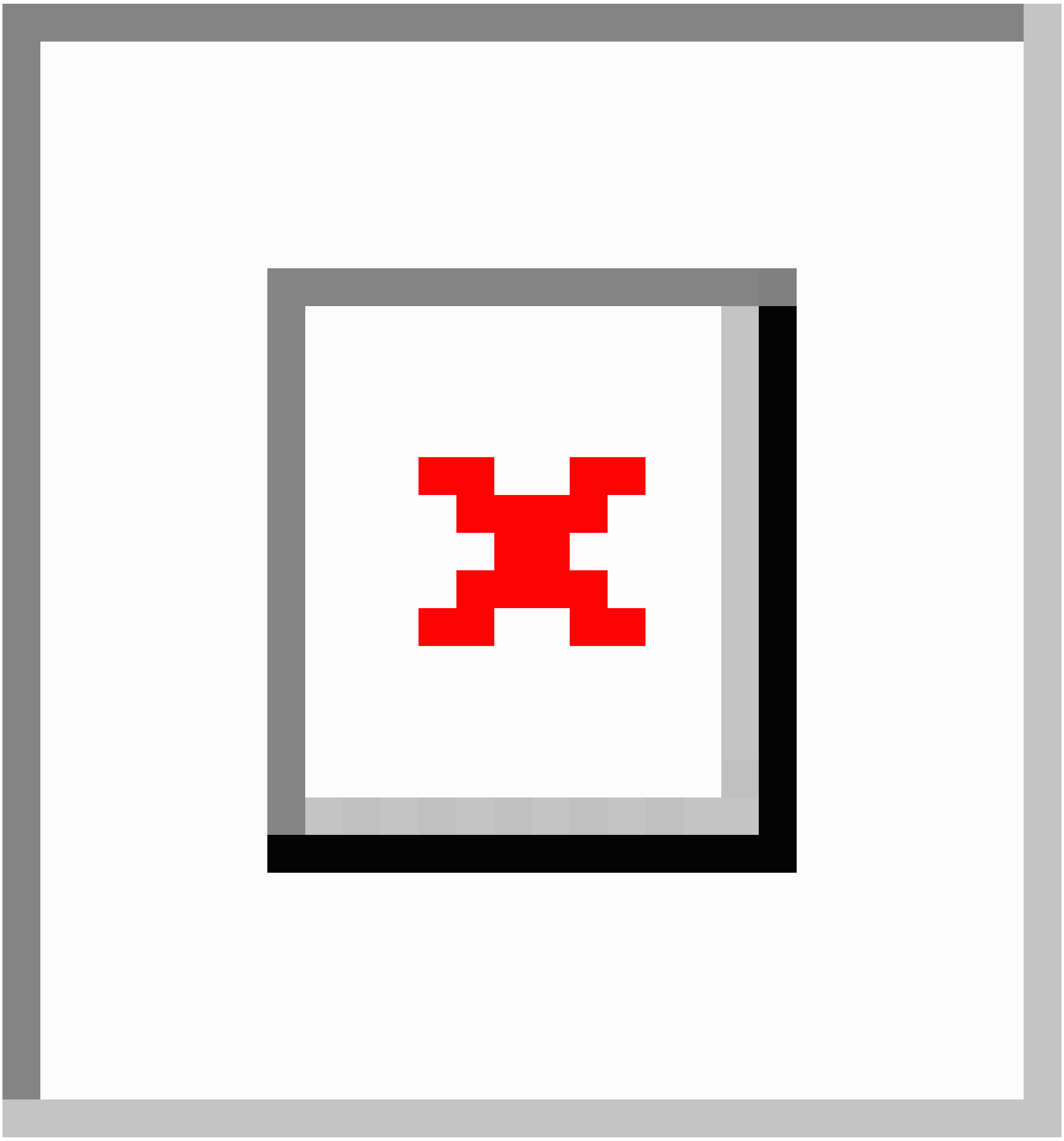
Figure 3. Social Interaction Anxiety Scale scores for the 3 groups. WCBT: web-based cognitive behavioral therapy.



The results of the comparison of pre- and posttests within the 3 groups showed that the difference between the pre- and posttest scores in the shame WCBT group was significant (mean deviation 20.07; *P*<.001), and the difference between the pre- and posttest scores in the normal WCBT group was significant (mean deviation 13.49; *P*<.001). There was no significant difference in the waiting group (mean deviation 3.07; *P*=.12). These results indicate that patients with SAD in both the normal and shame intervention groups experienced significant reductions in social anxiety levels (SIAS) at the end of the intervention and that the shame intervention group showed a higher reduction.

For the SPS, the results indicated a significant time×group interaction on the dependent variable SPS ($F_{2, 93}$=10.93; *P*<.001; ηp^2=0.19), and the time×age interaction was not significant ($F_{2, 93}$=1.16; *P*=.28; ηp^2=0.01). The interaction is shown in Figure 4. The main effect of time was significant ($F_{1, 93}$=6.67; *P*=.01; ηp^2=.07), the main effect of group was significant ($F_{2, 93}$=6.81; *P*<.001; ηp^2=0.13), and the main effect of age was not significant ($F_{1, 93}$=1.03; *P*=.31; ηp^2=0.01). Further simple effects analysis was conducted, and the difference in scores between the 2 comparisons of the 3 groups at pretest was not significant (all *P* values were>.99). At posttest, the difference in scores between the 3 groups was significant; the shame WCBT group scored significantly lower on the posttest than the waiting group (mean deviation −18.70; *P*<.001), the normal WCBT group scored significantly lower on the posttest than the waiting group (mean deviation −13.43; *P*<.001), and the shame WCBT group did not score significantly lower than the normal WCBT group (mean deviation −5.27; *P*=.20).

Figure 4. Social Phobia Scale scores for the 3 groups. WCBT: web-based cognitive behavioral therapy.



The results of the comparison of pre- and posttests within the 3 groups showed that there was a significant difference between the pre- and posttest scores in the shame WCBT group (mean deviation 17.12; *P*<.001) and the difference between the pre- and posttest scores in the normal WCBT group was significant (mean deviation 11.65; *P*<.001). There was no significant difference in the waiting group (mean deviation 0.52; *P*=.86). These results indicate that patients with SAD in both the normal and shame intervention groups experienced significant reductions in social phobia levels (SPS) at the end of the intervention and that there was no difference in the effects of the 2 interventions.

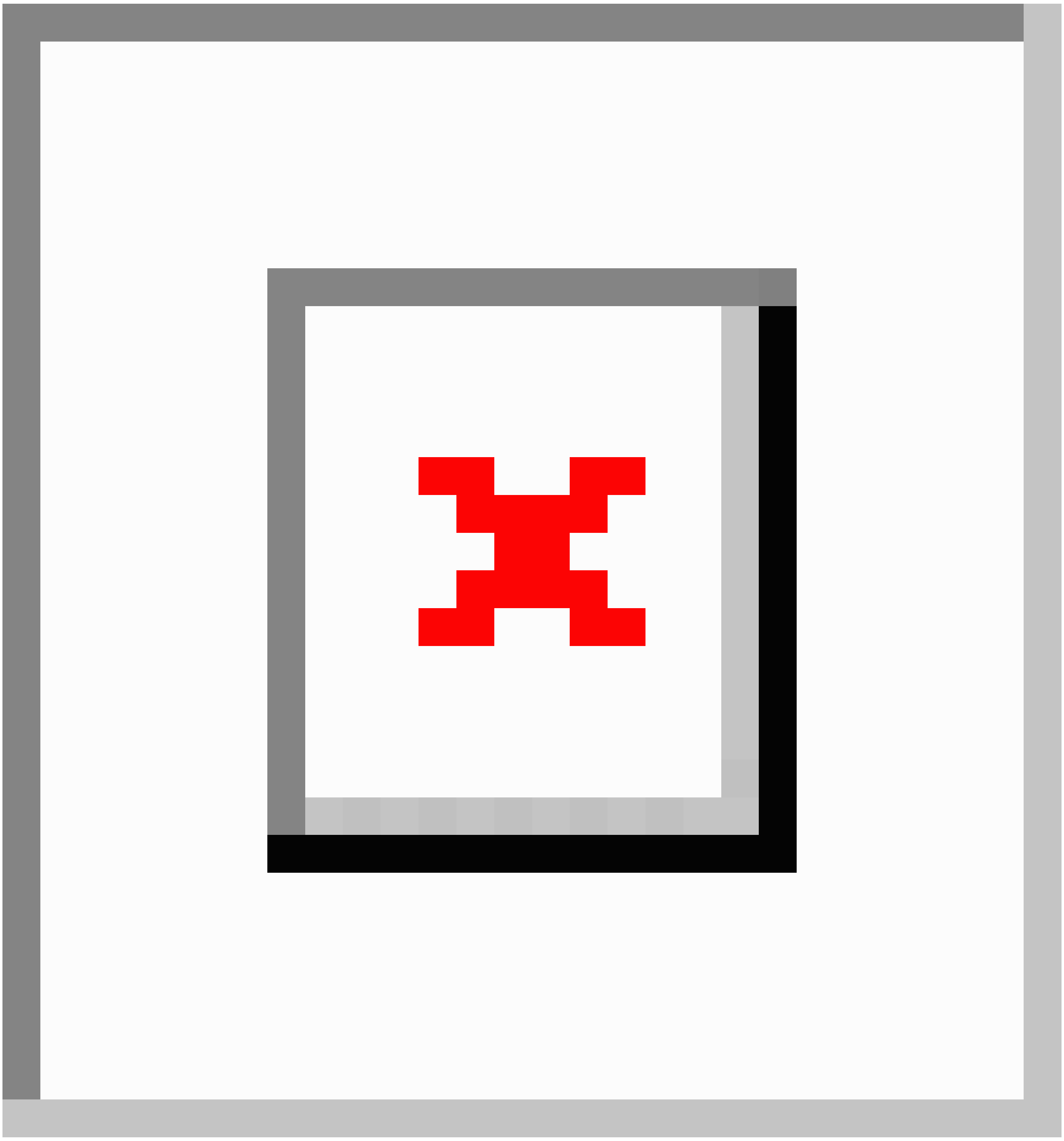
For the BDI scale, the results indicated a significant time×group interaction on the dependent variable BDI ($F_{2,93}$=6.90; *P*=.002; ηp2=0.13) and that the time×age interaction was not significant ($F_{2,93}$=0.20; *P*=.17; ηp^2=0.02). The main effect of time was significant ($F_{1,93}$=6.18; *P*=.02; ηp^2=0.06), the main effect of group was significant ($F_{2,93}$=6.37; *P*=.003; ηp^2=0.12), and the main effect of age was not significant ($F_{1, 93}$=0.18; P=.67; ηp^2=<0.01). Further simple effects analysis was conducted, and the difference in scores between the 2 comparisons of the 3 groups in the pretest was not significant (all P values were >.80). At posttest, the difference in scores between the 3 groups was significant; the shame WCBT group scored significantly lower on the posttest than the waiting group (mean deviation −10.46; P<.001), the normal WCBT group scored significantly lower on the posttest than the waiting group (mean deviation −10.26; P<.001), and the shame WCBT group did not score significantly lower than the normal WCBT group (mean deviation −0.02; P>.99).

The results of the comparison of pre- and posttests within the 3 groups showed that the difference between the pre- and posttest scores in the shame WCBT group was significant (mean deviation −7.14; P<.001) and that the difference between the pre- and posttest scores in the normal WCBT group was significant (mean deviation −7.54; P<.001). There was no significant difference in the waiting group (mean deviation 0.87; P=.65). These results indicate that patients with SAD in both the normal and shame intervention groups experienced significant reductions in depression levels (BDI) at the end of the intervention and that there was no difference in the effects of the 2 interventions.

Demographic Information and Mediation Analysis of Shame Intervention Design

A total of 136 longitudinal questionnaires were received (with 153 questionnaires collected at pretest); 107 (69%) participants who met the criteria for high social anxiety (SIAS) were screened for final data processing (mean age 23.67, SD 4.497 years; n=55 female). Their mean ESS score was 64.08 (SD 17.55) and mean SIAS score was 36.17 (SD 17.42). The mediation analysis focused on the 6 coping style variables (problem-solving, self-blame, help-seeking, fantasy, avoidance, and rationalization) as potential mediators between shame experience and social interaction anxiety (SIAS). The absolute values of the correlation coefficients of all independent variables for 2-by-2 comparisons were less than 0.7, indicating the absence of multicollinearity. The coefficients and significance results of each pathway are shown in Figure 5.

Figure 5. The coefficients and significance of coping style pathway. **P*<.05 and ***P*<.01.



When examining the results of the 6 mediating variables simultaneously, problem-solving and self-blame mediated the effect between shame experience and social interaction anxiety (SIAS). The 95% CI for problem-solving as a mediating variable was 0.025-0.217 and the indirect effect of the mediating variable was 0.115 (SE 0.049); the 95% CI for self-blame as a mediating variable was 0.024-0.339 and the indirect effect of the mediating variable was 0.165 (SE 0.082). Help-seeking, fantasy, avoidance, and rationalization were not significant mediating effects when used as mediating variables. The direct effect of shame experience on social interaction anxiety (SIAS) was 0.371 (SE 0.110; t_{96}=3.372; P=.003; R^2=0.26).

Discussion

This is the first study to design and incorporate a shame intervention component in WCBT and conduct a randomized controlled trial to validate the efficacy of the shame WCBT. Based on a comparison of the ESS scores, the shame WCBT group showed a significant reduction in shame level after treatment compared to the normal WCBT and waiting groups. The reduction of shame level in participants with social anxiety is in line with the treatment response of other CBT strategies with shame intervention [13,14]. Additionally, both the shame WCBT group and the normal WCBT group showed a significant reduction in social anxiety symptoms (SIAS and SPS scores) after receiving the 8-week WCBT intervention, while participants in the shame intervention group showed a more rapid reduction in SIAS scores. These results highlight the effectiveness of WCBT in alleviating their symptoms of social anxiety. Furthermore, the shame WCBT intervention was significantly more effective for SAD than the waiting control condition, and it prompted a more significant reduction in social interaction–related anxiety and shame experience than normal WCBT. This suggests that the shame WCBT had significant intervention effects for patients with SAD [18,20,21] and that the shame intervention component could enhance the efficacy of normal WCBT.

The design of the shame intervention component was derived from longitudinal data and the results of previous shame studies [4-7]. In the Chinese cultural context, there is a strong and significant correlation between shame and social anxiety [7,13,24]. One interesting finding is that problem-solving and self-blaming partially mediated the effect of shame on social anxiety. Specifically, the stronger the shame experience and the more negative the attitude toward problem-solving, the more likely the individual was to think and act in a self-blaming manner when faced with social difficulties and less likely to adopt the corresponding behavior. Consequently, social anxiety symptoms have been shown to intensify [41,42]. The positive effect of the shame intervention could be understood as follows: first, the new content concerning psychological education on shame helped participants to comprehend the significant positive association between shame and social anxiety symptoms [24]. Second, according to a previous study, self-blame leads to a worsening of social anxiety symptoms [41], while problem-solving is a specific mediating variable between shame and social anxiety in Chinese people [42]. After the shame WCBT, participants adjusted inward their attribution of shame and consolidated it with active coping with shame through cognition reconstruction and exposure. Therefore, this study supports previous findings that explicit intervention targeting shame could enhance the efficacy of CBT [9].

One interesting finding is the outcome with respect to the SPS, which is contrary to previous studies: the effects of the shame intervention were not significantly different from the normal WCBT intervention group. This difference may be attributed to the fact that the SIAS and SPS measure similar but different aspects. Although both scales showed a high correlation between the scores of socially anxious individuals [32], the focus of the 2 scales remains different. The SIAS focuses on individuals' anxiety in general social situations, and shame is more related to interpersonal interaction situations and tends to accompany social situations to avoid more experiences of social rejection [43]. However, the SPS focuses on anxiety levels when participants are observed or watched by others in different situations [31]. The significant changes in SPS scores after both WCBT interventions compared to the waiting group suggest the efficacy of WCBT, while the nonsignificant change in SPS score between the 2 different WCBTs may suggest that participants still worry about being evaluated in situations where they are observed [44].

It is noteworthy that there was no significant difference between the shame WCBT and the previous WCBT with respect to the reduction in depression levels. This might imply that the shame WCBT intervention did not specifically target depression. The reduction in depression levels in both WCBT groups is in agreement with previous findings that the alleviation of social anxiety levels can also, to a certain extent, contribute to the improvement of depression [6].

This study set out to develop the first shame-specific WCBT intervention for social anxiety and aimed to reduce social anxiety levels by psychoeducation and increasing individual initiative in problem-solving and reducing self-blame. The results show that the shame intervention was consistent with the traditional intervention on all indicators of social anxiety, shame, and depression. Further, it was more effective than the traditional WCBT program with respect to social anxiety in general situations, and participants who underwent the shame intervention also showed lower levels of shame. These results suggest that it is essential to explore shame-based WCBT for social anxiety; shame interventions could provide a deeper understanding of SAD [9].

There are some limitations and further directions that should be taken into consideration. First, the study did not examine changes in clinical outcomes over time of the 2 WCBT interventions. Although participants were screened with structured clinical interviews, the study relied on self-reported measures. This omission limits our understanding of long-term clinical changes, such as whether the participants made better use of problem-solving strategies to cope with shame, potentially leading to long-term benefits in managing SAD. Second, the relationship between shame and social anxiety is more prominent in Eastern cultures. Therefore, it is not known whether WCBT with a shame intervention in a Western cultural framework would yield better outcomes. It is also noteworthy that the dropout rate for the shame group was not significantly higher than that for the normal group, and adherence was comparable to a previous WCBT adherence study in an Eastern culture (for SAD, the dropout rate was 53.1%) [22]. While high dropout rates in WCBT studies are not unusual [45], they pose challenges to the reliability and validity of our findings. The shame intervention in this study primarily focused on cognitive reconstruction and exposure, aligning with the established WCBT framework. Future research could consider incorporating a broader range of therapeutic approaches to potentially enhance treatment adherence and reduce dropout rates.

Acknowledgments

We are grateful to Dr Chenghua Tian for the assistance with the recruitment phase. This study would not have been possible without the support of the Peking University and Ruhr University Bochum. We thank Ms Helen Vollrath for assistance with proofreading and language editing. This work was supported by the Nature Science Foundation of China (project U2336214).

Data Availability

The data sets generated during and/or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

XW contributed to conceptualization, data collection, data analysis, and writing of the paper. MG contributed to conceptualization and data collection and reviewed the paper. HC and TK contributed to conceptualization and data collection. MQ, JM, and TB contributed to conceptualization and reviewed the paper.

Conflicts of Interest

None declared.

Checklist 1
CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth; V 1.6.1)
[PDF File, 2903 KB - mental_v11i1e50535_app1.pdf]

References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, 5th edition: American Psychiatric Publishing; 2013. [doi: 10.1176/appi.books.9780890425596]
2. Stein DJ, Lim CCW, Roest AM, et al. The cross-national epidemiology of social anxiety disorder: data from the World Mental Health Survey Initiative. BMC Med 2017 Jul 31;15(1):143. [doi: 10.1186/s12916-017-0889-2] [Medline: 28756776]
3. Heimberg RG, Brozovich FA, Rapee RM. A cognitive-behavioral model of social anxiety disorder. In: Social Anxiety: Academic Press; 2014:705-728. [doi: 10.1016/B978-0-12-394427-6.00024-8]
4. Scheff TJ. Shame and conformity: the deference-emotion system. Am Soc Rev 1988 Jun;53(3):395-406. [doi: 10.2307/2095647]
5. Li B, Qian MY, Zhong J. Undergraduates' social anxiety: a shame proneness model. Chin Ment Health J 2005;19(5):304-306. [doi: 10.3321/j.issn:1000-6729.2005.05.003]
6. Zhao QX. Shame in social anxiety: the effect of ICBT [dissertation]. Peking University. 2014. URL: https://d.wanfangdata.com.cn/thesis/Y2666848 [accessed 2024-06-19]
7. Tangney JP. Assessing individual differences in proneness to shame and guilt: development of the Self-Conscious Affect and Attribution Inventory. J Pers Soc Psychol 1990 Jul;59(1):102-111. [doi: 10.1037//0022-3514.59.1.102] [Medline: 2213483]
8. Gilbert P. The relationship of shame, social anxiety and depression: the role of the evaluation of social rank. Clin Psychol Psychother 2000 Jul;7(3):174-189. [doi: 10.1002/1099-0879(200007)7:3<174::AID-CPP236>3.0.CO;2-U]
9. Swee MB, Hudson CC, Heimberg RG. Examining the relationship between shame and social anxiety disorder: a systematic review. Clin Psychol Rev 2021 Dec;90:102088. [doi: 10.1016/j.cpr.2021.102088] [Medline: 34598054]
10. Shi CS, Qian MY. Attribution styles and coping styles of shame-prone people. Chin Ment Health J 1998;12(4):193 CNKI:SUN:ZXWS.0.1998-04-000.
11. Piccirillo ML, Taylor Dryman M, Heimberg RG. Safety behaviors in adults with social anxiety: review and future directions. Behav Ther 2016 Sep;47(5):675-687. [doi: 10.1016/j.beth.2015.11.005] [Medline: 27816080]
12. Wong J, Gordon EA, Heimberg RG. Cognitive ‐ behavioral models of social anxiety disorder. In: Weeks JW, editor. The Wiley Blackwell Handbook of Social Anxiety Disorder: Wiley; 2014:1-23.
13. Li B, Qian MY, Ma CY. Group therapy on social anxiety of college students. Chin Ment Health J 2006;20(5):348-349. [doi: 10.3321/j.issn:1000-6729.2006.05.025]
14. Golden AS. Shame attacking exposure and experiential acceptance: a group treatment manual for social anxiety disorder [dissertation]. Palo Alto University. 2013. URL: https://www.proquest.com/openview/366836215ec5cd93ff0655fa83c519e9/1?pq-origsite=gscholar&cbl=18750 [accessed 2024-06-19]
15. Arditte KA, Morabito DM, Shaw AM, Timpano KR. Interpersonal risk for suicide in social anxiety: the roles of shame and depression. Psychiatry Res 2016 May 30;239:139-144. [doi: 10.1016/j.psychres.2016.03.017] [Medline: 27137975]
16. Broekhof E, Kouwenberg M, Oosterveld P, Frijns JHM, Rieffe C. Use of the Brief Shame and Guilt Questionnaire in deaf and hard of hearing children and adolescents. Assessment 2020 Jan;27(1):194-205. [doi: 10.1177/1073191117725169] [Medline: 28799409]
17. Hedman E, Andersson E, Ljótsson B, Andersson G, Rück C, Lindefors N. Cost-effectiveness of internet-based cognitive behavior therapy vs. cognitive behavioral group therapy for social anxiety disorder: results from a randomized controlled trial. Behav Res Ther 2011 Nov;49(11):729-736. [doi: 10.1016/j.brat.2011.07.009] [Medline: 21851929]

18. Berger T, Urech A, Krieger T, et al. Effects of a transdiagnostic unguided internet intervention ('velibra') for anxiety disorders in primary care: results of a randomized controlled trial. Psychol Med 2017 Jan;47(1):67-80. [doi: 10.1017/S0033291716002270] [Medline: 27655039]
19. Halmetoja CO, Malmquist A, Carlbring P, Andersson G. Experiences of internet-delivered cognitive behavior therapy for social anxiety disorder four years later: a qualitative study. Internet Interv 2014 Jul;1(3):158-163. [doi: 10.1016/j.invent.2014.08.001]
20. Williams AD, O'Moore K, Mason E, Andrews G. The effectiveness of internet cognitive behaviour therapy (iCBT) for social anxiety disorder across two routine practice pathways. Internet Interv 2014 Oct;1(4):225-229. [doi: 10.1016/j.invent.2014.11.001]
21. Kishimoto T, Krieger T, Berger T, Qian M, Chen H, Yang Y. Internet-based cognitive behavioral therapy for social anxiety with and without guidance compared to a wait list in China: a propensity score study. Psychother Psychosom 2016;85(5):317-319. [doi: 10.1159/000446584] [Medline: 27513757]
22. Chen H, Rodriguez MA, Qian M, Kishimoto T, Lin M, Berger T. Predictors of treatment outcomes and adherence in internet-based cognitive behavioral therapy for social anxiety in China. Behav Cogn Psychother 2020 May;48(3):291-303. [doi: 10.1017/S1352465819000730] [Medline: 31928568]
23. Wang H, Zhao Q, Mu W, Rodriguez M, Qian M, Berger T. The effect of shame on patients with social anxiety disorder in internet-based cognitive behavioral therapy: randomized controlled trial. JMIR Ment Health 2020 Jul 20;7(7):e15797. [doi: 10.2196/15797] [Medline: 32347799]
24. Zhong J, Wang A, Qian M, et al. Shame, personality, and social anxiety symptoms in Chinese and American nonclinical samples: a cross-cultural study. Depress Anxiety 2008;25(5):449-460. [doi: 10.1002/da.20358] [Medline: 17823960]
25. Hedman E, Ström P, Stünkel A, Mörtberg E. Shame and guilt in social anxiety disorder: effects of cognitive behavior therapy and association with social anxiety and depressive symptoms. PLoS One 2013;8(4):e61713. [doi: 10.1371/journal.pone.0061713] [Medline: 23620782]
26. Hoyer J, Wiltink J, Hiller W, et al. Baseline patient characteristics predicting outcome and attrition in cognitive therapy for social phobia: results from a large multicentre trial. Clin Psychol Psychother 2016;23(1):35-46. [doi: 10.1002/cpp.1936] [Medline: 25504802]
27. Lau AS, Fung J, Wang SW, Kang SM. Explaining elevated social anxiety among Asian Americans: emotional attunement and a cultural double bind. Cultur Divers Ethnic Minor Psychol 2009 Jan;15(1):77-85. [doi: 10.1037/a0012819] [Medline: 19209982]
28. Cox BJ, Fleet C, Stein MB. Self-criticism and social phobia in the US national comorbidity survey. J Affect Disord 2004 Oct 15;82(2):227-234. [doi: 10.1016/j.jad.2003.12.012] [Medline: 15488251]
29. Moscovitch DA. What is the core fear in social phobia? A new model to facilitate individualized case conceptualization and treatment. Cogn Behav Pract 2009 May;16(2):123-134. [doi: 10.1016/j.cbpra.2008.04.002]
30. Berger T, Hohl E, Caspar F. Internet-based treatment for social phobia: a randomized controlled trial. J Clin Psychol 2009 Oct;65(10):1021-1035. [doi: 10.1002/jclp.20603] [Medline: 19437505]
31. Mattick RP, Clarke JC. Development and validation of measures of social phobia scrutiny fear and social interaction anxiety. Behav Res Ther 1998 Apr;36(4):455-470. [doi: 10.1016/s0005-7967(97)10031-6] [Medline: 9670605]
32. Ye D, Qian M, Liu X, Chen X. Revision of Social Interaction Anxiety Scale and Social Phobia Scale. China J Clin Psychol 2007;15:115-119. [doi: 10.3969/j.issn.1005-3611.2007.02.002]
33. Sheehan DV, Lecrubier Y, Sheehan KH, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 1998;59 Suppl 20:22-33. [Medline: 9881538]
34. Si TM, Shu L, Dang WM, et al. Evaluation of the reliability and validity of Chinese version of the Mini-International Neuropsychiatric Interview in patients with mental disorders. Chin Ment Health J 2009;07:493-497. [doi: 10.3969/j.issn.1000-6729.2009.07.011]
35. Qian MY, Andrews B, Zhu RC, Wang A. The development of shame scale of Chinese college students. Chin Ment Health J 2000;14(4):217-221. [doi: 10.3321/j.issn:1000-6729.2000.04.001]
36. Xiao JH, Xu XF. Validity and reliability of the Coping Styles Questionnaire. Chin Ment Health J 1996;10(4):164-168 CNKI:SUN:ZXWS.0.1996-04-008.
37. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J. An inventory for measuring depression. Arch Gen Psychiatry 1961 Jun;4:561-571. [doi: 10.1001/archpsyc.1961.01710120031004] [Medline: 13688369]
38. Zhang YX, Wang Y, Qian MY. Reliability and validity of the Beck Depression Inventory. Chin Ment Health J 1990;4(4):164-168 CNKI:SUN:ZXWS.0.1990-04-006.
39. Preacher KJ, Hayes AF. SPSS and SAS procedures for estimating indirect effects in simple mediation models. Behav Res Methods Instrum Comput 2004 Nov;36(4):717-731. [doi: 10.3758/bf03206553] [Medline: 15641418]
40. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
41. Shahar B, Doron G, Szepsenwol O. Childhood maltreatment, shame-proneness and self-criticism in social anxiety disorder: a sequential mediational model. Clin Psychol Psychother 2015;22(6):570-579. [doi: 10.1002/cpp.1918] [Medline: 25196782]

42. Sun SJ, Zhao FC. A correlation study among general self-efficacy, coping styles and social anxiety in college students. Chin J Health Psychol 2008;16(3):305-307. [doi: 10.3969/j.issn.1005-1252.2008.03.027]
43. Michail M, Birchwood M. Social anxiety disorder and shame cognitions in psychosis. Psychol Med 2013 Jan;43(1):133-142. [doi: 10.1017/S0033291712001146] [Medline: 22608200]
44. Gao XD, Zhou AB, Xia RX. Research progress and future prospects on the relationship between guilt and shame. Chin Ment Health J 2008;22(7):534-537.
45. Beatty L, Binnion C. A systematic review of predictors of, and reasons for, adherence to online psychological interventions. Int J Behav Med 2016 Dec;23(6):776-794. [doi: 10.1007/s12529-016-9556-9] [Medline: 26957109]

Abbreviations

BDI: Beck Depression Inventory
CBGT: cognitive behavioral group therapy
CBT: cognitive behavioral therapy
CSQ: Coping Styles Questionnaire
ESS: Experience of Shame Scale
MINI: Mini-International Neuropsychiatric Interview
SAD: social anxiety disorder
SIAS: Social Interaction Anxiety Scale
SPS: Social Phobia Scale
WCBT: web-based cognitive behavioral therapy

Edited by S Gardezi, TDA Cardoso; submitted 04.07.23; peer-reviewed by L Yang, N Esfandiari; revised version received 29.03.24; accepted 25.05.24; published 27.06.24.

Please cite as:
Wen X, Gou M, Chen H, Kishimoto T, Qian M, Margraf J, Berger T
The Efficacy of Web-Based Cognitive Behavioral Therapy With a Shame-Specific Intervention for Social Anxiety Disorder: Randomized Controlled Trial
JMIR Ment Health 2024;11:e50535
URL: https://mental.jmir.org/2024/1/e50535
doi:10.2196/50535

Original Paper

Understanding Public Perceptions of Virtual Reality Psychological Therapy Using the Attitudes Towards Virtual Reality Therapy (AVRT) Scale: Mixed Methods Development Study

Aislinn D Gomez Bergin[1,2,3,4*], PhD; Aoife M Allison[3*], MSc; Cassie M Hazell[5*], PhD

[1]National Institute of Health and Care Research MindTech MedTech Co-operative, Institute of Mental Health, University of Nottingham, Nottingham, United Kingdom

[2]National Institute of Health and Care Research Nottingham Biomedical Research Centre, University of Nottingham, Nottingham, United Kingdom

[3]Mental Health and Clinical Neurosciences, School of Medicine, University of Nottingham, Nottingham, United Kingdom

[4]School of Computer Science, University of Nottingham, Nottingham, United Kingdom

[5]Department of Psychological Interventions, School of Psychology, University of Surrey, Guilford, United Kingdom

[*]all authors contributed equally

Corresponding Author:
Aislinn D Gomez Bergin, PhD
National Institute of Health and Care Research MindTech MedTech Co-operative
Institute of Mental Health
University of Nottingham
Jubilee Campus
Triumph Rd
Nottingham, NG7 2TU
United Kingdom
Phone: 44 115 82 30431
Email: aislinn.bergin@nottingham.ac.uk

Abstract

Background: Virtual reality (VR) psychological therapy has the potential to increase access to evidence-based mental health interventions by automating their delivery while maintaining outcomes. However, it is unclear whether these more automated therapies are acceptable to potential users of mental health services.

Objective: The main aim of this study was to develop a new, validated questionnaire to measure public perceptions of VR therapy (VRT) guided by a virtual coach. We also aimed to explore these perceptions in depth and test how aspects such as familiarity with VR and mental health are associated with these perceptions, using both quantitative and qualitative approaches.

Methods: We used a cross-sectional mixed methods design and conducted an exploratory factor analysis of a questionnaire that we developed, the Attitudes Towards Virtual Reality Therapy (AVRT) Scale, and a qualitative content analysis of the data collected through free-text responses during completion of the questionnaire.

Results: We received 295 responses and identified 4 factors within the AVRT Scale, including attitudes toward VRT, expectation of presence, preference for VRT, and cost-effectiveness. We found that being more familiar with VR was correlated with more positive attitudes toward VRT (factor 1), a higher expectation of presence (factor 2), a preference for VRT over face-to-face therapy (factor 3), and a belief that VRT is cost-effective (factor 4). Qualitative data supported the factors we identified and indicated that VRT is acceptable when delivered at home and guided by a virtual coach.

Conclusions: This study is the first to validate a scale to explore attitudes toward VRT guided by a virtual coach. Our findings indicate that people are willing to try VRT, particularly because it offers increased access and choice, and that as VR becomes ubiquitous, they will also have positive attitudes toward VRT. Future research should further validate the AVRT Scale.

(JMIR Ment Health 2024;11:e48537) doi:10.2196/48537

KEYWORDS

psychological interventions; digital; virtual reality; virtual agent; mental health; presence

Introduction

Background

Virtual reality (VR) is an immersive environment where people can interact using either computer equipment, such as a screen and mouse, or VR-enabled headsets and controllers, where additional sensors can track the users' actions in real time. The latter application provides people with a greater sense of presence, a term used to describe how closely a virtual environment is interpreted as real [1]. In recent years, VR has been used successfully in a range of health care settings to improve and increase access to treatment [2]. In particular, VR has been used in the delivery of psychological therapies for a range of mental health problems, with several decades of evidence demonstrating its clinical efficacy in the treatment of psychosis, depression, anxiety, and eating disorders [3-8].

VR therapies (VRTs) were initially developed to be used by therapists as an adjunct or tool in their delivery of therapy. However, the need for a real-world therapist to deliver VRTs presents a key challenge for their widespread implementation [9]. Researchers have shown that the automation of some therapeutic elements may overcome this barrier to meet the increase in demand for treatment globally [10,11]. Emerging evidence demonstrates that VRTs can be successfully delivered with little to no therapist involvement, with virtual coaches supporting people receiving therapy for fear of heights and agoraphobia in the context of psychosis [10,11]. Virtual coaches are also known as virtual agents [12]. These characters are not under human control and therefore offer automation of therapies, in which dialogue and responses are scripted instead of the formulation that is offered by real-world therapists.

There are financial and resource incentives for mental health services to offer more automated therapies [13]. Clinicians also appear to be in support of VRTs. For example, cognitive behavior therapists [14] and psychiatric health care staff [15] reported positive attitudes toward VRT, particularly when they were more familiar with VR. However, these studies do not consider how staff feel about VRT guided by a virtual coach and, notably, do not explore patient and public perceptions of VRT. VRT dropout rates have been used as a proxy measure of patient experience, and these figures show similar dropout rates to therapies delivered without VR [16]. However, dropout rates from research does not provide us with a clear picture of whether people will engage in therapies delivered using VR, including those guided by a virtual coach. A content analysis of social media posts by the public appears to suggest an interest in the application of VR in mental health care [17]. Staff and service users also have positive views toward their use in mental health inpatient facilities [18]. However, these studies still do not directly ask potential users of mental health services whether they would be willing to try VRT guided by a virtual coach or the factors that relate to this willingness.

Health care staff, when asked for their views regarding service users' opinions of VRT, had concerns regarding patients' willingness to accept their use as part of their mental health care package [19]. Furthermore, the literature has highlighted a lack of personalization as a barrier to engagement with digital mental health interventions [20]. It is unclear whether this indicates that automated VRTs can be sufficient when scripts are relevant to the experience of the individual. It is possible that the presence of a virtual coach may encourage more positive attitudes and a willingness to try VRTs. There is a need to understand service user and public perspectives on the use and delivery of VRTs and those guided by a virtual coach and how different factors may affect the uptake of such interventions.

Aim

The main aim of this study was to establish a new, validated questionnaire to measure the perceptions of VRT guided by a virtual coach. Second, we aimed to explore how these perceptions are associated with familiarity with VR and mental health, using both quantitative and qualitative approaches.

Methods

Study Design

This study used mixed methods with a cross-sectional design. Data were collected from a web-based questionnaire using Jisc software [21].

Participants

The participants were recruited via social media to complete the web-based questionnaire. We aimed to recruit a minimum of 200 participants in line with sample size recommendations for exploratory factor analysis (EFA) [22]. To be eligible to participate, persons were required to be a resident of the United Kingdom or Ireland and aged ≥18 years. A link to the web-based survey was included in all promotional materials.

Measures

Demographics

The participants were asked to provide basic demographic information, including their age and sex, as well as whether they were identified as having a mental health condition, had ever experienced therapy, or had supported anyone with a mental health condition. Furthermore, they were asked about their experience of VR (from never to ≥10 times) and their familiarity with VR, VRT, and mental health conditions.

The Attitudes Towards Virtual Reality Therapy Scale

The Attitudes Towards Virtual Reality Therapy (AVRT) Scale was developed by AMA and ADGB. Items were based on themes identified in previous literature that contribute to perceptions of VRT and digital mental health interventions [14,23-29]. Items surrounding the virtual coach drew on the literature related to therapeutic alliance and focused on trust, comfort, and need [30].

We designed 54 items all assessing different aspects of attitudes toward VRT, including 9 items related to attitudes toward VRT delivered by a virtual coach. Each item used a 7-point Likert scale where participants rated their agreement from "strongly agree" to "strongly disagree." Strong agreement or disagreement with 16 of these items triggered a free-text question for participants to provide context using free-text responses. Responses were scored from 1 (strongly disagree) to 7 (strongly agree). A higher score indicates more positive perceptions of VRT. A total of 27 items were reverse-worded and therefore reverse-coded.

Furthermore, participants were invited to respond to 3 additional free-text questions asking what they would like to know more about, how they think their level of VR experience has influenced their perceptions, and what they think the best setting for VRT would be.

Procedure

Upon opening the questionnaire, participants were first shown the information sheet, followed by a consent statement. After consenting, participants were asked to enter a unique identification code so that their anonymized responses could be identified later. Participants were then asked to provide basic demographic information, followed by an explanatory paragraph (Multimedia Appendix 1) about VRT and the virtual coach, which was described as “a computer-generated avatar” that “guides the patients through the scenarios and offers advice and encouragement.” This was followed by items on experience with VR and mental health, the AVRT Scale, and the 3 free-text questions. After completing these questionnaires, participants were presented with a debrief statement.

Ethics Approval

Ethics approval was granted by the Division of Psychiatry and Applied Psychology Ethics Subcommittee of the University of Nottingham (Project ID 1534).

Statistical Analysis

Raw data were downloaded from Jisc [21] into SPSS Statistics software (version 25; IBM Corp) [31]. We removed responses from participants who did not meet the inclusion criteria, did not provide consent, or had missing data. Sample characteristics were summarized using descriptive statistics.

To validate our new questionnaire, we conducted an EFA using principal component analysis with a varimax rotation. We assessed the suitability of the data for factor analysis using Bartlett’s test of sphericity and the Kaiser-Meyer-Olkin measure of sampling adequacy (acceptable adequacy ≥0.6). All 54 items using Likert scales were included in the EFA. Items were first screened to check for multicollinearity and poor correlations with the other items. We operationalized this screening by assessing the determinant and searching for any interitem correlations of ≥0.7 or where most coefficients were nonsignificant or <0.4. Any items that failed this initial screening were removed, and the EFA was rerun. Factors were derived using eigenvalues ≥1, where the Kaiser criterion [32] was met, and in combination with the point of inflection on the scree plot, where they were not. For factors to be retained, they must comprise at least 3 items. Where items loaded on >1 factor, the item was assigned to the factor with which it made the most thematic sense.

The questionnaire included both positively and negatively worded items. Once the factors were established, we reverse-scored negatively worded items so that a higher score indicated a more favorable attitude. We then assessed the internal consistency of the final factor structure using Cronbach α, with an acceptable internal consistency of ≥0.7 [33]. Further items may be removed at this point, where the scale reliability can be substantially improved if the item is removed. The factor scores were computed using the mean and SD of the scale sum.

We conducted a series of Pearson *r* correlations to assess whether there was a relationship among the scale totals of the derived factors and lived experience of VR, VRT, and mental health problems.

Qualitative Analysis

All responses to the free-text response questions (ie, 16 free-text boxes triggered by extreme responses to survey questions and 3 additional free-text questions) were uploaded to NVivo (version 12 for Mac; QSR International). Qualitative content analysis [34] was used to quantify and summarize the qualitative data within the broader context of the AVRT Scale. All data were coded inductively by a qualitative researcher (ADGB), where several codes could be applied to a single response. The codes were collated by questions or items. The study team met to review and revise any discrepancies or discuss any questions. Findings were then summarized according to each question or item and were presented within the factors of the AVRT Scale.

Results

Sample Characteristics

We collected 295 responses to the survey. Our sample reflected a range of age groups. The majority were female, had used VR at least once, and had no personal or professional experience with mental health problems. However, most participants had supported a friend or family member with poor mental health (Table 1).

Table 1. Sample characteristics (N=295).

Characteristics	Values
Age group (years), n (%)	
18-24	89 (30.2)
25-29	39 (13.2)
30-39	45 (15.3)
40-49	49 (16.6)
50-64	64 (21.7)
>65	9 (3.1)
Sex, n (%)	
Male	83 (28.1)
Female	209 (70.8)
Prefer not to say	3 (1)
Frequency of experiencing VR[a], n (%)	
Never	112 (38)
Once	48 (16.3)
<5 times	88 (29.8)
5-9 times	20 (6.8)
≥10 times	27 (9.2)
Participants identifying as having a mental health condition, n (%)	
Yes	102 (34.6)
No	181 (61.4)
Prefer not to say	12 (4.1)
Participants with experience in therapy for a mental health condition, n (%)	
Yes	131 (44.4)
No	160 (54.2)
Prefer not to say	4 (1.4)
Participants who have supported a friend or family member or colleague with a mental health condition, n (%)	
Yes	253 (85.8)
No	39 (13.2)
Prefer not to say	3 (1)
Participants who have worked in a caring role for people with mental health conditions, n (%)	
Yes	108 (36.6)
No	186 (63.1)
Prefer not to say	1 (0.3)
Participants familiar with VR, mean (SD)	2.33 (1.02)
Participants familiar with VR therapy, mean (SD)	1.35 (0.75)
Participants familiar with mental health conditions, mean (SD)	3.65 (0.96)

[a]VR: virtual reality.

Quantitative Analysis

Item Screening

The data were found to be appropriate for EFA (Kaiser-Meyer-Olkin=0.93; χ^2_{1431}=10,205.5, *P*<.001). The determinant suggested that there was multicollinearity, and inspection of the correlation coefficients revealed 2 pairs of items that were highly correlated (items 7 and 8=0.83; items 36 and 37=0.87); therefore, we removed 1 item from each pair of correlations (items 8 and 36). Items 13, 32, and 33 were removed, as the majority of interitem correlations were nonsignificant (P>.05). Furthermore, we removed items 6, 19, 25, 26, 28, 31, and 47 as either all or all but one of the correlation coefficients was <0.4. In total, we removed 12 items and then reran the EFA on the remaining 42 items.

The determinant again indicated that multicollinearity was an issue. We identified 3 pairs of correlations with coefficients >0.7 (items 10 and 11=0.76; items 29 and 49=0.76; items 35 and 37=0.71); therefore, we removed 1 item from each pair (items 10, 29, and 35) and reran the EFA on the remaining 39 items.

Exploratory Factor Analysis

The Kaiser criterion was met (n=295; average communalities 0.64) [32]. Therefore, the factor structure was determined based on eigenvalues >1. The rotated factor solution suggested 7 factors, which explained 63.64% of the variance. However, 3 factors were not retained because they contained <3 items. The removal of these factors resulted in the removal of items 7, 23, 24, 27, 48, and 49. The resulting 33 items were entered into a final EFA. A 4-factor solution was suggested based on the eigenvalues and the scree plot, which explained 58.61% of the variance.

Factor 1 had 13 items that assessed respondents' support for VRT, including 6 reverse-worded items (factor 1: attitude toward VRT). Factor 2 had 9 items. These items, including 7 reverse-worded items, assessed the extent to which the respondents expected VRT to be immersive (factor 2: the expectation of presence). Factor 3 had 7 items asking respondents to compare VRT to aspects of face-to-face therapies (factor 3: preference for VRT). Factor 4 had 4 items each assessing different aspects of the cost-effectiveness of VRT (factor 4: cost-effectiveness). Refer to Multimedia Appendix 2 for the final factor structure.

Scale Reliability

After reverse-scoring the reverse-worded items, we computed Cronbach α values for each scale. All scales had strong internal consistency (all Cronbach α≥.82). The scale reliabilities could not be improved by removing any of the items. A higher score on each of the subscales suggested a more favorable attitude (factor 1), increased perceived presence (factor 2), a preference for VRT over traditional therapies (factor 3), and agreement that VRT is cost-effective (factor 4). The desired direction for each subscale to demonstrate support for VRT was high for factors 1, 3, and 4 and low for factor 2.

Relationship Between Scales and Lived Experience

There was a significant relationship between the participants' familiarity with VR and their scores on all the factors. Familiarity with VR was positively associated with a more favorable attitude toward VRT (factor 1), higher expectations of presence (factor 2), a preference over face-to-face therapy (factor 3), and a belief that VRT is cost-effective (factor 4). We also found significant positive correlations between factors 1, 2, and 3, but not factor 4, and familiarity with the VRT. There was no significant correlation between mental health familiarity and the scores for any factors. Multimedia Appendix 2 presents the correlation coefficients and associated significance scores.

Qualitative Results

Qualitative Questions

Table 2 presents the initial qualitative questions that all participants were asked.

Table 2. Initial qualitative questions (N=295).

Type of data	Question	Response, n (%)
Qualitative question 1	Which aspects of virtual reality therapy guided by a virtual coach would you like to know more about?	226 (76.6)
Qualitative question 2	How has your previous experience of virtual reality (minimal or extensive) influenced your perceptions of virtual reality therapy guided by a virtual coach?	236 (80)
Qualitative question 3	If you were offered virtual reality therapy guided by a virtual coach, where do you think the best place to do the therapy would be?	245 (83)

Which Aspects of VRT Guided by a Virtual Coach Would You Like to Know More About?

Of the 226 participants, 19 (8.4%) indicated that they did not want to know anything more. Those who provided reasons indicated that they did not want to try or did not know enough. Of them, 22 (9.7%) participants indicated that, as they did not know enough, they would like to find out more; 5 (2%) indicated that they would like to try out VRT; and 9 (3.9%) asked regarding its cost. In total, 13 (5.8%) participants were curious about the conditions that could be targeted with the VRT, specifically regarding its use for anxiety disorders, depression, and emotion regulation.

Many participants asked how it could be tailored or personalized for them (29/226, 12.8%). This meant thinking about their position within the interaction, asking about safety or how much control they would have, and whether VRT might have a negative effect and how this would be monitored. Of 226 participants, 47 (20.8%) asked about the virtual coach, wanting to know how real it would be, how much of the language would be generic or responsive to them, and how they could build a relationship with a virtual coach. Of these, many wanted more information about whether there was a real therapist involved and how involved they would be (11/47, 23%), whether they would be able to meet them in person, whether they would deliver the therapy live or preprogram the coach, or whether the virtual coach would be completely artificially intelligent. The realness of the virtual coach and the VRT (16/226, 7.1%) was also an important question posed by participants, including asking whether it would be realistic enough and comparing it to "real" or face-to-face therapy.

Most additional responses indicated that participants would like to know more about the process (67/226, 29.6%). This included practical questions regarding the frequency of use, the length of sessions, and how it would be delivered (eg, in which location). Furthermore, many participants asked about the content (26/226, 11.5%), particularly not only the scenarios that could be represented but also other aspects including the appearance, the script, and how the content could link with face-to-face therapy. Of the 226 participants, 11 (4.9%) asked about the technical aspects including the development of the coach (eg, whether an algorithm or artificial intelligence was used) and what equipment would be used to deliver the VRT.

How Has Your Previous Experience of VR (Minimal or Extensive) Influenced Your Perceptions of VRT Guided by a Virtual Coach?

The largest group of those answering this question indicated that they had no previous experience (79/236, 33.5%). A few without experience were positive or curious (19/236, 8.1%) whereas others (8/236, 3.1%) expressed more negative perceptions about its effectiveness as a therapeutic tool, anticipating that it would not feel real or tailored enough to the individual. The second largest group (59/236, 25%) felt that their previous experience had helped them to be more positive and linked it to their own experience of mental health and how it could be used for treatment. Although several mentioned using VR for gaming, they felt that it was effective at producing a level of presence that would be conducive to therapy and help to invoke real emotions and responses. They felt it was easy to use, could potentially lower costs, make therapy more accessible, and even with negative experiences, such as motion sickness or technical difficulties, they still had a positive perception of VRT.

However, 34 (14.4%) of the 236 participants with a more negative perception reported nausea or dizziness, whereas others perceived VR as more suited to games. This included problems with the quality of their experience, feeling that the VRT had not offered enough presence. However, those with a negative experience comprised the smallest group (8/236, 3.1%). Finally, the third largest group felt that their previous experience would not influence how they felt about VRTs guided by a virtual coach (40/236, 16.9%). For some, their previous experiences could not inform their perception of VRTs as it had been for entertainment purposes or they had too little experience to be able to make a judgment (13/236, 5.5%).

If You Were Offered VRT Guided by a Virtual Coach, Where Do You Think the Best Place to Do the Therapy Would Be?

The largest group of respondents who identified a single location felt that it would be best delivered within the home (84/245, 34.3%), whereas the second largest group felt that it would be best delivered in a more professional location (43/245, 17.6%). Several felt that it could be offered in both settings (38/245, 15.5%), whereas others suggested that access could first be through a clinic (12/245, 4.9%), where they could access technical or therapeutic support, or from home (4/245, 1.6%), where they would feel more comfortable. When respondents highlighted delivery from home, they described it as being safe, comfortable, and familiar. They felt that they might feel susceptible or disorientated when coming out of a VRT session and that being at home would be preferable. More professional locations, such as physician surgeries or clinics, were also described as safe and familiar, although by fewer people. Professional or clinical settings were often viewed as a better location because of the presence of support. Other reasons included the level of cleanliness offered and that there would be fewer distractions.

Those without a preference identified elements of the location that were necessary to optimize the experience, including having a space to move, feeling safe and secure (eg, in an enclosed space), having privacy and quiet, and having few distractions. They also felt it would need to consider the condition being treated (including severity) and the individual's preferences.

Factors With Item Responses

Table 3 presents those items where either strong agreement or disagreement elicited a qualitative response.

Table 3. Qualitative responses to items.

Question	Strongly agree, n (%)	Strongly disagree, n (%)
Factor 1: attitudes toward VRT[a] (42 and 43)		
Item 41: If the virtual coach encouraged me to do something between sessions, I would try to do it. (n=19)	18 (95)	1 (5)
Item 51: I would never be willing to try virtual reality therapy. (n=63)	3 (5)	60 (95)
Item 50: I would be willing to try virtual reality therapy if I had more information about it. (n=24)	21 (87)	3 (13)
Item 52: I would encourage the people I care about to try virtual reality therapy, if it was offered to them. (n=16)	14 (87)	2 (13)
Item 53: I would discourage the people I care about to try virtual reality therapy, if it was offered to them. (n=25)	1 (4)	24 (96)
Item 54: I cannot imagine virtual reality therapy being useful for someone with mental health problems. (n=24)	2 (8)	22 (92)
Item 42: I would feel comfortable interacting with the virtual coach. (n=14)	10 (71)	4 (29)
Factor 2: expectation of presence (47, 49, 50, 52, and 40)		
Item 43: I would find the characters in the virtual reality therapy unsettling. (n=6)	1 (17)	5 (83)
Item 45: I am skeptical about the effectiveness of virtual reality therapy. (n=13)	6 (46)	7 (54)
Item 44: I think that the virtual reality therapy would make me feel present enough to be effective. (n=7)	3 (43)	4 (57)
Factor 3: preference for VRT (54, 39, 53, and 41)		
Item 39: I think virtual reality therapy would be better than face-to-face therapy. (n=29)	2 (7)	27 (93)
Item 40: I would trust a virtual coach the same amount as a real therapist. (n=20)	4 (20)	16 (80)
Factor 4: cost-effectiveness (46)		
Item 46: I think virtual reality therapy will be worth the cost. (n=9)	6 (66)	3 (34)
Nonfactor answers (45, 48, 44, and 51)		
Item 47: I think I would be able to use the virtual reality equipment easily. (n=26)	24 (92)	2 (8)
Item 49: I think that virtual reality equipment could spread diseases. (n=19)	0 (0)	19 (100)
Item 48: If my skills with technology were poor, I would feel confident using virtual reality therapy if the health care professional accompanying me was trained to a high standard. (n=19)	17 (89)	2 (11)

[a]VRT: virtual reality therapy.

Factor 1: Attitude Toward VRT

Individuals who scored highly on this factor had a positive attitude toward VRTs and VRTs delivered by a virtual coach, whereas those who scored low had a negative attitude. Within the items where strong agreement or disagreement elicited a text response (items 41, 51, 50, 52, 53, 54, and 42), those with positive attitudes highlighted the value of having a choice in mental health therapies. They emphasized the need to be willing to try different treatments to find the one that worked, reflecting on how more automated and digital options can help to increase access. Those with more negative attitudes indicated that it would be a type of therapy that they would not choose.

Factor 2: Expectation of Presence

Individuals who scored highly on this factor felt that VR would not be real, that is, low presence. Individuals who scored low felt that VR was immersive. Within the items where strong agreement or disagreement elicited a text response (items 43, 45, and 44), the respondents indicated several factors that affected their expectation that VR would be "real enough." Previous experiences appeared to be linked to the expectations of presence. People who enjoyed their experiences had higher expectations of presence. Those with lower expectations felt that VRT would be too much like a game, whereas others indicated that experiencing cybersickness meant they had not felt present.

Factor 3: Preference for VRT

Individuals who scored highly on this factor showed a preference for VRT, whereas those who scored low showed a preference for face-to-face therapy. Within the items where strong agreement or disagreement elicited a text response (items 39 and 40), there was a strong sense that those who preferred face-to-face therapy would feel the loss of human interaction most and feel that a real person was needed to build a relationship and trust. Those who were more in favor of VRT and the virtual coach felt that it would be more convenient and potentially enable more disclosures related to their mental health.

Factor 4: Cost-Effectiveness

Individuals who scored high on this factor felt that VRT was cost-effective, whereas those who scored low did not. Within the item where strong agreement or disagreement elicited a text response (item 46), those who felt it was cost-effective highlighted the decreasing costs of equipment and the benefits this could bring to mental health services. For those who felt that VR was still too expensive, there was also recognition of the difficulties that services might have in adopting VR.

Discussion

Overview

This study aimed to develop a new instrument for assessing the public perception of VRT delivered by a virtual coach. We received 295 responses. We found that a 4-factor solution was the best fit for the AVRT Scale, with all subscales having excellent internal consistency. The 4 factors were (1) attitudes toward VRT, (2) expectation of presence, (3) preference for VRT, and (4) cost-effectiveness. We found that being more familiar with VR was correlated with more positive attitudes toward VRT (factor 1), a higher expectation of presence (factor 2), preference for VRT over face-to-face therapy (factor 3), and belief that VRT is cost-effective (factor 4). Familiarity with mental health was not associated with any factor. The qualitative data supported the quantitative findings, with many respondents stating that their previous experience with VR may have affected their perception of VRT. Respondents identified their homes and spaces that felt safe and quiet as the best locations for delivering VRT. The virtual coach was a salient concept throughout the qualitative responses, with participants wanting to better understand it and the relationships that could be facilitated.

Principal Findings

Previous literature has indicated a correlation between VR familiarity and more positive attitudes toward VRTs [14]; however, this is the first study to demonstrate this through a public survey. Although we do not know the direction of this association, the qualitative findings suggest that as the reach of VR headsets increases, VRTs will likely be viewed more positively. The perceptions of potential patients are important in determining the efficacy of VRT, as positive expectations of any psychological therapy are associated with better treatment outcomes [35,36]. Therefore, an increase in the popularity of VR kits may indirectly improve the efficacy of VRTs.

This study is the first to explore peoples' perceptions of VRTs guided by a virtual coach. Although most participants had no personal or professional experience of mental health therapy, many mentioned aspects relating to the virtual coach that draw parallels with "therapeutic alliance"; in psychotherapy, this denotes the importance of the relationship between the therapist and service user. In psychotherapy research, a strong therapeutic alliance is associated with better treatment outcomes [37]. The concept of therapeutic alliance has been studied more broadly in relation to VR-assisted therapies [30] and digital mental health [38]. The effects are similar but may be predictive of treatment outcomes to a lesser extent and more so predict engagement.

Understanding whether it is possible to foster a "therapeutic alliance" with a virtual coach and, if so, the nature of this relationship is something that the public is concerned with and therefore requires further investigation. Our findings provide initial insights into how therapeutic alliances may operate in VRT with a virtual coach. Respondents who showed a preference for VRTs indicated that the presence of a virtual coach would aid disclosure. This may be because of the anonymity that this form of communication offers [39]. Furthermore, it is notable that many were curious about the level of automation and formulation offered by the virtual coach. Qualitative findings from a trial of VRT guided by a virtual coach found that the presence of a member of staff helped to reinforce learning, which may suggest that certain elements of therapy require a certain level of formulation [40]. Our data suggest that the public view personalization as an important component of therapy and that VRTs can be improved by offering a certain level of formulation.

Another novelty of this study is the exploration of presence in relation to VRTs. Previous research has suggested that increasing presence can increase the effectiveness of VRT [41]. Newer VR-enabled headsets provide a greater level of presence as the quality of graphics and functionality, such as interactivity and sensitivity of sensors (eg, eye tracking), have improved significantly. Therefore, we sought to explore the importance of this sense of presence in the general population. Our findings indicate that those with a greater expectation of presence are more positive and more likely to show a preference for VRTs. Notably, our findings indicate that even those who are familiar with VR share concerns regarding the lack of presence and immersion in VRTs. This suggests that developers and researchers must continue to develop and update their intervention designs to ensure that VRTs do not become stagnant and continue to elicit a sense of presence.

Most of those who viewed VRTs positively emphasized the need for a choice to help increase access to mental health treatment. More automated VRTs have been designed considering the pressure to deliver psychological therapies in mental health services and the lack of resources to meet this need [10,11]. Our findings indicate that the public is aware of this and views VRTs guided by a virtual coach as an acceptable solution. Our respondents also indicated that VRT guided by a virtual coach would be suitable for delivery at home, further alleviating resource pressures. However, this was not the case for all participants, with a notable proportion wanting to access VRTs in a location that was safe, familiar, free of distractions, and large enough to use the VRT. The flexibility of location in delivering VRT is an important consideration for increasing access and meeting the needs of service users, especially when considering the strong links between poor mental health and housing quality [42]. The delivery location for VRT should be considered on a case-by-case basis.

A further consideration for the implementation of VRT found in our study is the importance of information. Our qualitative findings indicate that people were keen to better understand what was involved in VRT and the virtual coach. Preintervention expectations are key in managing service users' expectations while also fostering hopefulness, which in turn improves engagement [43]. This information can also help to allay any concerns service users might have about VRT and help developers to understand their needs to improve the design and implementation of VRTs. For example, a small number of participants opposed the use of VRTs, which when expanded in our qualitative data collection, indicated certain ethical and moral concerns about its use in mental health care. All these views were valid, but a few may be rooted in misconceptions about VR or expectations about how it will be implemented. Therefore, potentially increasing the public's awareness and understanding of VR and VRT may help to appease them and improve how it is deployed.

We found mixed findings regarding the impact of cybersickness on willingness to engage in VRT. For some participants, cybersickness would dissuade people from using VRT. However, this finding was not ubiquitous, with some saying that they were still interested in trying VRT even if they had experienced cybersickness. This contradicts previous research, suggesting cybersickness is a considerable barrier [17]. As technology progresses, cybersickness might become less important. We also included a question on hygiene, as our questionnaire was shared during the COVID-19 pandemic. However, this was excluded, suggesting that it was not a significant concern for the public.

Limitations

The questionnaire has only been validated using the EFA. Further validation is required before we can confidently recommend its widespread use. Specifically, we must confirm the factor structure in a new sample using confirmatory factor analysis and assess its concurrent and discriminant validity. If we are able to replicate the strong psychometric properties found in this study, this questionnaire can be used to understand attitudes toward VRTs delivered by virtual coaches. The scale will also need to be adapted to contexts outside the United Kingdom, for example, by amending items and further validation.

Most of our respondents were female and had no previous experience with mental health conditions or therapy. However, men and those with more experience with mental health conditions or therapy may have different perceptions of VRTs. A recent review found that gender differences might affect the use and acceptability of VR, specifically that women are more susceptible to cybersickness and therefore may be less willing to use VR [44]. On the basis of this, it may be assumed that if we were to conduct a survey with male participants, the attitudes toward VRT guided by a virtual coach could be even more positive. We do not have any available evidence to indicate whether those who are living with or have lived with mental health conditions are likely to be more or less accepting of VRTs. Purposive sampling should be used in future studies to ensure that the views of these groups are included in future validation studies.

Furthermore, we sought text responses for strong agreement or disagreement with certain items. Notably, those with more neutral views may have offered additional insights, but we weighed this against the additional burden on respondents. This may also have led some participants to neutralize their views to avoid triggering a free-text question. However, there were no instructions in the questionnaire that responding differently removed the free-text responses. In addition, the range of scores indicated that this did not deter the participants from giving extreme answers. The qualitative data we captured were sufficient for our analysis.

Finally, the analysis of the relationship between familiarity and attitudes toward VR and VRT was correlational. We could not make any claim regarding the direction or causal nature of these associations. For example, those with more positive attitudes and a better understanding are more likely to become familiar with VR through continued use. However, our qualitative findings indicate that negative experiences with VR do not factor in a willingness to use VRT.

Recommendations

Future research should further validate this questionnaire. Once this has been accomplished, the questionnaire could be used to investigate the factors that improve or worsen attitudes toward VRT and VRT guided by a virtual coach. For example, asking questions such as whether trying VR improves perceptions or whether increasing sales of domestic VR kits is associated with improved attitudes. It is also important to explore how these attitudes translate into behavior, that is, whether positive attitudes predict patient preferences and engagement with VRTs. The impact of the level of automation versus the formulation of the virtual coach on attitudes should also be explored, as this was a salient concept within the qualitative data. The AVRT Scale could be adapted and applied to other areas where VR is used to deliver interventions, such as behavior change interventions, or as a training tool. The questionnaire can also be used alongside treatment development, evaluation, and implementation to explore the barriers and facilitators specific to VRT and VRT guided by a virtual coach or the perceptions of certain populations to aid the translation of research into practice [45]. In the long term, any research that considers barriers to the uptake, engagement, and adoption of VRT has the potential to alleviate the demand for trained therapists in clinical settings, thus improving access to psychological therapies.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. ADGB was funded by the National Institute for Health and Care Research MindTech MedTech Co-operative. The views expressed are those of the authors and not necessarily those of the National Health Service, the National Institute for Health and Care Research, or the Department of Health. These organizations had no role in the study design, including collection, management, analysis, and interpretation of data, writing of the report, or the decision to submit the report for publication.

ADGB would like to acknowledge the support of the National Institute for Health and Care Research Nottingham Biomedical Research Centre.

Data Availability

All relevant data have been included in this publication. Researchers who would like to access the scale for further validation or adaptation may contact the corresponding author with a methodologically sound proposal.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Description of virtual reality therapy and the virtual coach.
[PDF File (Adobe PDF File), 18 KB - mental_v11i1e48537_app1.pdf]

Multimedia Appendix 2
Factor structure and loadings.
[PDF File (Adobe PDF File), 172 KB - mental_v11i1e48537_app2.pdf]

References

1. Lee KM. Presence, explicated. Commun Theory 2004 Feb;14(1):27-50. [doi: 10.1111/j.1468-2885.2004.tb00302.x]
2. Kilkelly F, O'Brien R, Ticho S. The growing value of XR in healthcare in the United Kingdom. The XR Health Alliance. 2021. URL: https://www.xrhealthuk.org/the-growing-value-of-xr-in-healthcare [accessed 2023-12-16]
3. Eshuis LV, van Gelderen MJ, van Zuiden M, Nijdam MJ, Vermetten E, Olff M, et al. Efficacy of immersive PTSD treatments: a systematic review of virtual and augmented reality exposure therapy and a meta-analysis of virtual reality exposure therapy. J Psychiatr Res 2021 Nov;143:516-527 [FREE Full text] [doi: 10.1016/j.jpsychires.2020.11.030] [Medline: 33248674]
4. Falconer CJ, Rovira A, King JA, Gilbert P, Antley A, Fearon P, et al. Embodying self-compassion within virtual reality and its effects on patients with depression. BJPsych Open 2016 Jan;2(1):74-80 [FREE Full text] [doi: 10.1192/bjpo.bp.115.002147] [Medline: 27703757]
5. Ferrer-García M, Gutiérrez-Maldonado J. The use of virtual reality in the study, assessment, and treatment of body image in eating disorders and nonclinical samples: a review of the literature. Body Image 2012 Jan;9(1):1-11. [doi: 10.1016/j.bodyim.2011.10.001] [Medline: 22119329]
6. Morina N, Ijntema H, Meyerbröker K, Emmelkamp PM. Can virtual reality exposure therapy gains be generalized to real-life? A meta-analysis of studies applying behavioral assessments. Behav Res Ther 2015 Nov;74:18-24. [doi: 10.1016/j.brat.2015.08.010] [Medline: 26355646]
7. Opriş D, Pintea S, García-Palacios A, Botella C, Szamosközi Ş, David D. Virtual reality exposure therapy in anxiety disorders: a quantitative meta-analysis. Depress Anxiety 2012 Feb;29(2):85-93. [doi: 10.1002/da.20910] [Medline: 22065564]
8. Valmaggia L. The use of virtual reality in psychosis research and treatment. World Psychiatry 2017 Oct;16(3):246-247 [FREE Full text] [doi: 10.1002/wps.20443] [Medline: 28941114]
9. Hayward M. Increasing access to psychological treatments for mental illness. Lancet Psychiatry 2018 Aug;5(8):606-607 [FREE Full text] [doi: 10.1016/S2215-0366(18)30262-1] [Medline: 30007518]
10. Freeman D, Lambe S, Kabir T, Petit A, Rosebrock L, Yu LM, gameChange Trial Group. Automated virtual reality therapy to treat agoraphobic avoidance and distress in patients with psychosis (gameChange): a multicentre, parallel-group, single-blind, randomised, controlled trial in England with mediation and moderation analyses. Lancet Psychiatry 2022 May;9(5):375-388 [FREE Full text] [doi: 10.1016/S2215-0366(22)00060-8] [Medline: 35395204]
11. Freeman D, Haselton P, Freeman J, Spanlang B, Kishore S, Albery E, et al. Automated psychological therapy using immersive virtual reality for treatment of fear of heights: a single-blind, parallel-group, randomised controlled trial. Lancet Psychiatry 2018 Aug;5(8):625-632 [FREE Full text] [doi: 10.1016/S2215-0366(18)30226-8] [Medline: 30007519]
12. Rizzo A, Shilling R, Forbell E, Scherer S, Gratch J, Morency LP. Autonomous virtual human agents for healthcare information support and clinical interviewing. In: Luxton DD, editor. Artificial Intelligence in Behavioral and Mental Health Care. Amsterdam, Netherlands: Elsevier Academic Press; 2016:53-79.
13. Altunkaya J, Craven M, Lambe S, Beckley A, Rosebrock L, Dudley R, et al. Estimating the economic value of automated virtual reality cognitive therapy for treating agoraphobic avoidance in patients with psychosis: findings from the gamechange

randomized controlled clinical trial. J Med Internet Res 2022 Nov 18;24(11):e39248 [FREE Full text] [doi: 10.2196/39248] [Medline: 36399379]
14. Lindner P, Miloff A, Zetterlund E, Reuterskiöld L, Andersson G, Carlbring P. Attitudes toward and familiarity with virtual reality therapy among practicing cognitive behavior therapists: a cross-sectional survey study in the era of consumer VR platforms. Front Psychol 2019 Feb 8;10:176 [FREE Full text] [doi: 10.3389/fpsyg.2019.00176] [Medline: 30800086]
15. Chung OS, Johnson AM, Dowling NL, Robinson T, Ng CH, Yücel M, et al. Are Australian mental health services ready for therapeutic virtual reality? An investigation of knowledge, attitudes, implementation barriers and enablers. Front Psychiatry 2022 Feb 4;13:792663 [FREE Full text] [doi: 10.3389/fpsyt.2022.792663] [Medline: 35185649]
16. Benbow AA, Anderson PL. A meta-analytic examination of attrition in virtual reality exposure therapy for anxiety disorders. J Anxiety Disord 2019 Jan;61:18-26. [doi: 10.1016/j.janxdis.2018.06.006] [Medline: 30646997]
17. Keller MS, Park HJ, Cunningham ME, Fouladian JE, Chen M, Spiegel BM. Public perceptions regarding use of virtual reality in health care: a social media content analysis using Facebook. J Med Internet Res 2017 Dec 19;19(12):e419 [FREE Full text] [doi: 10.2196/jmir.7467] [Medline: 29258975]
18. Brown P, Waite F, Lambe S, Jones J, Jenner L, Diamond R, et al. Automated virtual reality cognitive therapy (gameChange) in inpatient psychiatric wards: qualitative study of staff and patient views using an implementation framework. JMIR Form Res 2022 Apr 12;6(4):e34225 [FREE Full text] [doi: 10.2196/34225] [Medline: 35412462]
19. Chung OS, Robinson T, Johnson AM, Dowling NL, Ng CH, Yücel M, et al. Implementation of therapeutic virtual reality into psychiatric care: clinicians' and service managers' perspectives. Front Psychiatry 2021 Jan 4;12:791123 [FREE Full text] [doi: 10.3389/fpsyt.2021.791123] [Medline: 35058823]
20. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, et al. Barriers to and facilitators of user engagement with digital mental health interventions: systematic review. J Med Internet Res 2021 Mar 24;23(3):e24387 [FREE Full text] [doi: 10.2196/24387] [Medline: 33759801]
21. JISC: online surveys. JISC. 2022. URL: https://www.jisc.ac.uk [accessed 2023-12-14]
22. Field A. Discovering Statistics Using IBM SPSS Statistics. Thousand Oaks, CA: Sage Publications; 2017.
23. Aref-Adib G, McCloud T, Ross J, O'Hanlon P, Appleton V, Rowe S, et al. Factors affecting implementation of digital health interventions for people with psychosis or bipolar disorder, and their family and friends: a systematic review. Lancet Psychiatry 2019 Mar;6(3):257-266. [doi: 10.1016/S2215-0366(18)30302-X] [Medline: 30522979]
24. Glegg SM, Levac DE. Barriers, facilitators and interventions to support virtual reality implementation in rehabilitation: a scoping review. PM R 2018 Nov;10(11):1237-51.e1 [FREE Full text] [doi: 10.1016/j.pmrj.2018.07.004] [Medline: 30503231]
25. March S, Day J, Ritchie G, Rowe A, Gough J, Hall T, et al. Attitudes toward e-mental health services in a community sample of adults: online survey. J Med Internet Res 2018 Feb 19;20(2):e59 [FREE Full text] [doi: 10.2196/jmir.9109] [Medline: 29459357]
26. Nichols S, Patel H. Health and safety implications of virtual reality: a review of empirical evidence. Appl Ergon 2002 May;33(3):251-271. [doi: 10.1016/s0003-6870(02)00020-0] [Medline: 12164509]
27. Schwartzman D, Segal R, Drapeau M. Perceptions of virtual reality among therapists who do not apply this technology in clinical practice. Psychol Serv 2012 Aug;9(3):310-315. [doi: 10.1037/a0026801] [Medline: 22867123]
28. Tinwell A, Nabi DA, Charlton JP. Perception of psychopathy and the uncanny valley in virtual characters. Comput Human Behav 2013 Jul;29(4):1617-1625. [doi: 10.1016/j.chb.2013.01.008]
29. Wrzesien M, Burkhardt JM, Botella C, Alcañiz M. Towards a virtual reality- and augmented reality-mediated therapeutic process model: a theoretical revision of clinical issues and HCI issues. Theor Issues Ergon Sci 2014 May 28;16(2):124-153 [FREE Full text] [doi: 10.1080/1463922x.2014.903307]
30. Miragall M, Baños RM, Cebolla A, Botella C. Working alliance inventory applied to virtual and augmented reality (WAI-VAR): psychometrics and therapeutic outcomes. Front Psychol 2015 Oct 08;6:1531 [FREE Full text] [doi: 10.3389/fpsyg.2015.01531] [Medline: 26500589]
31. IBM SPSS Statistics version 25.0. 2017. IBM Corp. 2017. URL: https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-25 [accessed 2023-12-14]
32. Kaiser HF. An index of factorial simplicity. Psychometrika 1974 Mar;39(1):31-36 [FREE Full text] [doi: 10.1007/bf02291575]
33. George D, Mallery P. SPSS for Windows Step by Step: A Simple Guide and Reference, 11.0 Update. Boston, MA: Allyn & Bacon; 2003.
34. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov 01;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
35. Gergov V, Lindberg N, Lahti J, Lipsanen J, Marttunen M. Effectiveness and predictors of outcome for psychotherapeutic interventions in clinical settings among adolescents. Front Psychol 2021 Feb 16;12:628977 [FREE Full text] [doi: 10.3389/fpsyg.2021.628977] [Medline: 33664698]
36. Karlsson H, Kronström K, Nabi H, Oksanen T, Salo P, Virtanen M, et al. Low level of optimism predicts initiation of psychotherapy for depression: results from the Finnish Public Sector Study. Psychother Psychosom 2011 Apr 18;80(4):238-244. [doi: 10.1159/000323613] [Medline: 21502771]
37. Falkenström F, Granström F, Holmqvist R. Therapeutic alliance predicts symptomatic improvement session by session. J Couns Psychol 2013 Jul;60(3):317-328. [doi: 10.1037/a0032258] [Medline: 23506511]

38. Tremain H, McEnery C, Fletcher K, Murray G. The therapeutic alliance in digital mental health interventions for serious mental illnesses: narrative review. JMIR Ment Health 2020 Aug 07;7(8):e17204 [FREE Full text] [doi: 10.2196/17204] [Medline: 32763881]
39. Clark-Gordon CV, Bowman ND, Goodboy AK, Wright A. Anonymity and online self-disclosure: a meta-analysis. Commun Rep 2019 May 07;32(2):98-111 [FREE Full text] [doi: 10.1080/08934215.2019.1607516]
40. Bond J, Kenny A, Pinfold V, Couperthwaite L, gameChange Lived Experience Advisory Panel, Kabir T, et al. A safe place to learn: peer research qualitative investigation of gameChange virtual reality therapy. JMIR Serious Games 2023 Jan 16;11:e38065 [FREE Full text] [doi: 10.2196/38065] [Medline: 36645707]
41. Price M, Anderson P. The role of presence in virtual reality exposure therapy. J Anxiety Disord 2007 Jan;21(5):742-751 [FREE Full text] [doi: 10.1016/j.janxdis.2006.11.002] [Medline: 17145164]
42. Evans GW, Wells NM, Chan HY, Saltzman H. Housing quality and mental health. Artic J Consult Clin Psychol 2000;68(3):526-530 [FREE Full text] [doi: 10.1037/0022-006X.68.3.526]
43. Bennett-Levy J, Richards D, Farrandds P, Christensen H, Griffiths K, Kavanagh D, et al. Oxford Guide to Low Intensity CBT Interventions. Oxford, UK: Oxford University Press; 2010.
44. Grassini S, Laumann K. Are modern head-mounted displays sexist? A systematic review on gender differences in HMD-mediated virtual reality. Front Psychol 2020 Aug 7;11:1604 [FREE Full text] [doi: 10.3389/fpsyg.2020.01604] [Medline: 32903791]
45. Andleeb H, Bergin AG, Robotham D, Brown S, Martin J. Relevant involvement and how it can improve implementation for novel digital mental health interventions. In: Proceedings of the 3rd UK Implementation Science Research Conference. 2020 Presented at: UKISRC '20; July 16-17, 2020; Virtual Event URL: https://www.researchgate.net/publication/343262737_Relevant_involvement_and_how_it_can_improve_implementation_for_novel_digital_mental_health_interventions [doi: 10.13140/RG.2.2.30467.68642]

Abbreviations

AVRT: Attitudes Towards Virtual Reality Therapy
EFA: exploratory factor analysis
VR: virtual reality
VRT: virtual reality therapy

Edited by J Torous; submitted 27.04.23; peer-reviewed by N Mungoli, T Cahill; comments to author 16.07.23; revised version received 29.08.23; accepted 21.10.23; published 12.01.24.

Please cite as:
Gomez Bergin AD, Allison AM, Hazell CM
Understanding Public Perceptions of Virtual Reality Psychological Therapy Using the Attitudes Towards Virtual Reality Therapy (AVRT) Scale: Mixed Methods Development Study
JMIR Ment Health 2024;11:e48537
URL: https://mental.jmir.org/2024/1/e48537
doi:10.2196/48537
PMID:38214958

Original Paper

Personalized Virtual Reality Compared With Guided Imagery for Enhancing the Impact of Progressive Muscle Relaxation Training: Pilot Randomized Controlled Trial

Susanna Pardini[1,2,3], MSc, PsyD; Silvia Gabrielli[2], PhD; Silvia Olivetto[1], MSc, PsyD; Francesca Fusina[1,4], PhD, PsyD; Marco Dianti[2], MSc; Stefano Forti[5], MSc; Cristina Lancini[1], MSc, ClinPsy; Caterina Novara[1], PhD

[1]Department of General Psychology, University of Padova, Padova, Italy

[2]Digital Health Research, Centre for Digital Health and Wellbeing, Fondazione Bruno Kessler, Trento, Italy

[3]Human Inspired Technology Research Centre (HIT), University of Padova, Padova, Italy

[4]Padova Neuroscience Center, University of Padova, Padova, Italy

[5]Centre for Digital Health and Wellbeing, Fondazione Bruno Kessler, Trento, Italy

Corresponding Author:
Susanna Pardini, MSc, PsyD
Department of General Psychology
University of Padova
8 Venezia St
Padova
Italy
Phone: 39 3335944315
Email: susanna.pardini@phd.unipd.it

Abstract

Background: Empirical evidence has shown that virtual reality (VR) scenarios can increase the effects of relaxation techniques, reducing anxiety by enabling people to experience emotional conditions in more vivid settings.

Objective: This pilot randomized controlled study aims to investigate whether the progressive muscle relaxation technique (PMRT) associated with a personalized scenario in VR promotes psychological well-being and facilitates the recall of relaxing images more than the standard complementary intervention that involves the integration of PMRT and guided imagery (GI).

Methods: On the basis of a longitudinal, between-subject design, 72 university students were randomly exposed to one of two experimental conditions: (1) standard complementary procedure (PMRT and GI exposure) and (2) experimental procedure (PMRT and personalized VR exposure). Individuals were assessed by a therapist before and after 7 training sessions based on measures investigating anxiety, depression, quality of life, coping strategies, sense of presence, engagement, and side effects related to VR exposure. Heart rate data were also collected.

Results: Differences in changes between the 2 groups after the in vivo PMRT session conducted by the psychotherapist (T1) were statistically significant for state anxiety ($F_{1,67}$=30.56; P<.001) and heart rate ($F_{1,67}$=4.87; P=.01). Individuals in the VR group obtained lower scores both before (t_{67}=−2.63; P=.01; Cohen d=0.91) and after (t_{67}=−7.23; P<.001; Cohen d=2.45) the relaxation session when it was self-administered by participants (T2). A significant reduction in perceived state anxiety at T1 and T2 was observed for both groups (P<.001). After the VR experience, individuals reported feeling higher engagement in the experience than what was mentioned by participants in the GI group ($F_{1,67}$=2.85; P=.03; η_p^2=0.15), and they experienced the environment as more realistic ($F_{1,67}$=4.38; P=.003; η_p^2=0.21). No differences between groups regarding sense of presence were found ($F_{1,67}$=1.99; P=.11; η_p^2=0.11). Individuals exposed before to the VR scenario (T1) referred to perceiving the scenario recalled in-imagination at T2 as more realistic than what those in the GI group experienced ($F_{1,67}$=3.21; P=.02; η_p^2=0.12). The VR group had lower trait anxiety levels than the GI group after the relaxation session during session 7 (T2; t_{67}=−2.43; P=.02).

Conclusions: Personalized relaxing VR scenarios can contribute to improving relaxation and decreasing anxiety when integrated with PMRT as a complementary relaxation method.

Trial Registration: ClinicalTrials.gov NCT05478941; https://classic.clinicaltrials.gov/ct2/show/NCT05478941

International Registered Report Identifier (IRRID): RR2-10.2196/44183

(JMIR Ment Health 2024;11:e48649) doi:10.2196/48649

KEYWORDS

digital health; progressive muscular relaxation technique; mental well-being; virtual reality therapy; anxiety; relaxation; e-therapy; eHealth; mobile phone

Introduction

Background

In working adults and university students, the prevalence of depression, anxiety, and stress increases, predisposing them to physical diseases and general repercussions on well-being [1,2].

The situation has worsened owing to the COVID-19 outbreak, which affected individuals' general well-being worldwide [3,4]. Different types of standardized relaxation interventions (eg, mindfulness-based stress reduction, progressive relaxation techniques by Jacobson, autogenic training by Schultz, abdominal relaxations, and visualizations) before, during, and after the COVID-19 pandemic have been shown to reduce anxiety and stress symptomatology in university students [4,5], with a more significant effect when integrated with other complementary techniques such as guided imagery (GI) [6,7]. GI is useful in creating mental imagery and refocusing attention on pleasant and relaxing imagined visual, auditory, tactile, or olfactory sensations, resulting in specific psychological and physiological responses such as relaxation and reduction of the autonomic nervous system responses [6,8,9].

The integration of the progressive muscle relaxation technique (PMRT) and GI can promote a higher sense of relaxation during training sessions, allowing for the application of the complementary techniques to cope with the stress experienced in daily activities [6,10-14].

Virtual reality (VR) is a usable, engaging, and user-friendly technology that promotes full immersion in the virtual context and facilitates the control of disturbing external stimuli [15,16].

Various studies have demonstrated that being immersed in virtual natural environments through a head-mounted display (HMD) facilitates the reduction of anxiety and stress symptoms in college students [17] as well as in association with different types of relaxation training (eg, body scan and muscle relaxation training) [18-21].

Owing to the positive impact of GI with PMRT in promoting relaxation, it may be hypothesized that exposure to a more vivid, closer-to-reality VR experience could be a helpful strategy for improving the relaxation learning promoted by the PMRT and decreasing stress and anxiety symptomatology.

GI offers an undefined number of possibilities for personalizing the imagined scenarios in a safe context. The personalization of content in VR should be helpful in recreating situations close to users' needs to promote relaxation and can override the limits of GI. The personalization of VR environments is a crucial element to consider as it allows for relaxation and the perception of safety in the virtual context [18,22] owing to the possibility of offering a more realistic emotional experience, reflecting the users' needs [23].

Furthermore, the first part of the PMRT, named "active PMRT," involves a series of active sessions in which people learn how to tense and release the different muscle groups from the bottom to the top of the body to recognize the subjective state of muscle relaxation and relax the muscle areas of the body, reducing tension interfering with skeletal muscle activity.

These previous active PMRT phases are essential and are typically administered face-to-face by a psychotherapist or health operator trained to conduct the PMRT. By considering the effective results of web-based interventions in reducing stress among college students [24] and the relevant advantages that web-based therapy can offer (eg, saving costs related to attending psychotherapy sessions, allowing for partial self-management of the relaxation training), we assumed that assessing the efficacy of PMRT in alternative settings may facilitate the administration of treatment when the implementation of the standard procedure is not possible.

In light of the evidence described, we hypothesized that VR is more effective than in-imagination exposure in allowing for relaxation and decreasing state anxiety because of a more realistic sensory experience, facilitating visualization.

Objectives

For this reason, our primary aim was to deploy the active PMRT sessions remotely via Zoom (Zoom Video Communications) and the last complementary session in the therapist's presence, exposing people to a passive progressive relaxation session with GI or VR.

More in detail, we investigated whether PMRT associated with a personalized relaxing scenario in VR can be more effective in reducing state anxiety, tension or activation, and heart rate.

The secondary purposes of this research pilot were to (1) understand whether VR promotes a better sense of presence and engagement in the scenario compared with GI after session 6 (T1) and whether it helps recall the image and be immersed in the relaxing scenario in session 7 (T2) and (2) investigate whether exposure to the VR scenario promotes a better perception of psychological well-being, stress, and trait anxiety symptoms.

Methods

Overview

This pilot study was a randomized, parallel-assignment, open-label, controlled, single-center trial based on a longitudinal, between-subject design conducted at the University of Padova (Italy) in collaboration with the Center for Digital Health and

Wellbeing–Fondazione Bruno Kessler (Italy). This study is part of a larger research protocol published by Pardini et al [25] (International Registered Report Identifier: RR2-10.2196/44183). The study's recruitment phase ended in February 2023.

Participants

Before study enrollment, during the T0 face-to-face assessment phase at the university, the participants signed a written informed consent form based on a paper-and-pencil form agreeing to participate in all the study sessions. They were informed that (1) their data would be confidential, (2) they could omit any information they did not wish to provide, and (3) they could withdraw from the study without explanation.

Study participants were recruited in Northeast Italy via social networking websites (on the web) and during university lectures (offline). Those individuals interested in participating were asked to attend a face-to-face assessment with the investigators to participate in the first evaluation phase (T0) to comply with the inclusion and exclusion criteria, provide written informed consent, and undergo a baseline evaluation.

Eligible participants were adults from the general population (aged ≥18 years) and native Italian speakers, owned a PC, and were able to use a PC and a smartphone.

Participants were excluded from the study if they had been diagnosed with a severe mental disorder or medical conditions (eg, neuromuscular disorders) or were undergoing current psychotherapeutic treatment.

Eligible participants were randomly allocated to one of the experimental conditions based on a simple blinded randomization via an Excel (Microsoft Corp) file using the "RAND" function. In total, 2 experimenters trained in cognitive and behavioral therapy conducted the relaxation sessions in a balanced way, each administering the sessions to 50% of the sample of each group to control for possible biases related to the therapist's personality and competencies. During the allocation to one of the experimental conditions, participants were unaware of whether they had to undergo the trial in the VR or GI condition. Participants were then informed about the type of procedure. The random allocation sequence and participant enrollment were conducted by 2 experimenters trained in cognitive and behavioral therapy.

Intervention

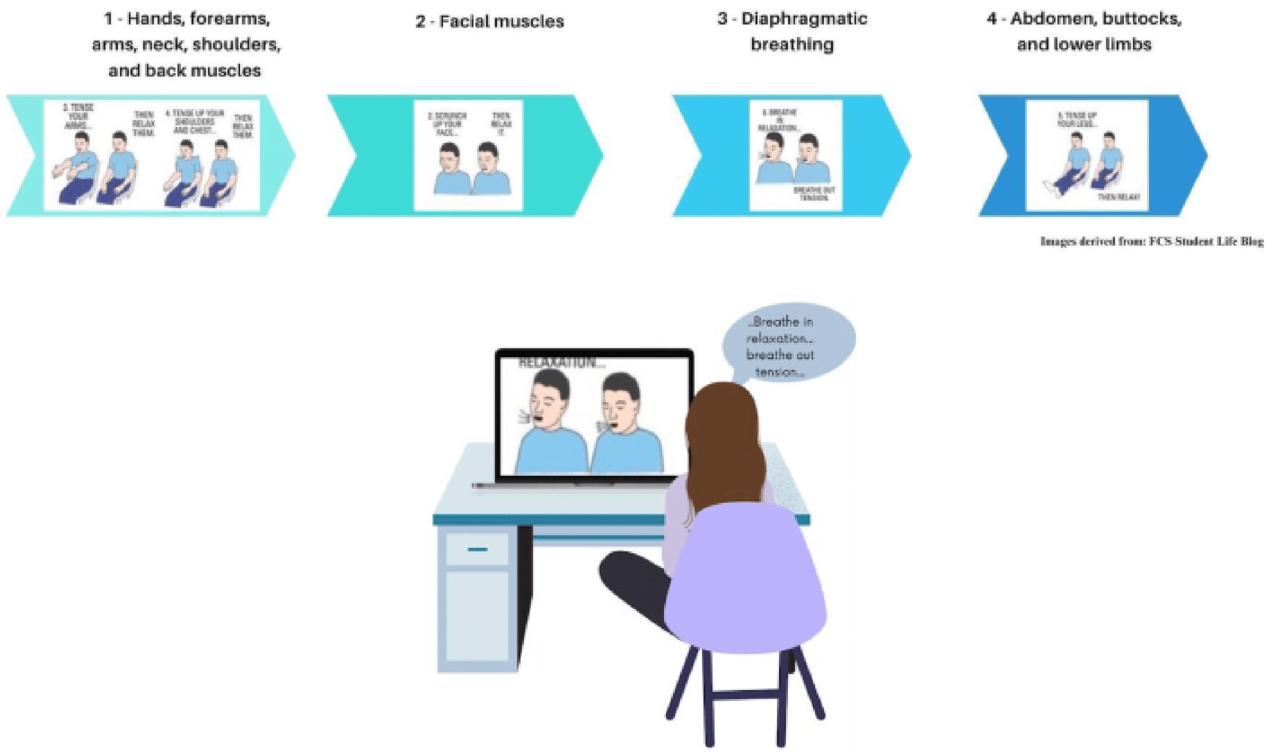
Hardware and Software Equipment

A commercially available VR headset (Meta Quest 2; Reality Labs) with an Alienware m15 Ryzen Edition R5" workstation and a link cable were used. The virtual scenarios were developed using the Unity framework (Unity Technologies) and the C# programming language (source of assets: Freesound [26], Unity, Poly Haven, and HDRIs). The code was versioned via GitLab. More detailed information about the virtual environment design is provided in the study by Pardini et al [18].

Procedure

For each group, the intervention implied previous learning of the adapted abbreviated PMRT [27] in 4 web-based active PMRT sessions. Each session required approximately 25 minutes (Figure 1). To check whether participants completed the active PMRT sessions, the therapists conducting the experimental procedure could check whether participants completed the assessment phases and the active PMRT sessions accessing Moodle (Moodle HQ). Therapists sent a reminder to participants a day before each session in the form of a direct message from the Moodle platform that was received as a personal email on the webmail university platform.

Figure 1. Illustrative examples of the relaxation exercises during the 4 web-based active progressive muscle relaxation technique sessions.



The first session focused on tensing and releasing the hands, forearms, arms, neck, shoulders, and back muscles; the second session focused on the facial muscles; the third session focused on teaching diaphragmatic breathing; and the fourth session focused on the abdomen, buttocks, and lower limbs (Figure 1).

The web-based sessions took place on the Zoom platform and were conducted by a cognitive behavioral therapist. Each Zoom session consisted of (1) filling out the State-Trait Anxiety Inventory–Form Y1 (STAI-Y1; state anxiety form), (2) sharing general standardized instructions for relaxation and the PMRT background with the participant, and (3) active PMRT session focused on a particular body section.

The fifth VR session was conducted at the VR laboratory at the University of Padova. All participants were asked to wear the Xiaomi 2 smartwatch to detect their heart rate frequency before, during, and after the relaxation session. A total of 5 measurements (1 per minute) were taken before the relaxation experience, 12 were taken during the entire exposure, and 5 were taken after the experience in the virtual context. Both the PMRT with GI and the PMRT with VR sessions were approximately 12 minutes long. This duration was also established based on previous studies' outcomes and to avoid potential cybersickness symptoms [28,29].

Specifically, the compared groups' conditions were characterized as follows: (1) the *PMRT and GI condition* consisted of the deployment of a standard behavioral intervention based on 4 individual active PMRT sessions via Zoom (sessions 2-5; T1-T4), an in vivo PMRT relaxation session with GI conducted by the psychotherapist (session 6; T1) a week after the baseline assessment (session 1; T0), and a follow-up phase (session 7; T2) after 2 weeks consisting of recovering the in-imagination relaxing scenario and a PMRT session (Figure 2); and (2) the *PMRT and VR condition* consisted of 4 active PMRT sessions administered via Zoom (sessions 2-5; T1-T4), a passive PMRT session integrated with the exposure to personalized VR scenarios deployed using the Meta Quest 2 HMD (session 6; T1), and a follow-up phase consisting of the same activities as the PMRT and GI condition (session 7; T2; Figure 3).

Figure 2. Progressive muscle relaxation technique (PMRT) and guided imagery (GI) exposure.

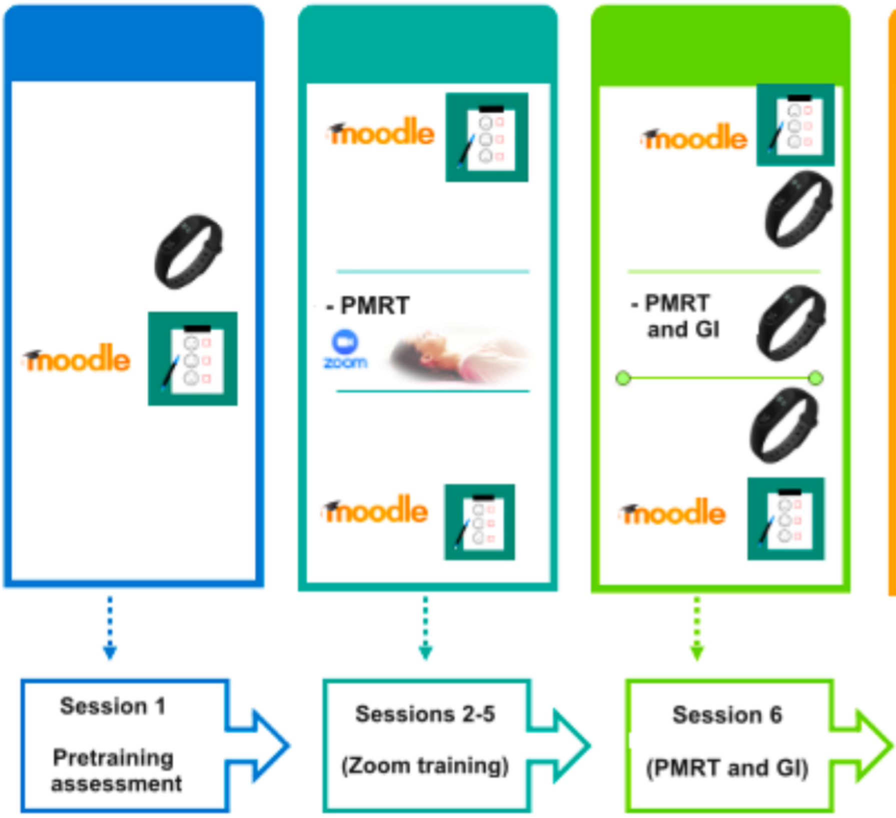
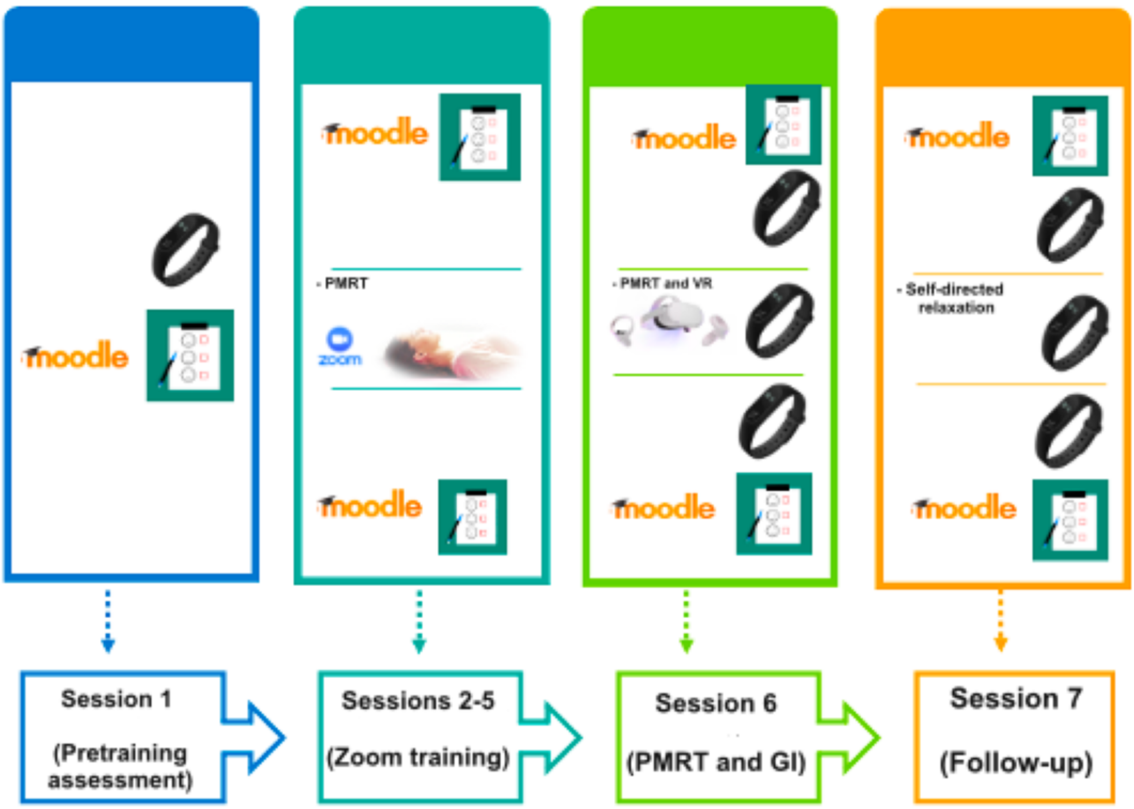


Figure 3. Progressive muscle relaxation technique (PMRT) and personalized virtual reality (VR) exposure.



The assessment phases were completed via the Moodle open-source learning platform. For this purpose, students could use their institutional accounts to access the platform. The administrator of the experimental procedure then created an ID profile for each participant. To access the Moodle platform, participants did not have to pay.

The assessment phases took place at baseline (T0); before and after each of the 4 active PMRT sessions; before and after the passive PMRT session administered with VR or GI (T1); and a week later, when participants were asked to lead the relaxation session autonomously (T2).

The T0 (baseline) assessment phase lasted approximately 40 minutes; it was the same for all participants and was administered at the VR laboratory at the University of Padova. It involved the administration of the following measures: (1) a demographic schedule [30]; (2) a series of self-report questionnaires investigating depression, anxiety, stress, quality of life, and distress coping strategies (State-Trait Anxiety Inventory–Form Y [STAI-Y]; Depression, Anxiety, and Stress Scale–21 [DASS-21]; Psychological General Well-Being Index [PGWBI]; and Coping Orientation to the Problems Experienced–*Nuova Versione Italiana* [COPE-NVI]); and (3) resting heart rate detection using the Xiaomi Mi Band 2.

Before and after each relaxation session (T1-T4), in approximately 20 minutes, the personal level of tension and relaxation was assessed using a visual analog scale (VAS) from 0 (no tension) to 10 (extreme tension level). The state anxiety level was evaluated based on the STAI-Y1. The 4 relaxation sessions were administered 2 to 3 days apart for all 3 groups. The assessment phase was web-based administered through Moodle, an e-learning platform used for data collection.

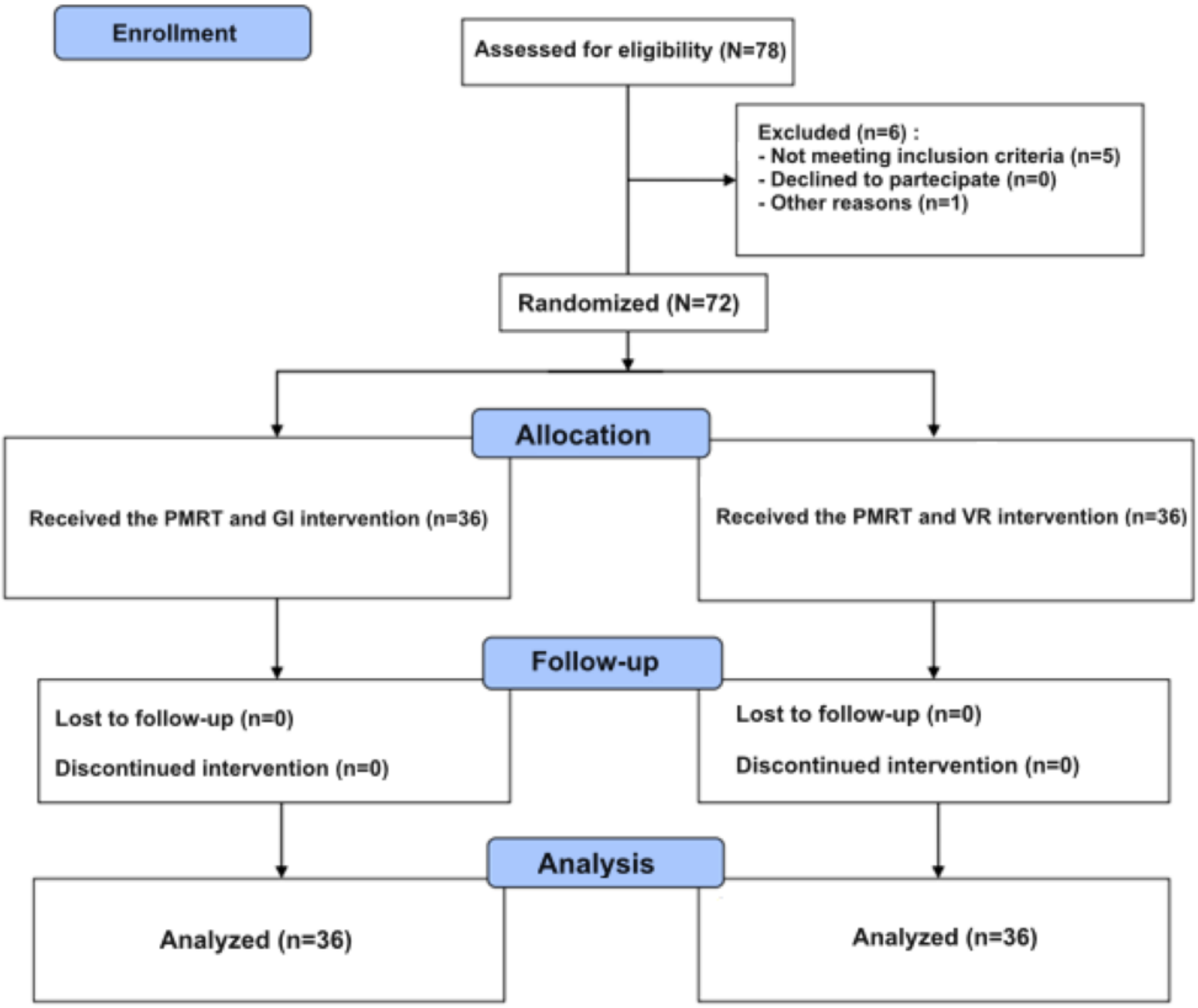
The T1 phase (day 7) took approximately 60 minutes and took place at the VR laboratory at the University of Padova. Before and after the relaxation session, states of tension and relaxation and anxiety were assessed using a 0 (no tension) to 10 (extreme tension level) VAS. The state anxiety level was evaluated based on the STAI-Y1. Participants were then exposed to a PMRT session merged with a VR or GI procedure. Before the in-imagination or VR experience, all the participants filled out the Vividness of Visual Imagery Questionnaire and the Test of Visual Imagery Control. After the PMRT session, users filled out a series of self-report questionnaires investigating depression, anxiety, and stress (STAI-Y and DASS-21), and only the VR group filled out the Virtual Reality Symptom Questionnaire (VRSQ) to monitor VR-related side effects (eg, sickness) and the International Test Commission–Sense of Presence Inventory (ITC-SOPI) to assess the sense of presence at the end of the T1 phase. The Xiaomi Mi Band 2 was used during the entire T1 phase to detect resting heart rate activity. The assessment phase was administered through the Moodle e-learning platform.

The T2 phase was deployed for approximately 45 minutes at the VR laboratory at the University of Padova. Before and after the relaxation session, states of tension and relaxation and anxiety were assessed using a VAS from 0 (no tension) to 10 (extreme tension level). The state anxiety level was evaluated based on the STAI-Y1. All users were exposed to a self-GI experience in which those who were part of the VR group were asked to recall the personalized VR scenario experienced during the T1 phase (day 7). The GI group retrieved instead the image that participants had used in association with the PMRT during the T1 phase (day 7). After the session, participants filled out a series of self-report questionnaires investigating depression, anxiety, stress, and quality of life (STAI-Y, DASS-21, and PGWBI) and an ad hoc version based on the ITC-SOPI to assess the sense of presence experienced during the imagery session. This assessment phase was administered based on the Moodle e-learning platform. The Mi Band 2 was used during the entire T1 phase administration (day 7) to detect resting heart rate activity.

Sample Size Estimation

To investigate whether the parameters based on our sample size could be acceptable, a formal sample size calculation was conducted using the G*Power (version 3.1) software [31]. As a statistical test, the repeated-measure ANOVA between the factors was considered. The effect size was 0.28, the Cronbach α was .05, and the power ($1 - \beta$ error probability) was 0.80. We had 2 separate groups and 3 measurements. On the basis of these parameters, it was estimated that at least 35 participants should be recruited for each group. The enrollment process is summarized in Figure 4.

Figure 4. CONSORT flow diagram. Progressive muscle relaxation technique (PMRT) and guided imagery (GI) condition: "As usual" intervention; PMRT and virtual reality (VR) condition: "Experimental" intervention.



Statistical Analysis

Quantitative statistical analyses were conducted using the SPSS (version 29.0; IBM Corp) software [32]. Frequencies, means, and SDs were measured to explore the sociodemographic features.

To examine the between- and within-subject differences, the 2 groups were evaluated using a repeated-measure multivariate analysis of covariance (MANCOVA) for mixed designs, multivariate MANCOVA, and *t* tests. To compare the differences between and within groups, pairwise comparisons with Bonferroni CI adjustment were calculated. Multiple linear regression analyses were conducted to investigate whether the sense of presence, engagement, and perception that the scenario was realistic could be predictive of the state anxiety level experienced after the VR exposure.

Outcome Measures

The measures were administered before and after each relaxation session. The heart rate frequency was recorded based on the Xiaomi Mi Band 2 before the session to obtain baseline data during and after the trials.

A sociodemographic schedule was filled out to obtain information about gender; age; mother tongue; marital status; years of education; occupation; psychological, medical, and neuromuscular problems; use of drugs; and whether participants had relaxation training experience or had used VR devices in the past.

The STAI-Y [33-35] is a self-report questionnaire that allows for the investigation of state and trait anxiety using 40 items on a 4-point Likert scale. Items are grouped into 2 scales focused on how participants generally feel (trait anxiety) or what they experience at particular times (state anxiety). The reliability and validity of the STAI-Y are good in the Italian sample. This study highlighted a moderate to acceptable internal consistency for the state and trait anxiety scales administered (.70<Cronbach α<.84).

The VAS for the tension and relaxation level is a measure of tension and relaxation intensity administered before and after each VR experience. Participants had to express how tense and activated they felt (0=*not at all*; 10=*completely*). Lower scores indicated higher relaxation levels.

The DASS-21 [36-38] is a self-report questionnaire based on 21 items that provides information about anxiety, depression, and stress symptomatology on a 4-point Likert scale from 0 to 3. Internal consistency and convergent, divergent, and criterion-oriented validity are adequate in the original and Italian versions. On the basis of our sample, a moderate to acceptable internal consistency emerged for the total and the 3 subscales (.77<Cronbach α<.87).

The PGWBI [39-41] is a self-report questionnaire consisting of 22 items that provides a general subjective assessment of psychological well-being. It comprises 6 subscales: anxiety, depression, positivity and well-being, self-control, general health, and vitality. The scores for all subscales can be summarized to provide a summary score, which reaches a maximum of 110 points representing the best achievable "well-being." The tool's psychometric properties are good for the original version and Italian validation. Considering our sample, an acceptable internal consistency emerged for the total score (Cronbach α=.72).

The COPE-NVI [42] is a 60-item self-report questionnaire on a 5-point Likert scale that investigates how often people use certain coping strategies with stressful or difficult events. Items are grouped into 5 subscales referring to different coping strategies: social support, avoidance strategies, positive attitude, problem-solving, and turning to religion. This tool is psychometrically valid to measure coping styles in the Italian context. Considering our sample, an acceptable internal consistency emerged for the total and the 6 subscales (.72<Cronbach α<.89).

The Vividness of Visual Imagery Questionnaire [43-46] comprises 16 items on a 5-point Likert scale and investigates individual differences regarding the ability to imagine visual contexts vividly. The participant is asked to generate 4 mental images and evaluate their vividness. On the basis of our sample, a good internal consistency was observed (Cronbach α=.89).

The Test of Visual Imagery Control [45-47] is a measure that evaluates individual differences in the ability to control and modify mental images intentionally. For example, participants are asked to visualize a car and then transform the image according to 10 different descriptions. Responses are recorded. On the basis of our sample, a good internal consistency was found (Cronbach α=.86).

The ITC-SOPI [48,49] is a questionnaire consisting of 42 items on a 5-point Likert scale that allows for the investigation of the sense of presence experienced in a VR context. It comprises 4 subscales investigating the sense of physical space, level of engagement experience in the virtual context, ecological validity, and negative effects of exposure. On the basis of our sample, a good internal consistency was found for all subscales (.67<Cronbach α<.78).

The VRSQ [50,51] assesses the general and eye-related physical symptoms of exposure to a VR environment. The score assigned to each item ranges from 0 to 6, with a maximum total score of 84 (48 for general symptoms and 36 for eye symptoms). Higher scores represent worse symptoms, with 0 corresponding to no adverse effects and 84 to serious adverse effects.

Ethical Considerations

The proposed study protocol was approved by the institutional review board of the Interdepartmental Ethical Committee of Psychology (17 Area) of the University of Padova (Italy; approval 4701; April 29, 2022). The study complied with the relevant ethical regulations of the Declaration of Helsinki (Italian law 196/2003; European Union General Data Protection Regulation 679/2016).

Results

Sociodemographic Features and Comparisons

The sociodemographic features of the study sample are outlined in Table 1.

Table 1. Demographic features and comparisons considering psychological constructs.

	Virtual reality group (n=36)	Guided imagery group (n=36)	F test (*df*)	Chi-square (*df*)	η_p^2
Gender (women), n (%)	28 (78)	27 (75)	N/A[a]	0.8 (2)	N/A
Age (years), mean (SD)	23.83 (6.10)	30.42 (8.36)	14.56 (1, 69)[b]	N/A	0.17
Years of education, mean (SD)	16.81 (1.31)	18.25 (2.60)	8.87 (1, 69)[b,c]	N/A	0.11
Marital status, n (%)			N/A	9.3 (2)[b,d]	N/A
Single	18 (50)	16 (44)			
Engaged, noncohabiting	17 (47)	10 (28)			
Married and cohabiting	1 (3)	10 (28)			
Medication (yes), n (%)	8 (22)	5 (14)	N/A	0.5 (2)	N/A
Psychological problems (yes), n (%)	15 (42)	12 (33)	N/A	0.5 (2)	N/A
Medical problems (yes), n (%)	6 (17)	10 (28)	N/A	1.3 (2)	N/A
Experienced relaxation protocol in the past (yes), n (%)	8 (22)	3 (8)	N/A	2.7 (2)	N/A
Experience with VR[e] in the past (yes), n (%)	10 (28)	8 (22)	N/A	0.3 (2)	N/A
STAI-Y2[f] (total), mean (SD)	48.06 (3.78)	47.56 (3.56)	0.96 (1, 69)	N/A	0.05
DASS-21[g] (anxiety), mean (SD)	3.61 (1.87)	3.69 (1.95)	0.03 (1, 69)	N/A	0.001
DASS-21 (depression), mean (SD)	4.46 (4.09)	5.47 (3.70)	1.18 (1, 69)	N/A	0.02
DASS-21 (stress), mean (SD)	8.56 (3.75)	7.69 (3.00)	1.16 (1, 69)	N/A	0.02
COPE-NVI[h] (social support), mean (SD)	34.92 (6.36)	32.11 (6.47)	3.44 (1, 69)	N/A	0.05
COPE-NVI (avoidance), mean (SD)	22.83 (4.91)	23.00 (4.04)	0.03 (1, 69)	N/A	0.001
COPE-NVI (positive attitude), mean (SD)	30.78 (5.03)	30.72 (5.26)	0.002 (1, 69)	N/A	0.001
COPE-NVI (problem-solving orientation), mean (SD)	31.97 (5.03)	31.69 (5.27)	0.05 (1, 69)	N/A	0.001
COPE-NVI (transcendental orientation), mean (SD)	17.44 (3.49)	17.31 (3.47)	0.03 (1, 69)	N/A	0.001
PGWBI-22[i] (total), mean (SD)	60.86 (6.97)	63.52 (9.48)	1.85 (1, 69)	N/A	0.03
VVIQ[j] (total), mean (SD)	58.53 (9.37)	58.39 (11.21)	0.003 (1, 69)	N/A	0.001
TVIC[k] (total), mean (SD)	41.56 (5.58)	43.44 (5.32)	2.16 (1, 69)	N/A	0.03

[a]N/A: not applicable.
[b]*P*<.001.
[c]*P*<.01.
[d]*P*<.05.
[e]VR: virtual reality.
[f]STAI-Y2: State-Trait Anxiety Inventory–Form Y2.
[g]DASS-21: Depression, Anxiety, and Stress Scale–21.
hCOPE-NVI: Coping Orientation to the Problems Experienced–*Nuova Versione Italiana*.
[i]PGWBI-22: Psychological General Well-Being Index–22.
[j]VVIQ: Vividness of Visual Imagery Questionnaire.
[k]TVIC: Test of Visual Imagery Control.

A multivariate ANOVA was conducted to investigate whether the groups differed in sociodemographic and psychological characteristics before the relaxation training. A difference between the groups was found only for age, years of education, and marital status (Table 1).

To control for the possible impact that baseline differences could have on the research outcomes, the effects of all these variables were controlled for in the subsequent analyses. The psychological problems referred to by participants were related to problematic relationships with parents, low self-esteem, and relational problems. Moreover, for individuals in the VR group, the VRSQ was administered to assess the possible collateral effects of VR exposure–related nausea. On average, individuals showed light susceptibility to general and eye-related motion sickness levels (VRSQ general score: mean 1.39, SD 1.32, range 0-4; VRSQ eye symptom score: mean 1.56, SD 1.32, range 0-5).

Objective 1: Investigate Whether PMRT Associated With a Personalized Relaxing Scenario in VR Can Be More Effective in Reducing State Anxiety, Tension and Activation, and Heart Rate Frequency

To investigate the differences between the “Virtual Reality” and “Guided Imagery” groups before and after the complementary relaxation experience in session 6 (T1) and session 7 (T2; Figures 2 and 3), we applied a repeated-measure MANCOVA. No significant within-subject main effect emerged ($F_{2,67}$=1.01; P=.37). A significant between-subject main effect was found ($F_{1,67}$=30.56; P<.001), and the VR group showed a greater change after the VR exposure than the GI group did in state anxiety (STAI-Y1; $F_{1,67}$=10.27; P<.001). After the relaxation experience at T1 (t_{67}=−7.82; P<.001), the VR group displayed lower state anxiety levels than the GI group both before (t_{67}=−2.63; P=.01; Cohen d=0.91) and after (t_{67}=−7.23; P<.001; Cohen d=2.45) the relaxation session during session 7 (T2; Table 2). Pairwise comparisons between and within the groups were performed before and after the relaxation experience at T1 and T2 (see also Multimedia Appendix 1). After the relaxation sessions at T1 and T2, the VR and GI groups experienced a significant decrease in state anxiety (P<.001). Pairwise comparisons within groups highlighted a significant reduction in perceived state anxiety at T1 and T2 for both the VR and GI groups (Table 3).

Table 2. Descriptive analysis for group differences.

Dependent variable, time, and group type	Values, mean (SD)
STAI-Y1[a] (T1)	
Before[b]	
Virtual reality (n=36)	48.06 (3.52)
Guided imagery (n=36)	50.16 (5.16)
After[c]	
Virtual reality (n=36)	31.17 (4.54)
Guided imagery (n=36)	40.00 (4.28)
STAI-Y1 (T2)	
Before	
Virtual reality (n=36)	48.89 (3.04)
Guided imagery (n=36)	51.78 (3.99)
After	
Virtual reality (n=36)	32.39 (4.33)
Guided imagery (n=36)	40.75 (4.74)

[a]STAI-Y1: State-Trait Anxiety Inventory–Form Y1.

[b]Refers to the assessment filled out before the relaxation experience at T1 and T2.

[c]Refers to the assessment filled out after the relaxation experience at T1 and T2.

Table 3. Within-group pairwise comparisons (Bonferroni CI adjustment).

Dependent variable, group type, and time	Mean difference (before − after)	SE	*t* test (*df*)	Cohen *d*
STAI-Y1 [a] (T1)				
Virtual reality (n=36)				
Before[b] and after[c]	17.13[d]	1.07	16.01 (67)	4.15
Guided imagery (n=36)				
Before and after	9.93[d]	1.07	9.30 (67)	2.14
STAI-Y1 (T2)				
Virtual reality (n=36)				
Before and after	16.58[d]	1.10	16.40 (67)	4.41
Guided imagery (n=36)				
Before and after	10.69[d]	1.10	10.33 (67)	2.52

[a]STAI-Y1: State-Trait Anxiety Inventory–Form Y1.

[b]Refers to the assessment filled out before the relaxation experience at T1 and T2.

[c]Refers to the assessment filled out after the relaxation experience at T1 and T2.

[d]P<.001.

Regarding the outcomes that emerged from the VAS, no significant differences were observed before the relaxation session at T1 (VR group: mean 3.71, SD 0.37; GI group: mean 4.13, SD 0.37; t_{67}=−0.76; P=.44; η_p^2=0.009), but the groups differed after the T1 session, with higher levels of tension experienced by participants in the GI group (VR group: mean 1.05, SD 0.30; GI group: mean 2.06, SD 0.30; t_{67}=−2.23; P=.03; η_p^2=0.069). The same pattern was observed at T2, with no differences between the groups before T2 (VR group: mean 3.52, SD 0.35; GI group: mean 4.17, SD 0.35; t_{67}=−1.24; P=.22; η_p^2=0.022) and higher scores in tension obtained by the individuals in the GI group after the relaxation session (VR group: mean 2.06, SD 1.71; GI group: mean 2.58, SD 1.80; t_{67}=−2.06; P=.04; η_p^2=0.060). Moreover, the levels of tension decreased in each group after the relaxation session at T1 (VR group: t_{67}=5.39, P<.001, and Cohen d=5.05; GI group: t_{67}=4.45, P<.001, and Cohen d=7.20) and T2 (VR group: t_{67}=8.55, P<.001, Cohen d=7.90; GI group: t_{67}=6.68, P<.001, and Cohen d<6.15).

Heart rate was assessed 3 times (before, during, and after the relaxation exposure). The Mauchly test of sphericity was adequate for the heart rate recorded in both sessions 6 and 7 (P=.09). Heart rate frequency was found to have a significant main effect between groups ($F_{1,67}$=4.87; P=.01) but to be not significant within groups ($F_{2,134}$=0.51; P=.73). An interaction between the “time of assessment” and “group” factors for heart rate ($F_{2,134}$=2.22; P=.01) was observed. Specifically, individuals in the VR group recorded lower levels of heart rate during (mean 68.26, SD 5.49) and after (mean 69.11, SD 4.98) the VR exposure at T1 than those of individuals in the GI condition (heart rate during: mean 72.86, SD 5.23; heart rate after: mean 72.22, SD 4.73; $-2.79<t_{1,\ 67}<-3.60$; P<.01). No significant outcomes were observed at T2.

Objective 2: Understand Whether VR Promotes a Better Sense of Presence and Engagement in the Scenario Compared With GI After Session 6 (T1) and Whether it Helps Recall the Image and Be Immersed in the Relaxing Scenario in Session 7 (T2)

Individuals in the VR group reported feeling higher engagement after the experience than participants in the GI group (ITC-SOPI–Engagement: $F_{1,67}$=2.85; P=.03; η_p^2=0.15). Moreover, the VR group participants reported experiencing a more realistic environment than that experienced by individuals who used imagination to create the scenario (ITC-SOPI–Ecological Validity: $F_{1,67}$=4.38; P=.003; η_p^2=0.21). No differences between the groups regarding sense of presence were found (ITC-SOPI–Sense of Presence: $F_{1,67}$=1.99; P=.11; η_p^2=0.11; Table 4).

Table 4. Multivariate analysis of covariance.

Dependent variable and group type	Values, mean (SD)
ITC-SOPI[a]–Sense of Presence (T1)	
Virtual reality (n=36)	60.06 (8.64)
Guided imagery (n=36)	57.00 (8.99)
ITC-SOPI–Engagement (T1)	
Virtual reality (n=36)	46.61 (4.66)
Guided imagery (n=36)	42.42 (5.71)
ITC-SOPI–Ecological Validity (T1)	
Virtual reality (n=36)	19.92 (2.12)
Guided imagery (n=36)	17.67 (2.75)
ITC-SOPI–Sense of Presence (T2)	
Virtual reality (n=36)	52.81 (12.56)
Guided imagery (n=36)	52.69 (11.55)
ITC-SOPI–Engagement (T2)	
Virtual reality (n=36)	40.40 (6.88)
Guided imagery (n=36)	40.36 (6.16)
ITC-SOPI–Ecological Validity (T2)	
Virtual reality (n=36)	17.44 (3.85)
Guided imagery (n=36)	14.17 (4.06)

[a]ITC-SOPI: International Test Commission–Sense of Presence Inventory.

To investigate whether VR facilitated image recall during session 7 more than the GI technique did, the sense of presence and engagement experiences were assessed after the relaxation experience in session 7 (T2). Individuals previously exposed to the personalized VR scenario referred to perceiving the same scenario recalled in T2 as more realistic than did individuals that, in the previous session (T1), were exposed to an imagined scenario (ITC-SOPI–Ecological Validity: $F_{1,67}$=3.21; P=.02; η_p^2=0.12). No differences emerged between the groups at session 7 (T2) for sense of presence (ITC-SOPI–Sense of Presence: $F_{1,67}$=0.76; P=.55; η_p^2=0.04) and engagement (ITC-SOPI–Engagement: $F_{1,67}$=2.30; P=.07; η_p^2=0.12; Table 4).

Objective 3: Investigate Group Differences Regarding Trait Anxiety, Depressive Symptoms, Stress, Coping, and Well-Being

To investigate the differences in trait anxiety, depressive symptoms, and stress between the VR and GI groups at baseline (T0), after session 6 (T1), and after session 7 (T2), we applied repeated-measure MANCOVA. The Mauchly test of sphericity was adequate only for the State-Trait Anxiety Inventory–Form Y2 (P>.05). The Hyunh-Feldt correction was adopted for the DASS-21 anxiety subscale. The Greenhouse-Geisser correction was adopted for the DASS-21 depression and stress subscales. A significant interaction between the variables "time of assessment" and "group" factors ($F_{2,140}$=3.62; P=.02) emerged. The "Virtual reality" group had lower trait anxiety levels than the "Guided Imagery" group after the relaxation session during session 7 (T2; t_{67}=−2.43; P=.02; Tables 5-7). No significant interactions between "time of assessment" and "group" were found for stress ($F_{1.16,81.41}$=1.15; P>.05), anxiety ($F_{1.89,132.58}$=3.11; P>.05), and depressive symptoms ($F_{1.16,81.23}$=1.09; P>.05) assessed using the DASS-21.

Table 5. Descriptive analysis of the State-Trait Anxiety Inventory–Form Y2 (STAI-Y2) at T1 and T2.

Dependent variable, time, and group type	Mean (SD)
STAI-Y2	
T1	
Virtual reality (n=36)	42.75 (3.07)
Guided imagery (n=36)	44.11 (4.58)
T2	
Virtual reality (n=36)	44.25 (4.07)
Guided imagery (n=36)	47.36 (4.29)

Table 6. Between-group pairwise comparisons for the State-Trait Anxiety Inventory–Form Y2 (STAI-Y2; Bonferroni CI adjustment).

Group type	Mean difference (before − after)	SE	*t* test (*df*)	Cohen *d*
STAI-Y2 (T0)	0.78	0.98	0.80 (67)	0.14
Virtual reality (n=36)				
Guided imagery (n=36)				
STAI-Y2 (T1)	−0.17	0.99	−2.18 (67)	−0.35
Virtual reality (n=36)				
Guided imagery (n=36)				
STAI-Y2 (T2)	*−2.76*[a]	1.13	−2.44 (67)	−0.74
Virtual reality (n=36)				
Guided imagery (n=36)				

[a]$P<.05$.

Table 7. Within-group pairwise comparisons for the State-Trait Anxiety Inventory–Form Y2 (STAI-Y2; Bonferroni CI adjustment).

Group type and time	Mean difference (before − after)	SE	*t* test (*df*)	Cohen *d*
Virtual reality (n=36)				
T0-T1	*4.83* [a,b]	0.94	5.14 (67)	1.54
T0-T2	*3.77* [a]	0.84	4.49 (67)	0.97
T1-T2	−1.06	0.96	−1.10 (67)	−0.42
Guided imagery (n=36)				
T0-T1	*3.90* [a]	0.95	4.41 (67)	0.84
T0-T2	0.23	0.84	0.27 (67)	0.05
T1-T2	*−3.67*	0.93	−3.95 (67)	−0.73

[a]$P<.001$.

[b]Italicization indicates that the *P* value is statistically significant.

Differences within and between groups in coping strategies and psychological well-being were investigated at baseline (T0) and after session 7 (T2). No significant differences were found between the "time of assessment" and "group" factors for both the PGWBI questionnaire ($F_{1,70}$=0.63; P>.05) and all the COPE-NVI subscales (COPE-NVI–Social Support: $F_{1,70}$=0.95 and P>.05; COPE-NVI–Avoidance Strategies: $F_{1,70}$=3.01 and P>.05; COPE-NVI–Positive Attitude: $F_{1,70}$=3.59 and P>.05; COPE-NVI–Problem-Solving: $F_{1,70}$=0.28 and P>.05; COPE-NVI–Turning to Religion: $F_{1,70}$=3.71 and P>.05).

Discussion

Principal Findings

Research on the use of personalized VR scenarios for promoting relaxation is growing, although few conclusions on their effectiveness have been reached. We know even less about the usefulness of integrating standardized and evidence-based relaxation techniques (eg, PMRT) with new technologies that could promote the independent use by patients, providing a more economical solution to treatment costs. As stated previously, the possibility of personalizing audio and visual stimuli in a VR environment is a promising approach for meeting users' preferences and needs [22], and it has shown encouraging results in terms of both feasibility and potential impact on psychological well-being [18].

The general aim of this pilot study was to evaluate the impact of a novel complementary relaxation training composed of PMRT and the exposure to a personalized relaxing scenario in VR aimed at reducing anxiety and promoting relaxation in a sample of university students.

First, we were interested in investigating the differences in state anxiety between and within groups in the T1 session, where participants were directly exposed to a virtual or imaginative relaxation session. Although a reduction in state anxiety was observed in both the VR and GI groups, our results highlighted that being in a personalized virtual scenario promoted a more significant reduction in self-perceived state anxiety and tension than being in the imaginative condition. Our results are consistent with those of other studies that showed that VR could be a useful tool in reducing stress and promoting relaxation from an assessment based on self-reported measures [52]. Although the target user sample in the study by Hoag et al [53] comprised children and young adults with acute and chronic illnesses, our results align with their results, showing a significant decline in state anxiety from before to after VR exposure. Our data support the role of VR in facilitating a powerful distraction effect on users, reducing their focus on thoughts and external events that would elicit anxiety responses [54,55].

In addition, our data highlighted how engagement with VR scenarios contributes to reducing perceived anxiety after the VR experience. Indeed, it seems that exposure to the VR scenario instead of the imaginative one facilitates the reduction in anxiety levels after the relaxation session. Even if our results need to be further investigated with larger user samples, they are aligned with those of previous studies that underlined the potential benefit of using pleasant, relaxing, and immersive VR scenarios for facilitating relaxation and engagement in individuals from the general population [18,56,57]. A point of strength of this experimental design is the assessment of the impact that the ability to control and vividly picture mental images may have. As we did not find differences between the 2 groups in these abilities before the exposure to the relaxation session in T1, we can support the observed positive effect that exposure to a VR context can have on reducing anxiety.

Considering the VR group, our data showed similar outcomes regarding state anxiety as those investigated at T2. Indeed, when both groups were asked to self-administer the relaxation session, which consisted of recalling the personalized image they had experienced in VR or imagination at T1, those who had been previously exposed to the personalized VR scenario obtained lower state anxiety scores than those in the other group after session 7 at T2. Considering that individuals in the VR group obtained significantly lower scores both before and after the relaxation session in T2, we confirm the hypothesis that being exposed to the self-managed session at T2 further contributed to lowering the anxiety level in the VR group.

Moreover, the VR group participants perceived the recalled scenario at T2 as more realistic than individuals in the GI group did. In addition, participants in the VR group obtained additional benefits in terms of relaxation and state anxiety reduction related to a more realistic sensorial experience that played a substantial role in facilitating the visualization of the scenario and enabled users to focus their attention on the relaxation activities. Consistent with other studies [55-57], this core outcome confirms the role of immersive VR in promoting relaxation through visualization, engagement, and immersion processes. The prominent impact of being exposed to realistic scenarios in VR on the enhanced visualization at T2 is a key contribution of our study, shedding light on the impact that exposure to VR may also have in more ecological everyday settings, such as when people are not wearing the HMD but have the chance to transfer the relaxation skills learned in VR to real-world situations. Our data are consistent with those of studies that show that a graphical representation of a scenario is more effective in the retention and recall processes than an imaginative representation [58,59].

Even if the aim was to compare VR scenarios with a scene on a PC, our outcome is in line with that of the study by Krokos et al [60], which highlighted the prominent impact of VR scenarios on memory recall ability. The hypotheses of this study and that by Krokos et al [60] are anchored on classic studies in cognitive psychology based on the method of loci [61] and the context-dependent memory theory [62]. These theories imply the essential role of learning and mnemonic processes in creating an association between the mnemonic content and a mental frame of scenarios and then recalling contents by mentally visualizing the scenarios in which the learning and memorization processes took place [60]. As that presence, immersion, and engagement in the VR scenarios imply sensorimotor contingencies similar to those in the real world [63] and the way we create and recall mental constructs is influenced by perception and action in the environment [64,65], our data coherently confirm the potential of immersive virtual environments in enhancing learning and recall for the intervention of vestibular and sensorimotor inputs [66].

Regarding our collected data on heart rate activity, our findings showed a more significant decrease in heart rate frequency in the VR group than in the GI group but only in the session in which participants were directly exposed to 1 of the 2 experimental conditions (at T1) and not when individuals had to self-administer the relaxation session (at T2). We derive from the differences found at T1 that were absent in T2 that the experience in VR was stronger and more engaging than in the condition in which individuals had to recall the immersive image without HMD support. Our data supporting previous studies' outcomes highlighted the impact of VR relaxing scenarios in maintaining lower levels of heart rate frequency than normal [67], but additional research is needed to deeply investigate whether and how naturalistic and relaxing VR scenarios induce relaxation and stress reduction by providing feedback on changes, for example, in heart rate frequency and variability, respiration rate, or skin conductance.

An interesting result is also related to differences over time and between groups in trait anxiety scores. Indeed, our data showed that, in the VR group, the decrease in trait anxiety scores found at T1 was maintained over time (at T2). Nevertheless, the same was not found for the GI group, in which a decrease in anxiety scores was observed at T1, but it significantly increased again at T2, returning to the baseline values (at T0). The 2 groups did not differ at T1, and both showed positive effects of the relaxation sessions. However, the most interesting fact is that, when anxiety was reevaluated after a week and we asked participants to respond considering how they had generally felt during the previous days, people who had used VR claimed to have a lower level of anxiety than that of participants in the other group. These data also need to be further investigated with more comprehensive samples of participants but support the claim that VR plays a crucial role in amplifying the effectiveness of already validated interventions, maintaining their effect over time.

No differences over time were highlighted in coping strategies and psychological well-being, and this could be because our sample was composed of individuals without a clinical diagnosis and high levels of distress at baseline, or it could even be traced back to the fact that these types of psychological constructs require more sessions and time to highlight an effective change.

Strengths and Limitations

This contribution adds new knowledge on the importance of customizing and personalizing digital interventions according to the users' perspectives, needs, and preferences [22]. Another point of strength of this experimental design is the assessment of the impact that the ability to control and vividly picture mental images may have. The outcomes were gathered from a well-designed pilot randomized controlled trial involving 2 selected groups of 72 university students whose sociodemographic and psychological characteristics were controlled for to balance their effect on the variables investigated. The experimental procedure adopted supports the reliability and validity of the results and conclusions presented in this paper. However, this study has some limitations that affect the generalizability of the findings. One limitation is that the results cannot be generalized to the entire nonclinical population as our sample consisted of students from a single university. Moreover, the fact that our sample comprised mainly young female participants may be considered as selection bias as the recruited sample may well be more receptive and motivated to participate in the intervention compared with people with other sociodemographic features [68]. In general, to further generalize and validate our results' effectiveness, the involvement of participants belonging to clinical and nonclinical populations, stratified according to different sociodemographic characteristics, should be considered. The use of VR can be helpful in overcoming several barriers that standard relaxation procedures present as it is less costly; promotes the availability of relaxing content that could be difficult to generate in the imagination; and is able to promote the sense of presence, immersion, and engagement closer to what can be obtained in a real-world situation but in a safer context. This is also a reason for considering its further investigation based on clinical and nonclinical samples in future studies. The preliminary results of our study highlight the potential of VR in reducing the number of psychotherapy sessions and their cost as it allows for partial self-management of the treatment.

Furthermore, in the case of patients with medical problems, the integration of VR could facilitate the administration of the relaxation intervention during specific invasive treatments (such as chemotherapy). If the merged administration of a customized VR relaxation scenario and PMRT is effective in obtaining a better subjective perception of relaxation than the standard procedure, it could allow the users to relax with greater autonomy. For this reason, assessing the efficacy of the PMRT and GI in alternative ways could extend treatment administration, especially in situations in which the standard procedure is more challenging. Future studies should consider structuring the relaxation protocol with more VR sessions to better understand the impact of a virtual scenario on relaxation. Another aspect to be considered in further experimental designs is the introduction of a control group that receives the training session with the physical presence of a therapist. Considering the impact of customization on anxiety and on engaging users in the virtual scenarios, another important aspect that needs to be considered is the opportunity to introduce customized stimuli based on the personal life experience of each participant in the virtual environment. As an example, introducing olfactory stimuli could be another essential aspect to enhance the sense of realism, immersion, and presence in the virtual scenarios [69,70]. A further limitation of this study is the possible exclusion of people who are not familiar with social media. Another limitation regards the limited use of objective data (eg, psychophysiological outcomes) for measuring anxiety and relaxation and the limited range of customization settings offered to users to adapt the VR environment to their preferences. The limitations of this study can be overcome through further investigation.

Conclusions

The main objective of psychological interventions is to offer people the opportunity to learn strategies to manage their daily lives independently and effectively. The opportunity to deploy VR for promoting self-management of state anxiety also in real-world situations constitutes an interesting line of investigation to be further explored in future studies involving larger nonclinical and clinical populations (eg, patients with chronic pain) to validate and standardize relaxation protocols integrated with new VR tools. This study has shown that personalized VR scenarios can be effective in improving relaxation and decreasing anxiety when integrated with the PMRT as a complementary relaxation method, thereby highlighting the need for further investigation.

Acknowledgments

This research project was funded by the Fondazione Valorizzazione Ricerca Trentina with a grant for the project "ViRLab_Salute," second call of the Next Generation 2021 grant. The authors wish to thank the students from the University of Padova for participating in the study and Dr Lucia Ronconi ("Statistical Consulting Service," University of Padova) for her contribution in revising the analysis.

Data Availability

The data set generated and analyzed during this study is available from the corresponding author upon reasonable request.

Authors' Contributions

SP wrote the manuscript, designed the study protocol, ran the intervention trial, conducted data analyses, and contributed to manuscript editing and revision. SG contributed to the design of the study protocol, revised the analysis and the original manuscript, and contributed to manuscript editing and revision. SO contributed to the curation and formal analysis, wrote the manuscript, and contributed to manuscript editing and revision. FF contributed to the curation and formal analysis and to manuscript editing and revision. MD worked on the hardware and software components. SF contributed to the design of the study protocol. CL contributed to manuscript editing and revision. CN contributed to the design of the study protocol, revised the analysis and the original manuscript, and contributed to manuscript editing and revision. All the authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Between-group pairwise comparisons (Bonferroni CI adjustment).
[DOCX File , 15 KB - mental_v11i1e48649_app1.docx]

Multimedia Appendix 2
CONSORT-EHEALTH (Consolidated Standards of Reporting Trials of Electronic and Mobile Health Applications and Online Telehealth) checklist (V 1.6.1).
[PDF File (Adobe PDF File), 750 KB - mental_v11i1e48649_app2.pdf]

References

1. Bayram N, Bilgel N. The prevalence and socio-demographic correlations of depression, anxiety and stress among a group of university students. Soc Psychiatry Psychiatr Epidemiol 2008 Aug;43(8):667-672. [doi: 10.1007/s00127-008-0345-x] [Medline: 18398558]
2. Wiegner L, Hange D, Björkelund C, Ahlborg Jr G. Prevalence of perceived stress and associations to symptoms of exhaustion, depression and anxiety in a working age population seeking primary care--an observational study. BMC Fam Pract 2015 Mar 19;16:38 [FREE Full text] [doi: 10.1186/s12875-015-0252-7] [Medline: 25880219]
3. Son C, Hegde S, Smith A, Wang X, Sasangohar F. Effects of COVID-19 on college students' mental health in the United States: interview survey study. J Med Internet Res 2020 Sep 03;22(9):e21279 [FREE Full text] [doi: 10.2196/21279] [Medline: 32805704]
4. Komariah M, Ibrahim K, Pahria T, Rahayuwati L, Somantri I. Effect of mindfulness breathing meditation on depression, anxiety, and stress: a randomized controlled trial among university students. Healthcare (Basel) 2022 Dec 22;11(1):26 [FREE Full text] [doi: 10.3390/healthcare11010026] [Medline: 36611488]
5. Ozamiz-Etxebarria N, Dosil Santamaría M, Eiguren Munitis A, Picaza Gorrotxategi M. Corrigendum: reduction of COVID-19 anxiety levels through relaxation techniques: a study carried out in northern Spain on a sample of young university students. Front Psychol 2020 Oct 28;11:609098 [FREE Full text] [doi: 10.3389/fpsyg.2020.609098] [Medline: 33250834]
6. Baird CL, Sands L. A pilot study of the effectiveness of guided imagery with progressive muscle relaxation to reduce chronic pain and mobility difficulties of osteoarthritis. Pain Manag Nurs 2004 Sep;5(3):97-104. [doi: 10.1016/j.pmn.2004.01.003] [Medline: 15359221]
7. Nasiri S, Akbari H, Tagharrobi L, Tabatabaee AS. The effect of progressive muscle relaxation and guided imagery on stress, anxiety, and depression of pregnant women referred to health centers. J Educ Health Promot 2018;7:41 [FREE Full text] [doi: 10.4103/jehp.jehp_158_16] [Medline: 29619392]
8. Lang PJ. Presidential address, 1978. A bio-informational theory of emotional imagery. Psychophysiology 1979 Nov;16(6):495-512. [doi: 10.1111/j.1469-8986.1979.tb01511.x] [Medline: 515293]
9. Benson H, Beary JF, Carol MP. The relaxation response. Psychiatry 1974 Feb;37(1):37-46. [doi: 10.1080/00332747.1974.11023785] [Medline: 4810622]
10. Syrjala KL, Abrams JR. Hypnosis and imagery in the treatment of pain. In: Gatchel RJ, Turk DC, editors. Psychological Approaches to Pain Management: A Practitioner's Handbook. New York, NY: The Guilford Press; 1996:231-258.

11. O'Neill LM, Barnier AJ, McConkey K. Treating anxiety with self‐hypnosis and relaxation. Contemp Hypn 2006 Feb 23;16(2):68-80 [FREE Full text] [doi: 10.1002/ch.154]
12. Arena JG, Blanchard EB. Biofeedback and relaxation therapy for chronic pain disorders. In: Gatchel RJ, Turk DC, editors. Psychological Approaches to Pain Management: A Practitioner's Handbook. New York, NY: The Guilford Press; 1996:179-230.
13. Bakke AC, Purtzer MZ, Newton P. The effect of hypnotic-guided imagery on psychological well-being and immune function in patients with prior breast cancer. J Psychosom Res 2002 Dec;53(6):1131-1137. [doi: 10.1016/s0022-3999(02)00409-9] [Medline: 12479996]
14. Myers CD, Robinson ME, Guthrie TH, Lamp SP, Lottenberg R. Adjunctive approaches for sickle cell chronic pain. Complement Health Pract Rev 1999 Dec 01;5(3):203-212 [FREE Full text] [doi: 10.1177/153321019900500303]
15. De Gauquier L, Brengman M, Willems K, Van Kerrebroeck H. Leveraging advertising to a higher dimension: experimental research on the impact of virtual reality on brand personality impressions. Virtual Real 2018 May 8;23(3):235-253 [FREE Full text] [doi: 10.1007/s10055-018-0344-5]
16. Lau KW, Lee PY. Shopping in virtual reality: a study on consumers’ shopping experience in a stereoscopic virtual reality. Virtual Reality 2018 Aug 14;23(3):255-268 [FREE Full text] [doi: 10.1007/s10055-018-0362-3]
17. Browning MH, Shin S, Drong G, McAnirlin O, Gagnon RJ, Ranganathan S, et al. Daily exposure to virtual nature reduces symptoms of anxiety in college students. Sci Rep 2023 Jan 23;13(1):1239 [FREE Full text] [doi: 10.1038/s41598-023-28070-9] [Medline: 36690698]
18. Pardini S, Gabrielli S, Dianti M, Novara C, Zucco GM, Mich O, et al. The role of personalization in the user experience, preferences and engagement with virtual reality environments for relaxation. Int J Environ Res Public Health 2022 Jun 13;19(12):7237 [FREE Full text] [doi: 10.3390/ijerph19127237] [Medline: 35742483]
19. Pizzoli SF, Triberti S, Monzani D, Mazzocco K, Kufel E, Porebiak M, et al. Comparison of relaxation techniques in virtual reality for breast cancer patients. In: Proceedings of the 5th Experiment International Conference. 2019 Presented at: exp.at '19; June 12-14, 2019; Funchal, Portugal p. 348-351 URL: https://ieeexplore.ieee.org/document/8876542 [doi: 10.1109/expat.2019.8876542]
20. Mazgelytė E, Rekienė V, Dereškevičiūtė E, Petrėnas T, Songailienė J, Utkus A, et al. Effects of virtual reality-based relaxation techniques on psychological, physiological, and biochemical stress indicators. Healthcare (Basel) 2021 Dec 14;9(12):1729 [FREE Full text] [doi: 10.3390/healthcare9121729] [Medline: 34946455]
21. Ch NA, Ansah AA, Katrahmani A, Burmeister J, Kun AL, Mills C, et al. Virtual nature experiences and mindfulness practices while working from home during COVID-19: effects on stress, focus, and creativity. Int J Hum Comput Stud 2023 Mar;171:102982 [FREE Full text] [doi: 10.1016/j.ijhcs.2022.102982] [Medline: 36536937]
22. Pizzoli SF, Mazzocco K, Triberti S, Monzani D, Alcañiz Raya ML, Pravettoni G. User-centered virtual reality for promoting relaxation: an innovative approach. Front Psychol 2019 Mar 12;10:479 [FREE Full text] [doi: 10.3389/fpsyg.2019.00479] [Medline: 30914996]
23. Holland AC, Kensinger EA. Emotion and autobiographical memory. Phys Life Rev 2010 Mar;7(1):88-131 [FREE Full text] [doi: 10.1016/j.plrev.2010.01.006] [Medline: 20374933]
24. Frazier P, Liu Y, Selvey A, Meredith L, Nguyen-Feng VN. Randomized controlled trials assessing efficacy of brief web-based stress management interventions for college students during the COVID pandemic. J Couns Psychol 2023 Apr;70(3):314-324. [doi: 10.1037/cou0000652] [Medline: 36689384]
25. Pardini S, Gabrielli S, Olivetto S, Fusina F, Dianti M, Forti S, et al. Personalized, naturalistic virtual reality scenarios coupled with web-based progressive muscle relaxation training for the general population: protocol for a proof-of-principle randomized controlled trial. JMIR Res Protoc 2023 Apr 17;12:e44183 [FREE Full text] [doi: 10.2196/44183] [Medline: 37067881]
26. Freesound. URL: https://freesound.org/ [accessed 2024-01-23]
27. Bernstein DA, Carlson CR, Schmidt JE. Progressive relaxation: abbreviated methods. In: Lehrer PM, Woolfolk RL, Sime WE, editors. Principles and Practice of Stress Management, 3rd Edition. New York, NY: The Guilford Press; 2007:88-122.
28. Basso JC, McHale A, Ende V, Oberlin DJ, Suzuki WA. Brief, daily meditation enhances attention, memory, mood, and emotional regulation in non-experienced meditators. Behav Brain Res 2019 Jan 01;356:208-220. [doi: 10.1016/j.bbr.2018.08.023] [Medline: 30153464]
29. Rooks JD, Morrison AB, Goolsarran M, Rogers SL, Jha AP. “We are talking about practice”: the influence of mindfulness vs. relaxation training on athletes’ attention and well-being over high-demand intervals. J Cogn Enhanc 2017 Apr 6;1(2):141-153. [doi: 10.1007/s41465-017-0016-5]
30. Pawlow LA, Jones GE. The impact of abbreviated progressive muscle relaxation on salivary cortisol and salivary immunoglobulin A (sIgA). Appl Psychophysiol Biofeedback 2005 Dec;30(4):375-387. [doi: 10.1007/s10484-005-8423-2] [Medline: 16385425]
31. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
32. IBM SPSS Statistics for Windows. IBM Corp. 2022. URL: https://www.ibm.com/products/spss-statistics [accessed 2024-01-18]

33. Spielberger CD, Gorsuch RL. Manual for the State-trait Anxiety Inventory (Form Y). Palo Alto, CA: Consulting Psychologists Press; 1983.
34. Spielberger CD. State-Trait Anxiety Inventory A Comprehensive Bibliography. 2nd edition. Palo Alto, CA: Consulting Psychologists Press; 1989.
35. Pedrabissi L, Santinello M. State-Trait Anxiety Inventory - Forma Y. Firenze, Italy: Giunti Psychometrics; 1989.
36. Lovibond SH, Lovibond PF. Manual for the Depression Anxiety Stress Scales. 2nd edition. Sydney, Australia: Psychology Foundation of Australia; 1995.
37. Lovibond PF, Lovibond SH. The structure of negative emotional states: comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. Behav Res Ther 1995 Mar;33(3):335-343. [doi: 10.1016/0005-7967(94)00075-u] [Medline: 7726811]
38. Bottesi G, Ghisi M, Altoè G, Conforti E, Melli G, Sica C. The Italian version of the Depression Anxiety Stress Scales-21: factor structure and psychometric properties on community and clinical samples. Compr Psychiatry 2015 Jul;60:170-181. [doi: 10.1016/j.comppsych.2015.04.005] [Medline: 25933937]
39. Dupuy HJ. The Psychological General Well-being (PGWB) index. In: Wenger NK, Mattson ME, Furburg CD, Elinson J, editors. Assessment of Quality of Life in Clinical Trials of Cardiovascular Therapies. New York, NY: Le Jacq Publishing; 1990:170-183.
40. Grossi E, Mosconi P, Groth N, Niero M, Apolone G. [Il Questionario Psychological General Well Being] Questionario per la valutazione dello stato generale di benessere psicologico. Versione Italiana. Milano, MI: Istituto di Ricerche Farmacologiche Mario Negri; 2002.
41. Grossi E, Groth N, Mosconi P, Cerutti R, Pace F, Compare A, et al. Development and validation of the short version of the Psychological General Well-Being Index (PGWB-S). Health Qual Life Outcomes 2006 Nov 14;4(1):88 [FREE Full text] [doi: 10.1186/1477-7525-4-88] [Medline: 17105655]
42. Sica C, Magni C, Ghisi M, Altoè G, Sighinolfi C, Chiri LR, et al. Coping orientation to problems experienced-Nuova versione Italiana (COPE-NVI): uno strumento per la misura degli stili di coping. Psicoter Cogn e Comportamentale 2008;14(1):27-53. [doi: 10.3280/pds2015-002007]
43. Marks DF. Visual imagery differences in the recall of pictures. Br J Psychol 1973 Feb;64(1):17-24 [FREE Full text] [doi: 10.1111/j.2044-8295.1973.tb01322.x] [Medline: 4742442]
44. Marks DF. Construct validity of the Vividness of visual imagery questionnaire. Percept Mot Skills 1989 Oct;69(2):459-465 [FREE Full text] [doi: 10.2466/pms.1989.69.2.459] [Medline: 2812992]
45. Antonietti A, Crespi M. Analisi di tre questionari per la valutazione della vividezza dell'immagine mentale [Analysis of three questionnaires for assessing the vividness of mental image]. Department of Psychology, Catholic University of The Sacred Heart. Technical report. 1995. URL: https://www.yumpu.com/it/document/view/34903893/analisi-di-tre-questionari-per-la-valutazione-della-vividezza-dell- [accessed 2024-01-24]
46. Iachini T, Maffei L, Masullo M, Senese VP, Rapuano M, Pascale A, et al. The experience of virtual reality: are individual differences in mental imagery associated with sense of presence? Cogn Process 2019 Aug;20(3):291-298 [FREE Full text] [doi: 10.1007/s10339-018-0897-y] [Medline: 30569268]
47. Gordon R. An investigation into some of the factors that favour the formation of stereotyped images. Br J Psychol Gen Sect 1949 Mar;39(3):156-167 [FREE Full text] [doi: 10.1111/j.2044-8295.1949.tb00215.x] [Medline: 18127238]
48. Lessiter J, Freeman J, Davidoff JB, Keogh E. A cross-media presence questionnaire: the ITC-sense of presence inventory. Presence (Camb) 2001 Jun;10(3):282-297. [doi: 10.1162/105474601300343612]
49. Chirico A, Gaggioli A. When virtual feels real: comparing emotional responses and presence in virtual and natural environments. Cyberpsychol Behav Soc Netw 2019 Mar;22(3):220-226. [doi: 10.1089/cyber.2018.0393] [Medline: 30730222]
50. Ames SL, Wolffsohn JS, McBrien NA. The development of a symptom questionnaire for assessing virtual reality viewing using a head-mounted display. Optom Vis Sci 2005 Mar;82(3):168-176. [doi: 10.1097/01.opx.0000156307.95086.6] [Medline: 15767873]
51. Chirico A, Maiorano P, Indovina P, Milanese C, Giordano GG, Alivernini F, et al. Virtual reality and music therapy as distraction interventions to alleviate anxiety and improve mood states in breast cancer patients during chemotherapy. J Cell Physiol 2020 Jun 20;235(6):5353-5362. [doi: 10.1002/jcp.29422] [Medline: 31957873]
52. Kim H, Kim DJ, Kim S, Chung WH, Park KA, Kim JD, et al. Effect of virtual reality on stress reduction and change of physiological parameters including heart rate variability in people with high stress: an open randomized crossover trial. Front Psychiatry 2021 Aug 10;12:614539 [FREE Full text] [doi: 10.3389/fpsyt.2021.614539] [Medline: 34447320]
53. Hoag JA, Karst J, Bingen K, Palou-Torres A, Yan K. Distracting through procedural pain and distress using virtual reality and guided imagery in pediatric, adolescent, and young adult patients: randomized controlled trial. J Med Internet Res 2022 Apr 18;24(4):e30260 [FREE Full text] [doi: 10.2196/30260] [Medline: 35436209]
54. McCaul KD, Malott JM. Distraction and coping with pain. Psychol Bull 1984;95(3):516-533 [FREE Full text] [doi: 10.1037//0033-2909.95.3.516]
55. Annerstedt M, Jönsson P, Wallergård M, Johansson G, Karlson B, Grahn P, et al. Inducing physiological stress recovery with sounds of nature in a virtual reality forest--results from a pilot study. Physiol Behav 2013 Jun 13;118:240-250. [doi: 10.1016/j.physbeh.2013.05.023] [Medline: 23688947]

56. Serrano B, Baños RM, Botella C. Virtual reality and stimulation of touch and smell for inducing relaxation: a randomized controlled trial. Comput Human Behav 2016 Feb;55:1-8 [FREE Full text] [doi: 10.1016/j.chb.2015.08.007]
57. Riches S, Azevedo L, Bird L, Pisani S, Valmaggia L. Virtual reality relaxation for the general population: a systematic review. Soc Psychiatry Psychiatr Epidemiol 2021 Oct 13;56(10):1707-1727 [FREE Full text] [doi: 10.1007/s00127-021-02110-z] [Medline: 34120220]
58. Fassbender E, Heiden W. The virtual memory palace. J Comput Inf Syst 2006;2(1):464 [FREE Full text]
59. Legge EL, Madan CR, Ng ET, Caplan JB. Building a memory palace in minutes: equivalent memory performance using virtual versus conventional environments with the Method of Loci. Acta Psychol (Amst) 2012 Nov;141(3):380-390. [doi: 10.1016/j.actpsy.2012.09.002] [Medline: 23098905]
60. Krokos E, Plaisant C, Varshney A. Virtual memory palaces: immersion aids recall. Virtual Real 2018 May 16;23(1):1-15 [FREE Full text] [doi: 10.1007/s10055-018-0346-3]
61. Knauf M. Space to Reason: A Spatial Theory of Human Thought. Cambridge, MA: MIT Press; 2013.
62. Godden DR, Baddeley AD. Context ‐ dependent memory in two natural environments: on land and underwater. Br J Psychol 2011 Apr 13;66(3):325-331. [doi: 10.1111/j.2044-8295.1975.tb01468.x]
63. Slater M. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. Philos Trans R Soc Lond B Biol Sci 2009 Dec 12;364(1535):3549-3557 [FREE Full text] [doi: 10.1098/rstb.2009.0138] [Medline: 19884149]
64. Barsalou LW. Grounded cognition. Annu Rev Psychol 2008 Jan 01;59(1):617-645. [doi: 10.1146/annurev.psych.59.103006.093639] [Medline: 17705682]
65. Shapiro L. Embodied Cognition. New York, NY: Routledge; 2010.
66. Hartley T, Lever C, Burgess N, O'Keefe J. Space in the brain: how the hippocampal formation supports spatial cognition. Philos Trans R Soc Lond B Biol Sci 2014 Feb 05;369(1635):20120510 [FREE Full text] [doi: 10.1098/rstb.2012.0510] [Medline: 24366125]
67. Varela-Aldás J, Palacios-Navarro G, García-Magariño I, Fuentes EM. Effects of immersive virtual reality on the heart rate of athlete’s warm-up. In: Proceedings of the 6th International Conference on Augmented Reality, Virtual Reality, and Computer Graphics. 2019 Presented at: AVR '19; June 24-27, 2019; Santa Maria al Bagno, Italy p. 175-185 URL: https://link.springer.com/chapter/10.1007/978-3-030-25965-5_14 [doi: 10.1007/978-3-030-25965-5_14]
68. Enzenbach C, Wicklein B, Wirkner K, Loeffler M. Evaluating selection bias in a population-based cohort study with low baseline participation: the LIFE-Adult-Study. BMC Med Res Methodol 2019 Jul 01;19(1):135 [FREE Full text] [doi: 10.1186/s12874-019-0779-8] [Medline: 31262266]
69. Pizzoli SF, Monzani D, Mazzocco K, Maggioni E, Pravettoni G. The power of odor persuasion: the incorporation of olfactory cues in virtual environments for personalized relaxation. Perspect Psychol Sci 2022 May 09;17(3):652-661 [FREE Full text] [doi: 10.1177/17456916211014196] [Medline: 34752166]
70. Abbott RW, Diaz-Artiles A. The impact of digital scents on behavioral health in a restorative virtual reality environment. Acta Astronautica 2022 Aug;197:145-153. [doi: 10.1016/j.actaastro.2022.05.025]

Abbreviations

COPE-NVI: Coping Orientation to the Problems Experienced–Nuova Versione Italiana
DASS-21: Depression, Anxiety, and Stress Scale–21
GI: guided imagery
HMD: head-mounted display
ITC-SOPI: International Test Commission–Sense of Presence Inventory
MANCOVA: multivariate analysis of covariance
PGWBI: Psychological General Well-Being Index
PMRT: progressive muscle relaxation technique
STAI-Y: State-Trait Anxiety Inventory–Form Y
STAI-Y1: State-Trait Anxiety Inventory–Form Y1
VAS: visual analog scale
VR: virtual reality
VRSQ: Virtual Reality Symptom Questionnaire

Edited by J Torous; submitted 02.05.23; peer-reviewed by P Watson, A Todorov; comments to author 11.08.23; revised version received 07.12.23; accepted 28.12.23; published 30.01.24.

Please cite as:
Pardini S, Gabrielli S, Olivetto S, Fusina F, Dianti M, Forti S, Lancini C, Novara C
Personalized Virtual Reality Compared With Guided Imagery for Enhancing the Impact of Progressive Muscle Relaxation Training: Pilot Randomized Controlled Trial
JMIR Ment Health 2024;11:e48649
URL: https://mental.jmir.org/2024/1/e48649
doi:10.2196/48649
PMID:38289673

Review

Effectiveness and User Experience of Virtual Reality for Social Anxiety Disorder: Systematic Review

Simon Shahid[1], BPsych, MPsych; Joshua Kelson[1], PhD; Anthony Saliba[1], PhD

Faculty of Business, Justice, and Behavioural Sciences, Charles Sturt University, Bathurst, Australia

Corresponding Author:
Joshua Kelson, PhD
Faculty of Business, Justice, and Behavioural Sciences
Charles Sturt University
Panorama Ave
Bathurst, 2795
Australia
Phone: 61 2 6338 4570
Email: jkelson@csu.edu.au

Abstract

Background: Social anxiety disorder (SAD) is a debilitating psychiatric disorder that affects occupational and social functioning. Virtual reality (VR) therapies can provide effective treatment for people with SAD. However, with rapid innovations in immersive VR technology, more contemporary research is required to examine the effectiveness and concomitant user experience outcomes (ie, safety, usability, acceptability, and attrition) of emerging VR interventions for SAD.

Objective: The aim of this systematic review was to examine the effectiveness and user experience of contemporary VR interventions among people with SAD.

Methods: The Cochrane Library, Emcare, PsycINFO, PubMed, ScienceDirect, Scopus, and Web of Science databases were searched between January 1, 2012, and April 26, 2022. Deduplicated search results were screened based on title and abstract information. Full-text examination was conducted on 71 articles. Studies of all designs and comparator groups were included if they appraised the effectiveness and user experience outcomes of any immersive VR intervention among people with SAD. A standardized coding sheet was used to extract data on key participant, intervention, comparator, outcome, and study design items.

Results: The findings were tabulated and discussed using a narrative synthesis. A total of 18 studies met the inclusion criteria.

Conclusions: The findings showed that VR exposure therapy–based interventions can generally provide effective, safe, usable, and acceptable treatments for adults with SAD. The average attrition rate from VR treatment was low (11.36%) despite some reported user experience difficulties, including potential simulator sickness, exposure-based emotional distress, and problems with managing treatment delivered in a synchronous group setting. This review also revealed several research gaps, including a lack of VR treatment studies on children and adolescents with SAD as well as a paucity of standardized assessments of VR user experience interactions. More studies are required to address these issues.

Trial Registration: PROSPERO CRD42022353891; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=353891

(JMIR Ment Health 2024;11:e48916) doi:10.2196/48916

KEYWORDS

social anxiety disorder; social phobia; virtual reality; VR; VR exposure therapy; effectiveness; user experience; safety; usability; acceptability; anxiety; phobia; exposure; systematic; review methods; review methodology; social; psychiatric; mental health; mobile phone

Introduction

Background

Social anxiety disorder (SAD; also known as social phobia) is a psychiatric disorder that is distinguished by a fear of humiliation or negative evaluation by others [1]. Current guidelines highlight the use of cognitive behavioral therapy to treat SAD [2]. This can involve activities such as psychoeducation, relaxation, distraction, cognitive restructuring, exposure therapy, and relapse prevention [2]. Despite the efficacy of cognitive behavioral therapy for SAD treatment [3,4], many who are diagnosed do not go on to seek help [5]. This is partly due to the in vivo (real-life) nature of the exposure therapy used to desensitize and habituate patients to feared situations. It takes significant time, effort, and resources to accurately recreate scenarios that will incite an appropriate level of fear response in social settings [6]. For example, to conduct in vivo exposure therapy with an individual with a fear of public speaking, a therapist would need to gather an audience in an appropriate context (ie, ensuring confidentiality and nonjudgment). Furthermore, social anxiety–provoking environments can be unpredictable, providing therapists with little control and a higher chance that a patient is embarrassed, leading to higher attrition rates [7]. To potentially overcome these issues with delivering in vivo exposure therapy, some researchers have examined the use of virtual reality (VR) technology [7].

VR Technology

VR technology provides a digital modality to deliver psychological interventions [7,8]. It involves the use of computer hardware and software technology (eg, stereoscopic displays of digital environments) to simulate real-world experiences [7]. For instance, one may enter a virtual environment that mimics a physical environment and could adopt a virtual avatar to interact with this virtual environment [9]. VR was first formulated in the 1960s, with the first commercial device developed in the 1980s [10]. As technology has developed, the quality of images has improved, and costs have been reduced.

VR systems can be divided into 2 categories: immersive and nonimmersive systems [11]. Immersive systems, such as head-mounted displays (HMDs) or cave automatic virtual environments (CAVEs), provide users with a realistic experience of VR environments, whereas nonimmersive systems, such as computer monitors, result in users not feeling as present [12]. Presence in VR refers to the extent of an individual's perception of being in a particular environment [13]. For VR therapy to be effective, an individual must feel present and immersed in the digital environment [14]. A CAVE system consists of an empty room with multiple screens arranged in a cubelike formation with users wearing stereoscopic glasses and interacting with virtual objects projected onto the screens [15]. Although CAVE systems have the potential to be more immersive than nonimmersive systems, they are expensive and complex to set up, require frequent physical and digital adjustments, and require dedicated personnel [16]. Conversely, a recent systematic review found that current HMDs offer a more immersive experience than CAVEs and are significantly more user-friendly in cost and setup, with a "plug-n-play" setup solution [15].

When integrated with therapy, VR technology can help address factors that influence the success of exposure-based treatments. For instance, VR allows for the creation of controlled digital environments, which enables therapists to predictably customize exposure scenarios to the specific needs and fears of individual clients [7]. VR can also improve accessibility to exposure therapy for individuals who find it logistically challenging or emotionally overwhelming to engage in real-world scenarios [17]. The immersive nature of VR helps bridge the gap between simulated experiences and real-life situations, fostering a sense of presence and engagement that can potentially enhance treatment adherence and effectiveness [17].

Effectiveness of VR for SAD

Virtual environments and avatars can be used to simulate socially distressing situations for SAD treatment. For example, a study immersed participants with SAD into a computer-generated classroom where they were asked to speak publicly on a topic while a therapist controlled the virtual audience's reactions according to the stage of therapy [18]. VR environments have also been shown to provide acceptable levels of presence and immersion that are necessary for exposure therapy in youth with social anxiety [19]. Several systematic reviews and meta-analyses have demonstrated the effectiveness of VR exposure therapy (VRET) in the treatment of SAD [6,20-24]. Indeed, researchers have established a large effect size for VRET versus waitlist (g=0.90), a medium to large effect size for VRET versus psychological placebo conditions (g=0.78) [21], a large overall effect size for VRET (g=0.82) [22], and a medium to large effect size for VRET at the 12-month follow-up (g=−0.74) [6]. This consistent pattern of symptom reduction can be observed across various contexts, such as participant countries (eg, the United States, France, Israel, and South Korea) and treatment settings (eg, universities, hospitals, and clinics) [6,20-24]. However, although existing reviews have explored VR-based therapy from an effectiveness standpoint (eg, reduction in anxiety symptoms), there are gaps in the literature on evaluating the VR user experience for people with SAD on key concomitant outcomes of safety, usability, acceptability, and attrition in different contexts.

User Experience of VR for SAD

Safety

Studies using VR for workplace training, physical rehabilitation, psychological therapy, and other settings highlight a significant safety issue: simulator sickness [25]. Simulator sickness (otherwise known as VR sickness or cybersickness) [26-28] is characterized by general discomfort, headache, eyestrain, nausea, difficulty concentrating, fatigue, blurred vision, dizziness, and vertigo. On the basis of postural instability theory, simulator sickness is arguably due to VR technology inducing sensory differences in the visual and vestibular systems, which coordinate balance and movement [28-30]. The human body may interpret these disparities as possibly deadly causes (ie, consuming poison) and seek to purge as a result [25]. Consequently, simulator sickness can have a negative impact on participants during VR use and for hours following use [31]. Other aspects of safety include physical injuries from repetitive strain, users colliding with objects in the real world, poor posture, headset discomfort, risk of inducing epileptic seizures, negative mood changes, and infection control [32]. Overall, these issues might put participants at risk of harm or cause them to discontinue using VR. Thus, a comprehensive examination of VR safety for SAD is necessary.

Usability

There does not yet appear to be a framework for the evaluation of VR usability in therapy-based applications. Nielsen [33] defines usability as a "quality attribute" that assesses how easy it is to interact with an interface. He highlighted 5 components: learnability (how easy it is for a beginner to use the interface), efficiency (once the user has learned to use the interface, how quickly they can perform tasks), memorability (re-establishing proficiency after a period of absence), errors (frequency, severity, and recoverability of errors), and satisfaction (level of pleasure from using the interface) [33]. Although this framework is applied to website design, it can also be applied to participants' perceptions of the usability of VR. Furthermore, the application of VR in real-world settings (eg, in a therapy room) would likely be performed by a clinician rather than a specialized technician. It is important to note that usability issues may arise among clinicians. For example, they may give up on the technology if components fail to load or connect to each other. For this reason, this review included both clinician experiences in administering VR-based therapy and client experiences.

Acceptability

Acceptability is a crucial consideration when evaluating VR interventions [34]. It involves assessing the degree to which the new intervention and its components are received and aligned with the needs of the target population [34]. For example, a study examining VR use in adults with SAD defined acceptability as a participant's willingness to use a VR program [35]. They measured acceptability by observing rates of attrition and responses to the following question—"Would you recommend this program to others who might have problems similar to yours?"—and inviting further feedback. Participants' additional feedback was coded into 2 themes: satisfaction (sense of realism, insight, and utility) and perceived effects of the treatment (impact on anxiety). The findings indicated that VR was considered acceptable by participants on all measures [35]. Nevertheless, although there are recent systematic reviews that have addressed the acceptability of VR use for the general population [36], psychosis [37], panic disorder [38], and posttraumatic stress disorder [39,40], a review of the literature on the acceptability of VR in individuals with SAD does not yet exist based on our current knowledge.

Attrition

Attrition, the discontinuation of therapy before treatment completion and resolution of symptoms, can have profound negative effects [41]. These can include the client not fully benefiting from therapy and being discouraged from seeking treatment in the future [41] as well as the effect that this may have on the therapist (eg, loss of revenue, demoralization, and feelings of failure) [42]. A recent meta-analysis of VRET showed significant heterogeneity in attrition rates in the treatment of anxiety disorders, highlighting reasons such as failure to immerse in the virtual environments, simulator sickness, vision complications, and difficulty communicating with a therapist that the participant could not see [43]. A systematic review examining the available literature on rates of attrition of VR-based interventions (both VRET and non-VRET) with participants with SAD does not yet appear to exist based on our current knowledge.

This Study

This study aimed to systematically identify and review available evidence regarding the effectiveness and user experience (ie, safety, usability, acceptability, and rates of attrition) of VR interventions in the treatment of SAD. The following objectives aided in the provision of a comprehensive and up-to-date account of the empirical status of VR therapy for SAD: (1) provide an overview of the existing literature and identify areas in which further research is needed on the treatment of SAD; (2) assess the potential of using VR as a treatment option for SAD, specifically in terms of effectiveness and user experience; and (3) provide guidance and recommendations for future research regarding the use of VR as a treatment option for SAD.

Methods

This systematic review was conducted using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist (Multimedia Appendix 1) [44]. The protocol of this systematic review was prospectively registered in the PROSPERO international database (CRD42022353891).

Eligibility Criteria

For articles to be included in this systematic review, the study participants needed to be people diagnosed with SAD regardless of age. If a study had a mix of people with and without SAD, that study would be included if subgroup analyses were available on the participants with SAD or if they made up the vast majority (ie, ≥80%). All studies needed to examine direct participant use of a VR intervention, which includes any system that incorporates immersive VR hardware (ie, HMD or CAVE systems). Only studies that were published after 2012 were included as this marked the introduction of widely available commercial HMD hardware such as the Oculus Rift [9]. Such hardware allowed for the delivery of VR experiences comparable with previously expensive commercial setups at a cheaper cost as well as easier accessibility to researchers [45]. Studies with research design comparators of any kind (eg, comparing VR with other non-VR interventions) were eligible for inclusion. All studies were required to report on VR intervention effectiveness and participant user engagement outcomes. This broadly included any standardized or unstandardized measure indicative of usability or acceptability (including attrition rates). Studies of all designs (ie, quantitative, qualitative, and mixed methods) were eligible for inclusion. No studies were excluded based on methodological quality. All the articles needed to be written in English and published in peer-reviewed journals.

Search Strategy

Prominent scientific research databases were searched between January 1, 2012, and April 26, 2022: Cochrane Library, Emcare, PsycINFO, PubMed, ScienceDirect, Scopus, and Web of Science. The following keywords were used to search the databases: ("virtual reality" or "VR") and ("social anxiety" or "social phobia"). The reference lists of eligible articles were also searched.

Article Selection

The search results for all databases were deduplicated, and the remaining article titles and abstracts were scanned. Full-text appraisal was performed on promising articles, and the final study inclusion was agreed upon by the researchers using the eligibility criteria. Divergent views on inclusion were resolved through discussion and mutual agreement.

Data Extraction

Data from the included studies were extracted by one reviewer (SS) into a standardized coding sheet and then checked by a second reviewer (JK). The data types extracted from eligible papers included the following:

1. Reference source: author surnames, year of publication, and paper title.
2. Sample: country; sample size; and nonidentifiable participant characteristics such as age, sex, and diagnosis.
3. Study design: methodology, comparator trial arms, and measurement points (pretest, midtest, and posttest measurement and follow-up).
4. VR intervention details: intervention program name, purpose of intervention (eg, exposure therapy, cognitive distraction, or relaxation), virtual environment type, hardware (eg, HMD or CAVE system), and treatment length.
5. Effectiveness: standardized measure names, outcomes, and effect sizes.
6. User experience: reported outcomes of intervention safety, usability, acceptability, attrition, and intention-to-treat analyses.

Attrition in this review was defined and measured as the relative number of participants who began using the VR intervention but did not complete measurements during or after the intervention.

Quality Assessment

This systematic review included randomized controlled trial (RCT) and nonrandomized studies. Therefore, the Mixed Methods Appraisal Tool (MMAT) was used to assess the quality of all the included studies [46]. The MMAT was used as it assesses methodological quality across 5 study categories: RCTs, nonrandomized quantitative studies, quantitative descriptive studies, qualitative studies, and mixed methods studies.

Data Analysis

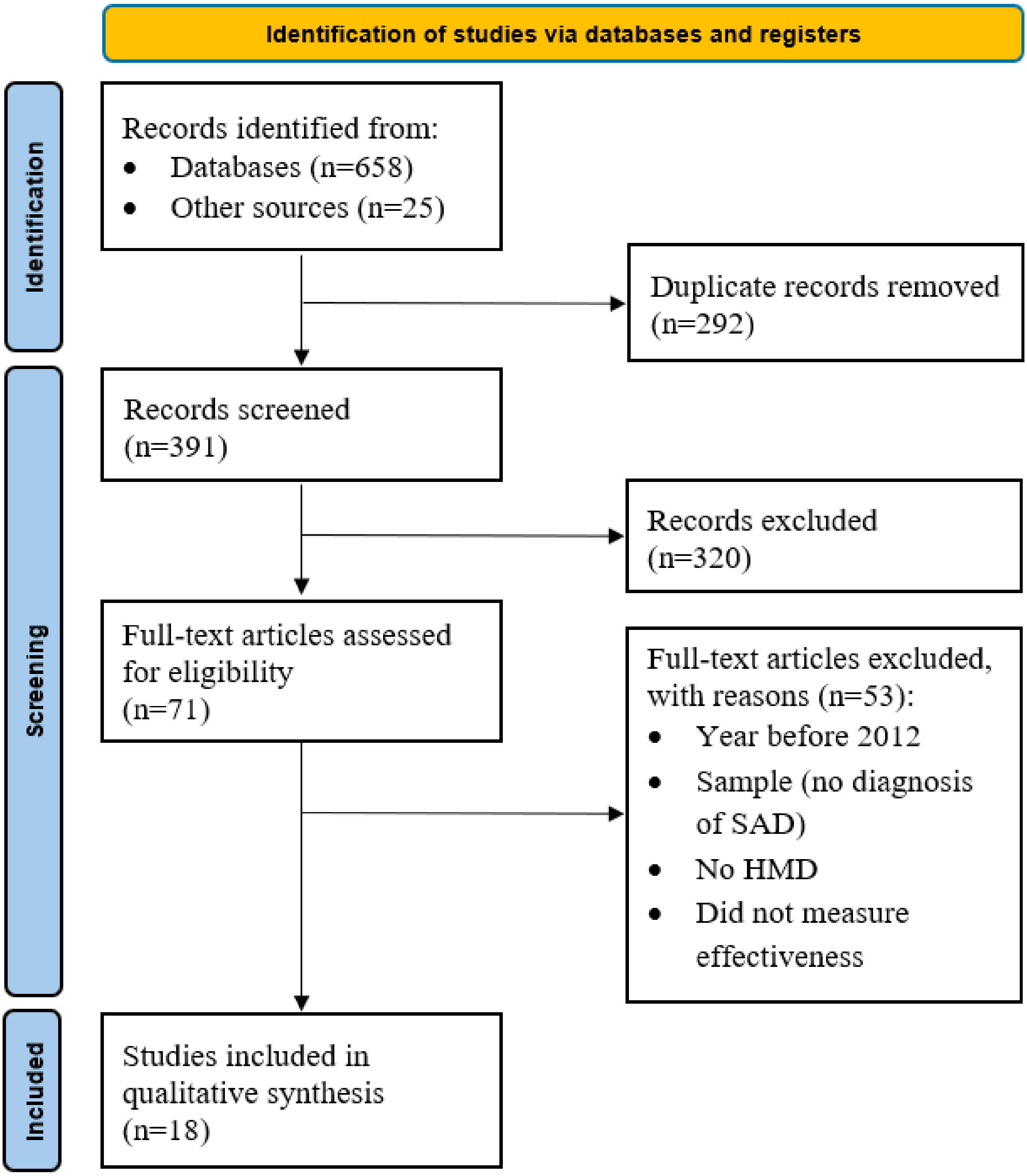
A narrative synthesis approach was used in this systematic review. This involved summarizing and explaining the findings using text as a statistical meta-analysis was not possible because of data heterogeneity across the included studies.

Results

Study Selection

Figure 1 shows that the literature search yielded 683 articles, of which 391 (57.2%) remained after deduplicating citations. Of these 391 records, 18 (4.6%) met the eligibility criteria.

Figure 1. Flowchart of the systematic review search results. HMD: head-mounted display; SAD: social anxiety disorder.



Participant Characteristics

A total of 808 participants were recruited for the VR studies (Table 1). They were largely from South Korea (n=368), followed by the United States (n=163), the Netherlands (n=60), Canada (n=59), Sweden (n=23), Czech Republic (n=10), Denmark (n=9), and Brazil (n=2). The country of origin was missing for some participants (n=114). It is unclear whether participants were unique in 11% (2/18) of studies conducted by research teams with some of the same researchers [47,48]. The sample sizes ranged from 1 to 115 participants, with a median of 48 participants. Participants' ages ranged from 18 to 65 years. Participants were mainly female, with an average sample proportion of 51.31% (SD 5.36%; range 0%-77.3%). Most participants were diagnosed with SAD. There were 9 participants with a diagnosis of flight phobia and 8 participants with a diagnosis of acrophobia; however, subgroup analyses were available for the participants with SAD in this study [49]. All participant diagnoses were obtained through clinical interviews delivered in person, by phone, or via videoconferencing. In total, 3 therapists were interviewed in addition to the sample of participants in one study [50].

Table 1. Description of participants and research designs in the reviewed studies.

Study	Country	Sample size[a]	Total sample mean age (years; SD)	Total sample age range (years)	Total sample percentage of female participants	Study design	Treatment conditions[b]	Measurements
Anderson et al [51]	United States	97	39.03 (11.26)	19-69	61.9	RCT[c]	EGT[d]: 25; VRET[e]: 25; WL[f]: 25	Pre- and posttest measurement and 3- and 12-month FU[g]
Arnfred et al [50]	Denmark	9	25.4 (6.54)	19-35	66.7	Interview	VRET: 9	Posttest measurement
Bouchard et al [52]	Canada	59	34.5 (11.9)	18-65	72.9	RCT	VRET: 17; in vivo: 22; WL: 20	Pre- and posttest measurement and 6-month FU
Geraets et al [53]	NR[h]	15	34.9 (12.4)	18-65	53.3	Single group	VRET: 15	Pre- and posttest measurement and 6-month FU
Hur et al [54]	South Korea	73	NR	NR	42.6	Case controlled	VRET: 25; HC[i]: 22	Pre- and posttest measurement
Jeong et al [55]	South Korea	115	NR	NR	34.8	Cohort	ET[j]: 52; NT[k]: 43; SE[l]: 20	Pre- and posttest measurement
Kampmann et al [56]	The Netherlands	60	36.9	18-65	63.3	RCT	VRET: 20; WL: 20; iVET[m]: 20	Pre- and posttest measurement and 3-month FU
Kim et al [47]	South Korea	54	23	NR	57.7	Controlled clinical trial	VRET: 22; HC: 30	Pre- and posttest measurement
Kim et al [57]	South Korea	74	NR	19-31	56.9	Longitudinal	VRET: 32; HC: 33	Pre- and posttest measurement
Kim et al [48]	South Korea	52	NR	19-30	NR	RCT	VRS[n]: 24; WL: 28	Pre- and posttest measurement
Kovar [58]	Czech Republic	10	34.6 (11.7)	19-51	50	Nonrandomized parallel comparison trial	Psychotherapy: 5; psychotherapy+VRET: 5	Pre- and posttest measurement
Lindner et al [59]	Sweden	23	40.61 (10.15)	≥18	57	Cohort	VRET: 23	Pre- and posttest measurement
Moldovan and David [49]	NR	32	NR	≥18	46.9	RCT	VRCBT[o]: 16; WL: 16	Pre- and posttest measurement and FU
Perandré and Haydu [60]	Brazil	2	23.5 (4.9)	20-27	0	Case study	VRET: 2	Pre- and posttest measurement and 1- and 3-month FU
Price and Anderson [61]	NR	67	40.31 (11.55)	NR	69	RCT	VRET: 33; EGT: 34	Pre-, mid-, and posttest measurement
Rubin et al [62]	United States	21	NR	18-65	61.9	RCT	VRET: 10; VRET+AGT[p]: 11	Pre- and posttest measurement and 1-week FU
Trahan et al [63]	United States	1	36	NR	0	Case study	VRET: 1	Pre- and posttest measurement
Zainal et al [64]	United States	44	23.3 (9.32)	18-53	77.3	RCT	VRET: 26; WL: 18	Pre- and posttest measurement and 3- and 6-month FU

[a]Refers to the total number of participants in the study.

[b]Refers to the number of participants in each treatment condition.

[c]RCT: randomized controlled trial.
[d]EGT: exposure group therapy.
[e]VRET: virtual reality exposure therapy.
[f]WL: waitlist control.
[g]FU: follow-up.
[h]NR: not reported.
[i]HC: healthy controls.
[j]ET: early termination.
[k]NT: normal termination.
[l]SE: session extension.
[m]iVET: in vivo exposure therapy.
[n]VRS: virtual reality self-training.
[o]VRCBT: virtual reality cognitive behavioral therapy.
[p]AGT: attention guidance training.

Details of the VR Interventions

All the studies included VR-based exposure therapy. Nearly all the studies (15/18, 83%) tested a unique VR intervention except for the studies by Jeong et al [55] and Kim et al [47,48] (Table 2). VR hardware included standard computers, smartphones, and HMDs. In total, 17% (3/18) of the studies [49,51,61] did not identify the headset brands. The studies used custom-built software that immersed participants in VR environments that simulated social situations increasing in difficulty with audio, video, text, and interactivity. The treatment lengths ranged from 1 to 14 sessions of exposure therapy, with a mode of 8. A total of 17% (3/18) of the studies [53,55,64] terminated the sessions early if habituation occurred before the completion of the sessions. Participant VR use time ranged from 5 minutes to 3 hours per session, and 17% (3/18) of the studies delivered the VR in a single session [49,59,62]. All VR interventions were tested with therapist or facilitator guidance even though 22% (4/18) [47,48,55,64] were designed to be delivered as self-help.

Table 2. Details of the virtual reality (VR) interventions.

Study	Virtual environments	Headset	Treatment length (duration)
Anderson et al [51]	Conference room, classroom, and auditorium	—[a]	4 exposure sessions (30 min each)
Arnfred et al [50]	Supermarket, meeting, cafeteria, party, and auditorium	Oculus Go	8 exposure sessions (45 min each)
Bouchard et al [52]	Meeting room, job interview, apartment, coffee shop, neighbors, store, and neutral	eMagin Z800	8 exposure sessions (20-30 min each)
Geraets et al [53]	Street, bus, café, and supermarket	Sony HMZ-T1	14 exposure sessions (40 min each)
Hur et al [54]	College student group	HTC Vive	6 exposure sessions (5-8 min each)
Jeong et al [55]	School, business, and daily life	Samsung Gear VR powered by Oculus	ET[b] (1-8 exposure sessions); NT[c] (9-10 exposure sessions); SE[d] (11-17 exposure sessions)
Kampmann et al [56]	Audience, stranger, clothes shopping, job interview, journalist interview, restaurant, and blind date	nVisor SX	7 exposure sessions (60 min each)
Kim et al [47]	School, business, and daily life	Samsung Gear VR powered by Oculus	8 exposure sessions
Kim et al [57]	College student group	HTC Vive	6 exposure sessions
Kim et al [48]	School, business, and daily life	Samsung Gear VR powered by Oculus	8 exposure sessions
Kovar [58]	Public speaking, telephone call, receiving criticism, job interview, refusal of job offer or unwanted product, and working lunch	HTC Vive	8 exposure sessions
Lindner et al [59]	Board room, conference room, and classroom	Oculus Go	1 exposure session (180 min)
Moldovan and David [49]	Presentation and interview	—	1 exposure session (90 min)
Perandré and Haydu [60]	Food court in shopping center	Oculus Rift	8 exposure sessions
Price and Anderson [61]	Conference room, classroom, and auditorium	—	8 exposure sessions
Rubin et al [62]	Conference room and auditorium	Oculus Rift DK2	1 exposure session (45 min)
Trahan et al [63]	Grocery store	Plastic HMD[e] bracket for mobile phone	12 exposure sessions (12-15 min each)
Zainal et al [64]	Dinner party and job interview	Pico Goblin VR	8 exposure sessions (25-30 min each)

[a]Brand not reported.

[b]ET: early termination.

[c]NT: normal termination.

[d]SE: session extension.

[e]HMD: head-mounted display.

Research Designs and Comparators

Table 1 summarizes the research designs and comparators. Nearly half (8/18, 44%) of the studies appraised participant VR use through RCT designs. Comparators included exposure group therapy, in vivo exposure, early and extended termination, attention guidance training using VR, psychotherapy, and waitlist control. All studies (18/18, 100%) had pre- and posttest assessments of user outcomes, although 39% (7/18) also had follow-up assessments, with the longest being 12 months [51].

Effectiveness Measures and Outcomes

The details of the effectiveness measures and outcomes of VR treatment for SAD are summarized in Table 3. VR treatment effect sizes across all studies that reported them ranged from medium to large. Almost all studies (15/18, 83%) demonstrated a decrease in symptoms following VR treatment.

Table 3. Details on social anxiety measures and virtual reality (VR) effectiveness outcomes.

Study and measures	VR effectiveness outcomes
Anderson et al [51]	
PRCS[a]	Significant improvement in confidence as a speaker from before to after treatment (*d*[b]=1.19; *P*=.01), with benefits maintained at the 3- and 6-month follow-ups.
FNE-B[c]	Significant decrease in fear of negative evaluation from before to after treatment (*d*=0.29; *P*=.01), with benefits maintained at the 3- and 6-month follow-ups.
BAT[d]	Significant improvement in speech length (*d*=0.78; *P*=.01) and peak anxiety (*d*=0.70; *P*=.02) from before to after treatment.
Arnfred et al [50]	
NSQ[e]	The virtual environments effectively induced immersion and anxiety in some but not all participants with social anxiety disorder.
Bouchard et al [52]	
LSAS-SR[f]	Significant decrease in social anxiety symptoms from before to after treatment compared with waitlist (*P*<.001) that was maintained at the 6-month follow-up.
BAT	Significant decrease in behavioral avoidance from before to after treatment (*P*<.001).
SPS[g]	Significant decrease in social phobia from before to after treatment (*P*<.001) that was maintained at the 6-month follow-up.
SIAS[h]	Significant decrease in social anxiety symptoms from before to after treatment (*P*<.001) that was maintained at the 6-month follow-up.
FNE[i]	Significant decrease in fear of negative evaluation from before to after treatment (*P*<.001) that was maintained at the 6-month follow-up.
Geraets et al [53]	
SIAS	Significant decrease in social interaction anxiety from before to after treatment (*d*=0.9; *P*=.008) that was maintained at the 6-month follow-up (*d*=1.3; *P*=.003).
Hur et al [54]	
SPS	Significant decrease in social phobia symptoms from before to after treatment (*P*=.005).
PERS[j]	Significant decrease in negative postevent rumination from before to after treatment (*P*<.001).
Jeong et al [55]	
FNE-B	Significant decrease in fear of negative evaluation from first to last session for the early, normal, and extended termination groups (*P*<.001).
LSAS	Significant decrease in social anxiety symptoms from first to last session for the normal (*P*<.001) and extended termination groups (*P*=.002).
SPS	Significant decrease in social phobia symptoms from first to last session for the normal (*P*=.001) and extended termination groups (*P*<.001).
SIAS	Significant decrease in social interaction anxiety from first to last session for the normal (*P*<.001) and extended termination groups (*P*=.006).
Kampmann et al [56]	
LSAS-SR	Significant decrease in social anxiety symptoms from before to after treatment compared with waitlist (*d*=0.55; *P*=.01) that was maintained at the 3-month follow-up.
FNE-B	No significant change in fear of negative evaluation compared with waitlist group from before to after treatment or the 3-month follow-up.
BAT	Significant increase in speech length from before to after treatment compared with waitlist (*d*=0.56; *P*=.02) that was maintained at the 3-month follow-up; however, there was no significant difference in speech performance.
Kim et al [47]	
HADS[k]	Significant decrease in anxiety symptoms from before to after treatment (*P*<.001).
LSAS-SR	Significant decrease in social anxiety symptoms from before to after treatment (*P*<.001).
SIAS	Significant decrease in social interaction anxiety from before to after treatment (*P*<.001).
Kim et al [57]	
SPS	Significant decrease in social phobia symptoms from before to after treatment (*P*<.001).

Study and measures	VR effectiveness outcomes
SIAS	Significant decrease in social interaction anxiety from before to after treatment (*P*<.001).
FNE-B	Significant decrease in fear of negative evaluation from before to after treatment (*P*=.004).
KSAD[l]	Significant decrease in social avoidance and distress from before to after treatment (*P*<.001).
LSAS[m]	Significant decrease in social anxiety symptoms from before to after treatment (*P*=.04).
Kim et al [48]	
HADS	No significant changes in anxiety symptoms from before to after treatment.
LSAS-SR	Significant decrease in social anxiety symptoms from before to after treatment (*P*<.01).
Kovar [58]	
FNE-B, SPIN[n], SIAS, SADS[o], and BAI[p]	Higher average decrease in symptoms on all these measures in the VR treatment group compared with the non-VR treatment group.
Lindner et al [59]	
PSAS[q]	Significant decrease in self-rated public speaking anxiety following the first 3-hour session (*d*=0.77; *P*=.006).
LSAS-SR	Significant decrease in social anxiety symptoms from before to after treatment (*P*=.001).
FNE-B	Significant decrease in fear of negative evaluation from before to after treatment (*P*=.04).
Moldovan and Price [49]	
FNE-B	Significant decrease in fear of negative evaluation from before to after treatment (*P*<.05).
SSPS[r]	Significant decrease in negative self-statements from before to after treatment (*P*<.05).
LSAS	Significant decrease in social anxiety symptoms from before to after treatment (*P*<.05).
Perandré and Haydu [60]	
SPIN and BAI	Decrease in anxiety symptoms reported on both measures from the pretest measurement to the 3-month follow-up from treatment for both participants.
Price and Anderson [61]	
SSPS	Significant improvements on positive and negative self-statements from before to after treatment (*P*<.01).
PRCA-SF[s]	Significant decrease in public speaking anxiety from before to after treatment (*P*<.01).
Rubin et al [62]	
PRPSA[t]	Significant decrease in fear of public speaking from before to after treatment (*d*=−1.11) and at the 1-week follow-up (*d*=−1.68).
LSAS-SR	Significant decrease in general symptoms of social anxiety from before to after treatment (*d*=−0.60) and at the 1-week follow-up (*d*=−2.07).
Trahan et al [63]	
SUDS[u]	No significant change in subjective distress from before to after treatment for the participant (*P*=.21).
SADS	Score decrease of 52.6% in social anxiety from before to after treatment for the participant.
Zainal et al [64]	
SAD composite[v]	Significant decrease in social anxiety symptoms from before to after treatment compared with the waitlist group (*g*[w]=−4.77; *P*<.001). No significant within-group change at the 3- (*g*=0.12) and 6-month follow-ups (*g*=−0.13).

Study and measures	VR effectiveness outcomes
MASI[x]	Significant decrease in job interview anxiety from before to after treatment compared with the waitlist group (*g*=−4.17; *P*<.001). No significant within-group change at the 3- (*g*=−0.10) and 6-month follow-ups (*g*=−0.53).

[a]PRCS: Personal Report of Confidence as a Speaker.

[b]Cohen *d* effect size.

[c]FNE-B: brief Fear of Negative Evaluation Scale.

[d]BAT: behavioral avoidance task.

[e]NSQ: nonstandardized questions.

[f]LSAS-SR: Liebowitz Social Anxiety Scale–Self-Report.

[g]SPS: Social Phobia Scale.

[h]SIAS: Social Interaction Anxiety Scale.

[i]FNE: Fear of Negative Evaluation Scale.

[j]PERS: Post-Event Rumination Scale.

[k]HADS: Hospital Anxiety and Depression Scale.

[l]KSAD: Korean Social Avoidance and Distress Scale.

[m]LSAS: Liebowitz Social Anxiety Scale.

[n]SPIN: Social Phobia Inventory.

[o]SADS: Social Avoidance and Distress Scale.

[p]BAI: Beck Anxiety Inventory.

[q]PSAS: Public Speaking Anxiety Scale.

[r]SSPS: Self-Statements During Public Speaking scale.

[s]PRCA-SF: Personal Report of Communication Apprehension–Short Form.

[t]PRPSA: Personal Report of Public Speaking Anxiety.

[u]SUDS: Subjective Units of Distress Scale.

[v]SAD composite: average standardized scores of the Social Phobia Diagnostic Questionnaire and the SIAS.

[w]Hedges *g* effect size.

[x]MASI: Measure of Anxiety in Selection Interviews.

User Experience With the VR Interventions

The average attrition rate was 11.36% across all studies in the active VR treatment phase, with a range of 0% to 45.2% (Table 4). A total of 22% (4/18) of the studies reported the use of an intention-to-treat analysis. To measure VR user experience, 56% (10/18) of the studies used standardized measures, and 11% (2/18) of the studies used nonstandardized questions. A total of 67% (8/12) of these studies reported positive VR user experience findings in various areas of presence, usability, acceptability, or satisfaction. Low levels of simulator sickness were reported in 75% (3/4) of the studies that used standardized questions; however, 25% (1/4) of these studies reported higher levels of simulator sickness in participants with SAD than in controls without SAD [47]. No other safety issues, such as physical injury, user collision, postural complaints, headset discomfort, seizures, or infection, were reported.

Table 4. Virtual reality (VR) interventions and user experience outcomes.

Study	Measures	VR user experience findings	Attrition (%)	ITT[a]
Anderson et al [51]	CSQ[b]	High satisfaction with VR was reported after treatment and maintained at the 12-month follow-up.	5/30 (17)	Yes
Arnfred et al [50]	NSQ[c]	A high level of presence in virtual environments for some participants but not all. There were technical issues with setting up and storing away equipment for the group. Wearing the HMD[d] in front of strangers was more anxiety provoking than the virtual environments for some participants. All patients found VR to be a meaningful addition to their therapy sessions, with several wanting more exposure.	0/9 (0)	—[e]
Bouchard et al [52]	SSQ[f], PQ[g], and GPQ[h]	No significant increases in simulator sickness after exposure sessions (*P*>.20). Good level of presence that increased with a higher number of exposures.	2/17 (12)	Yes
Geraets et al [53]	—	VR treatment was well tolerated and deemed acceptable for most participants.	2/15 (13)	—
Hur et al [54]	—	—	16/73 (21)	—
Jeong et al [55]	—	—	52/115 (45)	—
Kampmann et al [56]	—	Simulator sickness led one patient to drop out.	5/20 (25)	Yes
Kim et al [47]	SSQ	Participants with SAD[i] experienced significantly more simulator sickness than participants without SAD (*P*=.003).	2/54 (4)	—
Kim et al [57]	—	—	9/74 (12)	—
Kim et al [48]	SSQ	Low levels of simulator sickness.	3/24 (13)	—
Kovar [58]	—	—	0/10 (0)	—
Lindner et al [59]	NEQ[j]	High stress levels and low levels of satisfaction.	3/23 (13)	Yes
Moldovan and David [49]	ITQ[k] and PQ	No moderating effect of immersion and presence on pre- and posttest anxiety.	0/32 (0)	—
Perandré and Haydu [60]	SPI[l]	High sense of presence reported by both participants.	0/2 (0)	—
Price and Anderson [61]	—	From a randomly selected subset of videotaped sessions (14%), high participant compliance was found, with 92% of the VR treatment protocol being completed.	0/33 (0)	—
Rubin et al [62]	—	—	2/21 (10)	—
Trahan et al [63]	SUS[m]	High usability reported by the participant.	0/1 (0)	—
Zainal et al [64]	NSQ, IPQ[n], and SSQ	Acceptable presence and low levels of simulator sickness. High levels of homework compliance. Participants (85%) would recommend it to others with SAD. High levels of acceptability and usability.	9/44 (21)	—

[a]ITT: intention-to-treat analysis.
[b]CSQ: Client Satisfaction Questionnaire.
[c]NSQ: nonstandardized questions.
[d]HMD: head-mounted display.
[e]Not reported.
[f]SSQ: Simulator Sickness Questionnaire.
[g]PQ: Presence Questionnaire.
[h]GPQ: Gatineau Presence Questionnaire.
[i]SAD: social anxiety disorder.
[j]NEQ: Negative Effects Questionnaire.
[k]ITQ: Immersive Tendencies Questionnaire.
[l]SPI: Sense of Presence Inventory.
[m]SUS: System Usability Scale.
[n]IPQ: Igroup Presence Questionnaire.

Quality Assessment Results

Multimedia Appendix 2 [47-64] contains a table of quality assessment results for the included studies. In all RCT studies [48,49,51,52,56,61,62,64], randomization was reported, but schedule details were unclear in 11% (2/18) of the studies [48,61]. All RCT studies reported comparable baseline group analyses. In total, 38% (3/8) of the RCT studies reported complete outcome data, which is defined as ≥80% [49,56,64]. All but the RCT studies by Kim et al [48], Moldovan and David [49], Price and Anderson [61], and Rubin et al [62] reported blinding of outcome assessors, which was applied at the pretest measurements. All RCT studies except those by Bouchard et al [52], Kampmann et al [56], and Rubin et al [62] reported that participants adhered to their assigned VR interventions.

In the quantitative descriptive studies [53,60,63], the sampling strategy was relevant to the research question except in 33% (1/3) of the studies, in which details were unclear [53]. All quantitative descriptive study samples were representative of the target population, and the measures fulfilled the inclusion criteria. Nonresponse bias was low in all studies except one (2/3, 67%) [60]. Statistical analyses were appropriate to answer the research question in 33% (1/3) of the studies [53] but unclear in the other 2 [60,63].

In the quantitative nonrandomized studies [47,54,55,57-59], participants were representative of the target population, measurements were appropriate regarding both the outcome and intervention, and there were complete outcome data (defined as ≥80%) in all but 2 studies (4/6, 67%) [54,55]. Confounds were accounted for in the design and analysis of 50% (3/6) of the studies [54,55,59]. In total, 67% (2/3) of the quantitative nonrandomized studies reported that the intervention was administered as intended [54,55].

In the single qualitative interview study [50], the qualitative approach was appropriate to answer the research question; the data collection methods were adequate to address the research question; findings were adequately derived from the data; the interpretation of the results was sufficiently substantiated by the data; and there was coherence between qualitative data sources, collection, analysis, and interpretation.

Discussion

Principal Findings

Overview

SAD is a common and debilitating anxiety disorder that affects occupational and social functioning [2]. Current in vivo–based exposure therapies require significant time, resources, and effort, which results in limited treatment dissemination [6]. VR technology provides an alternative modality for treating SAD [19]; however, contemporary evidence on the user experience of VR for SAD is sparse. This systematic review was conducted to provide a comprehensive and up-to-date account of the available evidence regarding the effectiveness and user experience (ie, safety, usability, acceptability, and attrition) of VR interventions for the treatment of SAD.

Effectiveness of VR Interventions for SAD

Our review found that VR interventions can effectively treat SAD in adult populations, which is congruent with the existing literature [6,20-24]. It is interesting to note that, although our search terms and inclusion criteria were open to any VR-based intervention for treating people with SAD (eg, providing relaxation, cognitive distraction, exposure therapy, and psychoeducation), all the included interventions were intended for exposure therapy. This indicates that VRET dominates the research field of VR-based interventions for SAD.

Studies including follow-up measures highlight the maintenance of SAD symptom improvement from 1 week [62] to 1 year [51], indicating that VRET can provide effective short- and long-term treatment for SAD symptoms. This is impressive given that the study showing maintained benefits for up to 1 year involved only 4 treatment sessions [51]. However, it is important to note that this study only included participants with a fear of public speaking as the primary social fear as opposed to other social situations (eg, going to dinner with friends), limiting the generalizability of the findings [65]. Nevertheless, our findings suggest that VRET can be a rapidly effective treatment for SAD with the potential to provide long-term symptom improvement.

Safety of VR Interventions for SAD

Simulator sickness was a common measure of safety in the reviewed studies. Participant simulator sickness was reportedly low in most studies. However, it was found that participants with an SAD diagnosis were more prone to simulator sickness when compared with participants without SAD in one study [47]. This could be because patients with anxiety tend to experience greater motion discomfort [66,67]. For example, patients with anxiety may be more susceptible to irregular breathing and hyperventilation, leading to dizziness and nausea when exposed to fear-inducing cues. This may exacerbate the body's interpretation of disparities in visual and vestibular systems as possible deadly causes (ie, poison) and potentially lead to nausea and vomiting [25]. Another safety consideration is the absence of other physical injuries (eg, collisions with real-world objects, poor posture, headset discomfort, and seizures) reported in the reviewed studies, which supports VR as a safe SAD treatment.

However, although the research safety findings are encouraging, the limitations of these studies are important to note. For example, most studies screened out participants who were unable to tolerate the VR environment and HMD or those who had a history of seizures. This would result in a sampling bias in favor of VR safety. Furthermore, all studies except for one [63] were conducted in controlled settings (ie, hospitals and clinics) that were supervised by clinicians, further reducing risks that would otherwise be significant when using VR alone. For example, an individual purchasing and using a VR system at home may collide with real-world objects without the intervention of a third party. As such, more research is required to evaluate the safety of VR for SAD in nonclinical, unsupervised settings.

Usability of VR Interventions for SAD

There was large variability in the VR software used for SAD. This is likely due to the infancy of VR for SAD. With such variability, it is inevitable that reports of usability will vary according to the hardware and software used, with some programs being easier to use than others.

A distinct hindrance in evaluating the usability of VR for SAD was the lack of an existing framework. The studies largely used nonstandardized questions and qualitative feedback to determine usability, making it difficult to generalize findings across multiple studies. Although most studies did not comment on aspects of usability, those that did provided valuable information on the usability of VR for SAD. Studies in which practitioners delivered VR therapy to individual participants reported high levels of usability, such as the ease of setting up and navigating the hardware and software. However, reports of VR use in a group setting described low levels of usability, significant amounts of time spent on setting up and storing the equipment, and loss of focus on the exposure experience when therapists were helping others with their HMDs [50].

The differences in usability between individualized and group settings highlight important requirements for the use of VR interventions for SAD. Primarily, VR technology for SAD needs to be easy to learn by patients, and it is important that errors are limited in frequency and severity and that patients can recover from errors largely autonomously. As such, we propose a "VR usability framework" for the measurement of usability of VR for SAD that borrows elements from the usability heuristics by Nielsen [33]: (1) "learnability," assessing how easy it is for a patient to set up and learn the VR technology; (2) "errors," assessing the frequency, severity, and recoverability of errors autonomously by the patient; and (3) "memorability," how easy it is to re-establish proficiency after a period of absence.

Using the VR usability framework, current trends show variability in the usability of VR for SAD. VR used in group therapy has a steep learning curve and requires substantial input from therapists to work through errors, and it is difficult to re-establish proficiency in it after a period of absence (eg, some participants wished they could take the equipment home) [50]. In contrast, VR used in individualized therapy is easy to learn, patients can autonomously handle errors, and they are familiar with the technology upon return [63,64]. Thus, VR may be more user-friendly in one-on-one therapy as opposed to a group setting, as articulated by the VR usability framework.

Acceptability of VR Interventions for SAD

The results we found regarding high VR acceptability in adult patients with SAD are congruent with earlier research by Saxena [35]. Empirical findings indicate that VR for SAD is generally acceptable to adult patients with SAD, with high scores on standardized measures of satisfaction reported by most patients. Positive qualitative responses suggest that VR allowed patients to gain more insights into their anxiety and a better understanding of the social situations that they would normally avoid or be too emotionally activated to observe [50]. For example, in real life, an individual with social anxiety may avoidantly play with their phone when someone sits next to them in a cafeteria rather than perceive the encounter as a valued learning experience. Therefore, it is likely that many adults with SAD who willingly undergo VR therapy will find the experience acceptable.

Conversely, one study [59] found that some patients reported that their expectations for the treatment were not fulfilled, and some reported feeling more stress during VR. It was also found that positive expectations of VR effectiveness as well as a positive working alliance with the therapist were significantly correlated with positive emotional changes [49]. Therefore, VR treatment may not be acceptable for all adults with SAD based on individual differences regarding their previous VR experience, their perceptions of VR therapy helpfulness, their level of distress tolerance to exposure to digital stimuli before habituation [68], and the nature of their relationship with the therapist offering VR treatment.

Attrition of VR Interventions for SAD

This review found that the attrition rate across most studies was relatively low and within acceptable levels (≤20%) [46]. Indeed, attrition rates for the use of VR interventions for SAD were found to be substantially lower than estimates from VRET in anxiety disorders [43]. Considering this, it appears that patients with SAD continue with treatment more than other patients with anxiety.

There may be several reasons for the low average attrition rate finding. First, patients with SAD may prefer to learn more about social situations in a VR space. An individual with SAD may be curious about learning about social situations but may struggle to overcome the anxiety associated with placing themselves in an environment where negative evaluation is possible. By engaging with VR, patients with SAD have the knowledge that they can exit the simulation at any point, giving them the opportunity to learn about social situations without real-world social consequences. Second, patients with SAD may be more tolerant of the potentially negative effects of VR (eg, simulator sickness) when compared with the general population with anxiety [43]. Third, patients with SAD may be more hesitant to drop out of therapy for fear of negative evaluation by examiners. For instance, patients with SAD may be more likely to remain in a study because of social desirability bias—the tendency to respond in a certain way to avoid criticism [69].

It is important to note that the observed attrition rates are heterogeneous. Some studies reported proportionally higher attrition rates than others [54,55]. This may be due to the differences in the number of sessions involved in different studies. For example, some studies were composed of single sessions [49,59,62], whereas the study with the highest attrition had 9 to 17 sessions [55]. As it takes longer to deliver all sessions, there is more opportunity for participants to drop out. Furthermore, attrition was defined in this review as those who did not complete measurements during or after intervention use, including completion of follow-up measures. Considering that some studies included follow-up measures of 3 months after the intervention or longer, it is plausible that participants may not have re-engaged in these measures for several reasons. These could potentially include both therapy-related factors (eg, intolerance of VR-induced anxiety, simulator sickness, and low satisfaction levels) or factors outside of therapy (eg, moving away, becoming too busy in everyday life, and major life events).

Recommendations

Safety

With regard to safety, the primary issue identified in this review was simulator sickness. Several factors appear to be related to the susceptibility to simulator sickness. If simulator sickness is exacerbated by physiological symptoms of anxiety (eg, hyperventilation leading to dizziness) [66,67], it may be helpful to target these symptoms with clinical treatment before using VR technology. This is in line with other studies exploring attrition in anxiety disorders [43], which found that VRET attrition occurs early in treatment because of factors such as dizziness. As such, VR protocols for SAD should aim to improve retention at the beginning of treatment using a phase-based approach that includes strategies to tolerate negative emotions before immersion in VR. These may include implementation of relaxation and grounding techniques [70] or the prescription of antinausea medications. Changes in VR technology can also be applied to reduce simulator sickness [48]. Blurring or lowering the resolution of a VR image has been shown to reduce simulator sickness and improve the sense of reality [71,72].

Future Research

Presently, there are many research gaps in the literature regarding the user experience of VR for SAD. The development of a standardized measure to assess the usability of VR for SAD has the potential to identify prominent issues with usability and aid in the development of future VR programs. This measure may include elements identified in the VR usability framework discussed previously to assess learnability, error recoverability, and memorability. This may be applied to technical developments in VR that would likely improve VR's "plug-n-play" capability for SAD and other anxiety disorder treatments. Future research should also delve deeper into the study of simulator sickness in patients with SAD when compared with both healthy controls and patients with other anxiety disorders. This may lead to valuable information on reducing simulator sickness, thereby reducing the levels of attrition and improving the user experience of VR for SAD. Finally, there were no studies found that specifically targeted a child or adolescent population. Given that the onset of SAD typically occurs around adolescence [1], future studies should evaluate the efficacy of early intervention of VR for SAD, particularly given adolescents' success with VR for psychological distress [32].

Limitations

This review has several limitations. First, our review did not perform a cost-benefit analysis of the hardware identified (eg, HMDs). Affordability could have implications for the acceptability of VR among consumers with SAD. Second, we included only English-language studies, and there may have been pertinent articles published in other languages. Third, this study conducted a qualitative review of studies with different designs. Although the MMAT [46] was used to assess the quality of the studies, there is still a risk of subjective reviewer bias in addressing its criteria. Finally, studies may have been missed in our search because of obscure nomenclature (eg, research publications that did not clearly specify the use of a VR intervention for SAD in their title and abstract).

Conclusions

Our review findings showed that VRET interventions can generally provide an effective, safe, usable, acceptable, and low-attrition treatment option for adults with SAD. Nevertheless, there are research gaps evident when appraising user experience outcomes. These include the need to conduct more VR research with children and adolescents with SAD. We also do not yet know the specific causes of elevated simulator sickness in patients with SAD compared with participants without SAD or how effective other VR-based interventions beyond exposure therapy (eg, focused on mindfulness, relaxation, or cognitive distraction) are in the treatment of SAD. Further experimental studies (eg, pilot feasibility studies and RCTs) are required to explore these domains.

Acknowledgments

The authors acknowledge the financial support provided by Charles Sturt University.

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2009 checklist.

[PDF File (Adobe PDF File), 146 KB - mental_v11i1e48916_app1.pdf]

Multimedia Appendix 2

Quality assessment results of the included studies.

[PDF File (Adobe PDF File), 174 KB - mental_v11i1e48916_app2.pdf]

References

1. Leigh E, Clark DM. Understanding social anxiety disorder in adolescents and improving treatment outcomes: applying the cognitive model of Clark and Wells (1995). Clin Child Fam Psychol Rev 2018 Sep;21(3):388-414 [FREE Full text] [doi: 10.1007/s10567-018-0258-5] [Medline: 29654442]
2. Social anxiety disorder: recognition, assessment and treatment. National Institute for health and care excellence. 2013. URL: https://www.nice.org.uk/guidance/cg159 [accessed 2023-12-09]

3. Carpenter JK, Andrews LA, Witcraft SM, Powers MB, Smits JA, Hofmann SG. Cognitive behavioral therapy for anxiety and related disorders: a meta-analysis of randomized placebo-controlled trials. Depress Anxiety 2018 Jun;35(6):502-514 [FREE Full text] [doi: 10.1002/da.22728] [Medline: 29451967]
4. Hans E, Hiller W. A meta-analysis of nonrandomized effectiveness studies on outpatient cognitive behavioral therapy for adult anxiety disorders. Clin Psychol Rev 2013 Dec;33(8):954-964. [doi: 10.1016/j.cpr.2013.07.003] [Medline: 23988455]
5. Crome E, Grove R, Baillie AJ, Sunderland M, Teesson M, Slade T. DSM-IV and DSM-5 social anxiety disorder in the Australian community. Aust N Z J Psychiatry 2015 Mar;49(3):227-235 [FREE Full text] [doi: 10.1177/0004867414546699] [Medline: 25122449]
6. Horigome T, Kurokawa S, Sawada K, Kudo S, Shiga K, Mimura M, et al. Virtual reality exposure therapy for social anxiety disorder: a systematic review and meta-analysis. Psychol Med 2020 Nov;50(15):2487-2497. [doi: 10.1017/S0033291720003785] [Medline: 33070784]
7. Bush J. Viability of virtual reality exposure therapy as a treatment alternative. Comput Human Behav 2008 May;24(3):1032-1040 [FREE Full text] [doi: 10.1016/j.chb.2007.03.006]
8. Riches S, Jeyarajaguru P, Taylor L, Fialho C, Little J, Ahmed L, et al. Virtual reality relaxation for people with mental health conditions: a systematic review. Soc Psychiatry Psychiatr Epidemiol 2023 Jul 20;58(7):989-1007 [FREE Full text] [doi: 10.1007/s00127-022-02417-5] [Medline: 36658261]
9. Davis BA, Bryla K, Benton PA, Rosedale P. Oculus Rift in Action. Shelter Island, NY: Manning; 2015.
10. Cipresso P, Giglioli IA, Raya MA, Riva G. The past, present, and future of virtual and augmented reality research: a network and cluster analysis of the literature. Front Psychol 2018 Sep 02;9(4870):2086-2085 [FREE Full text] [doi: 10.3389/fpsyg.2018.02086] [Medline: 30459681]
11. Shahrbanian S, Ma X, Aghaei NN, Simmonds M. Use of virtual reality (immersive vs. non immersive) for pain management in children and adults: a systematic review of evidence from randomized controlled trials. Eur J Exp Biol 2012;2(5):1408-1422 [FREE Full text]
12. Ventura S, Brivio E, Riva G, Baños RM. Immersive versus non-immersive experience: exploring the feasibility of memory assessment through 360° technology. Front Psychol 2019 Nov 14;10:2509 [FREE Full text] [doi: 10.3389/fpsyg.2019.02509] [Medline: 31798492]
13. Rizzolatti G, Craighero L. The mirror-neuron system. Annu Rev Neurosci 2004;27:169-192. [doi: 10.1146/annurev.neuro.27.070203.144230] [Medline: 15217330]
14. Sherman WR, Craig AB. Understanding Virtual Reality: Interface, Application, and Design. 2nd edition. Cambridge, MA: Morgan Kaufmann; 2018.
15. Nunes de Vasconcelos G, Malard M, van Stralen M, Campomori M, Canavezzi de Abreu S, Lobosco T, et al. Do we still need CAVEs? In: Proceedings of the 37th eCAADe and 23rd SIGraDi Conference. 2019 Presented at: eCAADe-SIGraDi '19; September 11-13, 2019; Porto, Portugal p. 133-142 URL: https://papers.cumincad.org/cgi-bin/works/paper/ecaadesigradi2019_474 [doi: 10.5151/proceedings-ecaadesigradi2019_474]
16. Havig P, McIntire J, Geiselman E. Virtual reality in a cave: limitations and the need for HMDs? Proc. SPIE 8041, Head- and Helmet-Mounted Displays XVI: Design and Applications,. 2011. URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/8041/1/Virtual-reality-in-a-cave--limitations-and-the-need/10.1117/12.883855.short [accessed 2024-01-08]
17. Maples-Keller JL, Bunnell BE, Kim SJ, Rothbaum BO. The use of virtual reality technology in the treatment of anxiety and other psychiatric disorders. Harv Rev Psychiatry 2017 May;25(3):103-113 [FREE Full text] [doi: 10.1097/HRP.0000000000000138] [Medline: 28475502]
18. Anderson P, Rothbaum BO, Hodges LF. Virtual reality exposure in the treatment of social anxiety. Cogn Behav Pract 2003 Jun;10(3):240-247 [FREE Full text] [doi: 10.1016/s1077-7229(03)80036-6]
19. Parrish DE, Oxhandler HK, Duron JF, Swank P, Bordnick P. Feasibility of virtual reality environments for adolescent social anxiety disorder. Res Soc Work Pract 2016 Aug 03;26(7):825-835 [FREE Full text] [doi: 10.1177/1049731514568897]
20. Caponnetto P, Triscari S, Maglia M, Quattropani MC. The simulation game-virtual reality therapy for the treatment of social anxiety disorder: a systematic review. Int J Environ Res Public Health 2021 Dec 15;18(24):13209 [FREE Full text] [doi: 10.3390/ijerph182413209] [Medline: 34948817]
21. Carl E, Stein AT, Levihn-Coon A, Pogue JR, Rothbaum B, Emmelkamp P, et al. Virtual reality exposure therapy for anxiety and related disorders: a meta-analysis of randomized controlled trials. J Anxiety Disord 2019 Jan;61:27-36. [doi: 10.1016/j.janxdis.2018.08.003] [Medline: 30287083]
22. Chesham RK, Malouff JM, Schutte NS. Meta-analysis of the efficacy of virtual reality exposure therapy for social anxiety. Behav change 2018 Jun 18;35(3):152-166 [FREE Full text] [doi: 10.1017/bec.2018.15]
23. Emmelkamp PM, Meyerbröker K, Morina N. Virtual reality therapy in social anxiety disorder. Curr Psychiatry Rep 2020 May 13;22(7):32 [FREE Full text] [doi: 10.1007/s11920-020-01156-1] [Medline: 32405657]
24. Morina N, Kampmann I, Emmelkamp P, Barbui C, Hoppen TH. Meta-analysis of virtual reality exposure therapy for social anxiety disorder. Psychol Med 2023 Apr;53(5):2176-2178 [FREE Full text] [doi: 10.1017/S0033291721001690] [Medline: 34001293]

25. Howard MC, Van Zandt EC. A meta-analysis of the virtual reality problem: unequal effects of virtual reality sickness across individual differences. Virtual Real 2021 May 11;25(4):1221-1246 [FREE Full text] [doi: 10.1007/s10055-021-00524-3]
26. Cobb SV, Nichols S, Ramsey A, Wilson JR. Virtual reality-induced symptoms and effects (VRISE). Presence (Camb) 1999 Apr;8(2):169-186 [FREE Full text] [doi: 10.1162/105474699566152]
27. Kennedy RS, Fowlkes JE, Berbaum KS, Lilienthal MG. Use of a motion sickness history questionnaire for prediction of simulator sickness. Aviat Space Environ Med 1992 Jul;63(7):588-593. [Medline: 1616434]
28. Dennison MS, D'Zmura M. Cybersickness without the wobble: experimental results speak against postural instability theory. Appl Ergon 2017 Jan;58:215-223. [doi: 10.1016/j.apergo.2016.06.014] [Medline: 27633216]
29. Koslucher F, Haaland E, Stoffregen TA. Sex differences in visual performance and postural sway precede sex differences in visually induced motion sickness. Exp Brain Res 2016 Jan;234(1):313-322. [doi: 10.1007/s00221-015-4462-y] [Medline: 26466829]
30. Li R, Peterson N, Walter HJ, Rath R, Curry C, Stoffregen TA. Real-time visual feedback about postural activity increases postural instability and visually induced motion sickness. Gait Posture 2018 Sep;65:251-255. [doi: 10.1016/j.gaitpost.2018.08.005] [Medline: 30558940]
31. Kemeny A, Chardonnet JR, Colombet F. Getting Rid of Cybersickness: In Virtual Reality, Augmented Reality, and Simulators. Cham, Switzerland: Springer; 2020.
32. Kelson JN, Ridout B, Steinbeck K, Campbell AJ. The use of virtual reality for managing psychological distress in adolescents: systematic review. Cyberpsychol Behav Soc Netw 2021 Oct 01;24(10):633-641. [doi: 10.1089/cyber.2021.0090] [Medline: 34558970]
33. Nielsen J. Usability 101: introduction to usability. Nielsen Norman Group. 2012 Jan 03. URL: https://www.nngroup.com/articles/usability-101-introduction-to-usability/ [accessed 2023-12-09]
34. Ayala GX, Elder JP. Qualitative methods to ensure acceptability of behavioral and social interventions to the target population. J Public Health Dent 2011;71 Suppl 1(0 1):S69-S79 [FREE Full text] [doi: 10.1111/j.1752-7325.2011.00241.x] [Medline: 21656958]
35. Saxena A. Acceptability and tolerability of omnidirectional virtual reality exposure therapy for generalized social anxiety disorder. Palo Alto University ProQuest Dissertations Publishing. 2021. URL: https://www.proquest.com/docview/2604476479 [accessed 2024-01-18]
36. Riches S, Azevedo L, Bird L, Pisani S, Valmaggia L. Virtual reality relaxation for the general population: a systematic review. Soc Psychiatry Psychiatr Epidemiol 2021 Oct 13;56(10):1707-1727 [FREE Full text] [doi: 10.1007/s00127-021-02110-z] [Medline: 34120220]
37. Rus-Calafell M, Garety P, Sason E, Craig TJ, Valmaggia LR. Virtual reality in the assessment and treatment of psychosis: a systematic review of its utility, acceptability and effectiveness. Psychol Med 2018 Feb;48(3):362-391. [doi: 10.1017/S0033291717001945] [Medline: 28735593]
38. Papola D, Ostuzzi G, Tedeschi F, Gastaldon C, Purgato M, Del Giovane C, et al. Comparative efficacy and acceptability of psychotherapies for panic disorder with or without agoraphobia: systematic review and network meta-analysis of randomised controlled trials. Br J Psychiatry 2022 Sep;221(3):507-519 [FREE Full text] [doi: 10.1192/bjp.2021.148] [Medline: 35049483]
39. Coventry PA, Meader N, Melton H, Temple M, Dale H, Wright K, et al. Psychological and pharmacological interventions for posttraumatic stress disorder and comorbid mental health problems following complex traumatic events: systematic review and component network meta-analysis. PLoS Med 2020 Aug 19;17(8):e1003262 [FREE Full text] [doi: 10.1371/journal.pmed.1003262] [Medline: 32813696]
40. Melton H, Meader N, Dale H, Wright K, Jones-Diette J, Temple M, et al. Interventions for adults with a history of complex traumatic events: the INCiTE mixed-methods systematic review. Health Technol Assess 2020 Sep;24(43):1-312 [FREE Full text] [doi: 10.3310/hta24430] [Medline: 32924926]
41. Björk T, Björck C, Clinton D, Sohlberg S, Norring C. What happened to the ones who dropped out? Outcome in eating disorder patients who complete or prematurely terminate treatment. Eur Eat Disord Rev 2009 Mar;17(2):109-119. [doi: 10.1002/erv.911] [Medline: 19142975]
42. Barrett MS, Chua WJ, Crits-Christoph P, Gibbons MB, Casiano D, Thompson D. Early withdrawal from mental health treatment: implications for psychotherapy practice. Psychotherapy (Chic) 2008 Jun 01;45(2):247-267 [FREE Full text] [doi: 10.1037/0033-3204.45.2.247] [Medline: 19838318]
43. Benbow AA, Anderson PL. A meta-analytic examination of attrition in virtual reality exposure therapy for anxiety disorders. J Anxiety Disord 2019 Jan;61:18-26. [doi: 10.1016/j.janxdis.2018.06.006] [Medline: 30646997]
44. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
45. Jerry P, Tavares-Jones N. Virtual Worlds: The Virtual Reality and Augmented Reality Intersections. Oxford, UK: Inter-Disciplinary Press; 2016.

46. Hong QN, Fàbregues S, Bartlett G, Boardman F, Cargo M, Dagenais P, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf 2018 Dec 18;34(4):285-291 [FREE Full text] [doi: 10.3233/EFI-180221]
47. Kim HE, Hong YJ, Kim MK, Jung YH, Kyeong S, Kim JJ. Effectiveness of self-training using the mobile-based virtual reality program in patients with social anxiety disorder. Comput Human Behav 2017 Aug;73:614-619 [FREE Full text] [doi: 10.1016/j.chb.2017.04.017]
48. Kim MK, Eom H, Kwon JH, Kyeong S, Kim JJ. Neural effects of a short-term virtual reality self-training program to reduce social anxiety. Psychol Med 2022 May;52(7):1296-1305. [doi: 10.1017/S0033291720003098] [Medline: 32880252]
49. Moldovan R, David D. One session treatment of cognitive and behavioral therapy and virtual reality for social and specific phobias. J Evid Based Psychother 2014;14(1):67-83 [FREE Full text]
50. Arnfred BT, Bang P, Winther Davy J, Larsen LQ, Hjorthøj C, Christensen AB. Virtual reality exposure in cognitive behavioral group therapy for social anxiety disorder: a qualitative evaluation based on patients' and therapists' experiences. Transl Issues Psychol Sci 2021 Sep;7(3):229-247 [FREE Full text] [doi: 10.1037/tps0000291]
51. Anderson PL, Price M, Edwards SM, Obasaju MA, Schmertz SK, Zimand E, et al. Virtual reality exposure therapy for social anxiety disorder: a randomized controlled trial. J Consult Clin Psychol 2013 Oct;81(5):751-760. [doi: 10.1037/a0033559] [Medline: 23796315]
52. Bouchard S, Dumoulin S, Robillard G, Guitard T, Klinger É, Forget H, et al. Virtual reality compared with in vivo exposure in the treatment of social anxiety disorder: a three-arm randomised controlled trial. Br J Psychiatry 2017 Apr;210(4):276-283. [doi: 10.1192/bjp.bp.116.184234] [Medline: 27979818]
53. Geraets CN, Veling W, Witlox M, Staring AB, Matthijssen SJ, Cath D. Virtual reality-based cognitive behavioural therapy for patients with generalized social anxiety disorder: a pilot study. Behav Cogn Psychother 2019 Nov;47(6):745-750. [doi: 10.1017/S1352465819000225] [Medline: 30915939]
54. Hur JW, Shin H, Jung D, Lee HJ, Lee S, Kim GJ, et al. Virtual reality-based psychotherapy in social anxiety disorder: fMRI study using a self-referential task. JMIR Ment Health 2021 Apr 14;8(4):e25731 [FREE Full text] [doi: 10.2196/25731] [Medline: 33851931]
55. Jeong HS, Lee JH, Kim HE, Kim JJ. Appropriate number of treatment sessions in virtual reality-based individual cognitive behavioral therapy for social anxiety disorder. J Clin Med 2021 Feb 26;10(5):915 [FREE Full text] [doi: 10.3390/jcm10050915] [Medline: 33652739]
56. Kampmann IL, Emmelkamp PM, Hartanto D, Brinkman WP, Zijlstra BJ, Morina N. Exposure to virtual social interactions in the treatment of social anxiety disorder: a randomized controlled trial. Behav Res Ther 2016 Feb;77:147-156. [doi: 10.1016/j.brat.2015.12.016] [Medline: 26752328]
57. Kim HJ, Lee S, Jung D, Hur JW, Lee HJ, Lee S, et al. Effectiveness of a participatory and interactive virtual reality intervention in patients with social anxiety disorder: longitudinal questionnaire study. J Med Internet Res 2020 Oct 06;22(10):e23024 [FREE Full text] [doi: 10.2196/23024] [Medline: 33021481]
58. Kovar I. Virtual reality as support of cognitive behavioral therapy in social anxiety disorder. Int J Adv Sci Eng Inf Technol 2018;8(4):1343-1349 [FREE Full text] [doi: 10.18517/ijaseit.8.4.5904]
59. Lindner P, Dagöö J, Hamilton W, Miloff A, Andersson G, Schill A, et al. Virtual Reality exposure therapy for public speaking anxiety in routine care: a single-subject effectiveness trial. Cogn Behav Ther 2021 Jan;50(1):67-87 [FREE Full text] [doi: 10.1080/16506073.2020.1795240] [Medline: 32870126]
60. Perandré YH, Haydu VB. A treatment program for social anxiety disorder by using virtual reality. Temas Psicol 2018;26(2):867-882 [FREE Full text] [doi: 10.9788/tp2018.2-12en]
61. Price M, Anderson PL. Outcome expectancy as a predictor of treatment response in cognitive behavioral therapy for public speaking fears within social anxiety disorder. Psychotherapy (Chic) 2012 Jun;49(2):173-179 [FREE Full text] [doi: 10.1037/a0024734] [Medline: 21967073]
62. Rubin M, Muller K, Hayhoe MM, Telch MJ. Attention guidance augmentation of virtual reality exposure therapy for social anxiety disorder: a pilot randomized controlled trial. Cogn Behav Ther 2022 Sep;51(5):371-387 [FREE Full text] [doi: 10.1080/16506073.2022.2053882] [Medline: 35383544]
63. Trahan MH, Morley RH, Nason EE, Rodrigues N, Huerta L, Metsis V. Virtual reality exposure simulation for student veteran social anxiety and PTSD: a case study. Clin Soc Work J 2021;49(2):220-230 [FREE Full text] [doi: 10.1007/s10615-020-00784-7] [Medline: 33487778]
64. Zainal NH, Chan WW, Saxena AP, Taylor CB, Newman MG. Pilot randomized trial of self-guided virtual reality exposure therapy for social anxiety disorder. Behav Res Ther 2021 Dec;147:103984 [FREE Full text] [doi: 10.1016/j.brat.2021.103984] [Medline: 34740099]
65. Blöte AW, Kint MJ, Miers AC, Westenberg PM. The relation between public speaking anxiety and social anxiety: a review. J Anxiety Disord 2009 Apr;23(3):305-313. [doi: 10.1016/j.janxdis.2008.11.007] [Medline: 19117721]
66. Jacob RG, Redfern MS, Furman JM. Optic flow-induced sway in anxiety disorders associated with space and motion discomfort. J Anxiety Disord 1995 Sep;9(5):411-425 [FREE Full text] [doi: 10.1016/0887-6185(95)00021-f]
67. Brandt T. Phobic postural vertigo. Neurology 1996 Jun;46(6):1515-1519. [doi: 10.1212/wnl.46.6.1515] [Medline: 8649539]

68. Benito KG, Walther M. Therapeutic process during exposure: habituation model. J Obsessive Compuls Relat Disord 2015 Jul 01;6:147-157 [FREE Full text] [doi: 10.1016/j.jocrd.2015.01.006] [Medline: 26258012]
69. Hebert JR, Ma Y, Clemow L, Ockene IS, Saperia G, Stanek EJ, et al. Gender differences in social desirability and social approval bias in dietary self-report. Am J Epidemiol 1997 Dec 15;146(12):1046-1055. [doi: 10.1093/oxfordjournals.aje.a009233] [Medline: 9420529]
70. Staab JP, Ruckenstein MJ. Which comes first? Psychogenic dizziness versus otogenic anxiety. Laryngoscope 2003 Oct;113(10):1714-1718. [doi: 10.1097/00005537-200310000-00010] [Medline: 14520095]
71. Blum T, Wieczorek M, Aichert A, Tibrewal R, Navab N. The effect of out-of-focus blur on visual discomfort when using stereo displays. In: Proceedings of the 2010 IEEE International Symposium on Mixed and Augmented Reality. 2010 Presented at: ISMAR '10; October 13-16, 2010; Seoul, South Korea p. 13-17 URL: https://ieeexplore.ieee.org/document/5643544 [doi: 10.1109/ismar.2010.5643544]
72. Carnegie K, Rhee T. Reducing visual discomfort with HMDs using dynamic depth of field. IEEE Comput Graph Appl 2015;35(5):34-41. [doi: 10.1109/MCG.2015.98] [Medline: 26416360]

Abbreviations

CAVE: cave automatic virtual environment
HMD: head-mounted display
MMAT: Mixed Methods Appraisal Tool
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT: randomized controlled trial
SAD: social anxiety disorder
VR: virtual reality
VRET: virtual reality exposure therapy

Edited by J Torous; submitted 21.05.23; peer-reviewed by S Triscari, Q Xu, N Benson; comments to author 30.09.23; revised version received 24.11.23; accepted 05.12.23; published 08.02.24.

Please cite as:
Shahid S, Kelson J, Saliba A
Effectiveness and User Experience of Virtual Reality for Social Anxiety Disorder: Systematic Review
JMIR Ment Health 2024;11:e48916
URL: https://mental.jmir.org/2024/1/e48916
doi:10.2196/48916
PMID:38329804

Review

Immersive Technologies for Depression Care: Scoping Review

C Mahony Reategui-Rivera[1], MD; David Villarreal-Zegarra[2,3], MPH; Kelly De La Cruz-Torralva[4], BSc; Paquita Díaz-Sánchez[4], MD; Joseph Finkelstein[1], MD, PhD

[1]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, United States

[2]Instituto Peruano de Orientación Psicólogica, Lima, Peru

[3]Escuela de Psicología, Universidad Continental, Lima, Peru

[4]Unidad de Telesalud, Facultad de Medicina, Universidad Nacional Mayor de San Marcos, Lima, Peru

Corresponding Author:
C Mahony Reategui-Rivera, MD
Department of Biomedical Informatics
University of Utah
421 Wakara Way, Ste 140
Salt Lake City, UT, 84108
United States
Phone: 1 801 581 4080
Email: mahony.reategui@utah.edu

Abstract

Background: Depression significantly impacts quality of life, affecting approximately 280 million people worldwide. However, only 16.5% of those affected receive treatment, indicating a substantial treatment gap. Immersive technologies (IMTs) such as virtual reality (VR) and augmented reality offer new avenues for treating depression by creating immersive environments for therapeutic interventions. Despite their potential, significant gaps exist in the current evidence regarding the design, implementation, and use of IMTs for depression care.

Objective: We aim to map the available evidence on IMT interventions targeting depression treatment.

Methods: This scoping review followed a methodological framework, and we systematically searched databases for studies on IMTs and depression. The focus was on randomized clinical trials involving adults and using IMTs. The selection and charting process involved multiple reviewers to minimize bias.

Results: The search identified 16 peer-reviewed articles, predominantly from Europe (n=10, 63%), with a notable emphasis on Poland (n=9, 56%), which contributed to more than half of the articles. Most of the studies (9/16, 56%) were conducted between 2020 and 2021. Regarding participant demographics, of the 16 articles, 5 (31%) exclusively involved female participants, and 7 (44%) featured participants whose mean or median age was >60 years. Regarding technical aspects, all studies focused on VR, with most using stand-alone VR headsets (14/16, 88%), and interventions typically ranging from 2 to 8 weeks, predominantly in hospital settings (11/16, 69%). Only 2 (13%) of the 16 studies mentioned using a specific VR design framework in planning their interventions. The most frequently used therapeutic approach was Ericksonian psychotherapy, used in 56% (9/16) of the studies. Notably, none of the articles reported using an implementation framework or identified barriers and enablers to implementation.

Conclusions: This scoping review highlights the growing interest in using IMTs, particularly VR, for depression treatment but emphasizes the need for more inclusive and comprehensive research. Future studies should explore varied therapeutic approaches and cost-effectiveness as well as the inclusion of augmented reality to fully realize the potential of IMTs in mental health care.

(JMIR Ment Health 2024;11:e56056) doi:10.2196/56056

KEYWORDS

depression; immersive technologies; virtual reality; augmented reality; mobile phone

Introduction

Background

Depression is a debilitating disorder characterized by a persistent low mood and a loss of interest in everyday activities, significantly affecting various dimensions of life [1]. Globally, approximately 280 million people are afflicted by this condition [2]. However, only 16.5% of people with depression worldwide receive treatment, indicating a substantial treatment gap [3]. The scarcity of mental health professionals exacerbates this issue, with figures in low- and middle-income countries being particularly low at 1.4 to 3.8 per 100,000 population [4]. This shortage of resources highlights the urgent need for innovative solutions in mental health care [5].

Digital technologies, now more crucial than ever, have emerged as vital tools in bridging health care gaps [6]. Among these, immersive technologies (IMTs) such as virtual reality (VR) and augmented reality (AR) stand out for their potential to revolutionize depression care. These technologies offer computer-generated immersive experiences that blend virtual and real environments, with VR providing entirely virtual experiences and AR overlaying virtual objects onto the real world [7].

The application of IMTs in mental health leverages their ability to create controlled immersive environments, offering a safe space for individuals to explore coping exercises and techniques [8,9]. This digital modality encompasses immersive sensory experiences that allow users to interact with a virtual environment [10,11]. Such interactions have been shown to increase engagement in health care–related tasks [12,13], which is a crucial challenge in the treatment of anxiety and depression [14], providing a novel approach to mental health care. Furthermore, continual improvements in IMT device technology, exemplified by the Meta Quest VR headsets, have further broadened the accessibility of these interventions globally. Moreover, IMTs have proven effective in treating a wide range of mental health conditions, such as anxiety, posttraumatic stress disorder, autism spectrum disorders, and various phobias [15-22].

Objectives

Although some reviews have examined the use of IMTs in treating depression [23-25], they have not focused primarily on depression as the assessment goal; nor have they focused on IMT applications specifically aimed at treating depression or on the psychotherapeutic aspects of these interventions. Moreover, the available literature did not address relevant elements, such as the design or implementation of IMT interventions.

Therefore, we aim to map the most rigorous available evidence on IMT interventions targeting depression treatment and identify the gaps related to the design and implementation of these interventions. Given the emerging nature of IMTs in mental health and our specific research focus, a scoping review was deemed the most appropriate methodology. This approach allows for a broad overview of the existing literature, identifying key concepts and highlighting gaps in the research. Furthermore, we decided to focus on randomized clinical trials (RCTs) to ensure a robust and reliable evidence base. RCTs are considered the gold standard in clinical research, providing high-quality data to inform clinical practice and guide future research. By concentrating on RCTs, we aim to capture the most rigorous and scientifically valid studies, thereby enhancing the credibility and applicability of our findings in this emerging area of mental health care.

Methods

Overview

This scoping review adheres to the framework formulated by Arksey and O'Malley [26], expanded upon by Levac et al [27] and Daudt et al [28] and summarized by the JBI Manual for Evidence Synthesis [29]. Accordingly, we followed five main steps for conducting the scoping review: (1) identifying the research questions; (2) identifying relevant studies; (3) selecting the studies; (4) charting the data; and (5) collating, summarizing, and reporting the results.

Regarding reporting, our study aligns with the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) 2020 guidelines [30] (Multimedia Appendix 1). The protocol was preregistered on Open Science Framework [31].

Step 1: Identifying the Research Questions

The primary research question guiding this study is *What is the available scientific evidence and what are the gaps that exist in this evidence concerning the use of IMTs to address depression among adults?*

The study also seeks to address the following secondary questions:

- From which regions or countries does the evidence come?
- Which technical aspects of IMTs have been reported in the evidence?
- What therapeutic approaches were used?
- What are the barriers and facilitators to implementing IMT interventions for depression treatment?
- What outcomes have been evaluated in studies examining the impact of IMT interventions on addressing depression?

Step 2: Identifying Relevant Studies

We conducted a systematic search of the following electronic databases: MEDLINE (via PubMed), Scopus, Web of Science, Embase, PsycINFO (via EBSCO), IEEE Xplore, and Cochrane Library. The search strategy incorporated the keywords "virtual reality," "augmented reality," "depression," and "randomized clinical trial." An example of the search query performed in PubMed is presented in Textbox 1, and the search queries for each database are detailed in Multimedia Appendix 2. The search was limited to articles published in English and spanned from the inception of each database to October 10, 2023.

Textbox 1. PubMed search query.

Search query

- (“Depression”[MeSH] OR “Depressive Disorder”[MeSH] OR depressive*[tiab] OR depression[tiab]) AND “Virtual Reality”[MeSH] OR “virtual reality”[tiab] OR “Augmented Reality”[MeSH] OR “augmented reality”[tiab] OR “VR headset”[tiab] OR “VR glasses”[tiab] OR “virtual environment”[tiab] OR “virtual world”[tiab] OR metaverse[tiab] OR meta-verse[tiab]) AND (“Randomized Controlled Trial”[Publication Type] OR “Randomized Controlled Trials as Topic”[MeSH] OR “randomized clinical trial”[tiab] OR RCT[tiab] OR (randomized[tiab] AND “clinical trial”[tiab]))

Step 3: Study Selection

The inclusion and exclusion criteria are detailed in Textbox 2. Secondary studies were excluded, but their references were consulted to identify primary research studies that fulfilled our selection criteria. Similarly, protocols were not included, but registration IDs were consulted in the web to find preliminary or primary results published in articles.

Textbox 2. Inclusion and exclusion criteria (we defined immersive technologies as all augmented reality and virtual reality–only applications that belong to the degree of full immersion, according to the definitions provided in the literature [7,10]).

Inclusion criteria

- Articles must include randomized clinical trials and be published in peer-reviewed journals.
- All participants must be aged at least 18 y.
- Depression must be a primary outcome measured either through clinical assessment or validated screening tests.
- At least 1 group of participants should have received, or should have been exposed to, immersive technologies using glasses, headsets, or other head-mounted display devices, with or without using other complementary devices.

Exclusion criteria

- Secondary studies (systematic, umbrella, narrative, and scoping reviews) and protocols were not considered.
- Articles in which the immersive technology interventions only focused on exercise as a treatment were not considered.

We searched the various databases of scientific articles and exported all records as RIS format files. These records were imported into EndNote X9 (Clarivate) for automatic and manual duplicate checking. Subsequently, the selection process was carried out on the Rayyan web platform (Rayyan Systems Inc) in 2 phases. First, the records were screened by title and abstract by independent reviewers using the platform. In the second phase, full-text evaluations were conducted to determine compliance with the inclusion criteria. Each document was assessed independently by a pair of reviewers (CMRR and KDC as well as CMRR and PDS) to ensure that they met the eligibility criteria. Review disagreements were solved through consensus, and a third reviewer (DVZ) made a final decision in case disagreements persisted. The reasons for exclusion were documented (Multimedia Appendix 3). Before the selection process, reviewers undertook a pilot test with 10 articles to standardize the process and gain expertise in using the Rayyan platform.

Step 4: Charting the Data

Two pairs of reviewers (CMRR and KDC as well as CMRR and PDS) independently collected data using a collection form developed for the study protocol and refined at the data collection stage. The reviewers performed a pilot test with 2 documents to standardize information extraction criteria. The collected data included general and study characteristics (country of study, study design, participants’ characteristics, the type of depression outcome, and intervention and control descriptions), IMT intervention technical aspects (devices, the amount of time used and frequency of use, the setting of use, duration, IMT design framework consulted, and the degree of guidance), therapeutic approach used, and implementation characteristics (implementation framework used, implementation stage, and barriers and enablers). Study designs were categorized following the clinical trial classification formulated by Hopewell et al [32]. The implementation stages were defined as follows: (1) *preliminary*, if it was a pilot or feasibility study; (2) *implementation*, if it was mentioned that the RCT had been developed after a pilot or feasibility study; and (3) *unclear*, if there was no mention of it being a pilot or feasibility study, and there was no reference either to the results from these studies.

Step 5: Collating, Summarizing, and Reporting the Results

We used a narrative approach to synthesize data [33]. We describe the information in the *Results* section using frequencies and percentages. Detailed information for each included article is presented in cross-tables. In addition, the geographic location of the studies is visualized as a bubble plot, categorized by year.

Results

Overview

Our search strategy identified 1052 records; after removing 477 (45.34%) duplicates, the remaining 575 (54.66%) records were screened by title and abstract. Of these 575 records, 52 (9%) underwent full-text review. Of these 52 records, 36 (69%) were excluded, mainly on account of being the wrong type of publication (n=15, 42%); thus, 16 (31%) reports [34-49] from 15 studies were included in the review (Figure 1).

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flowchart outlining the search and selection process.

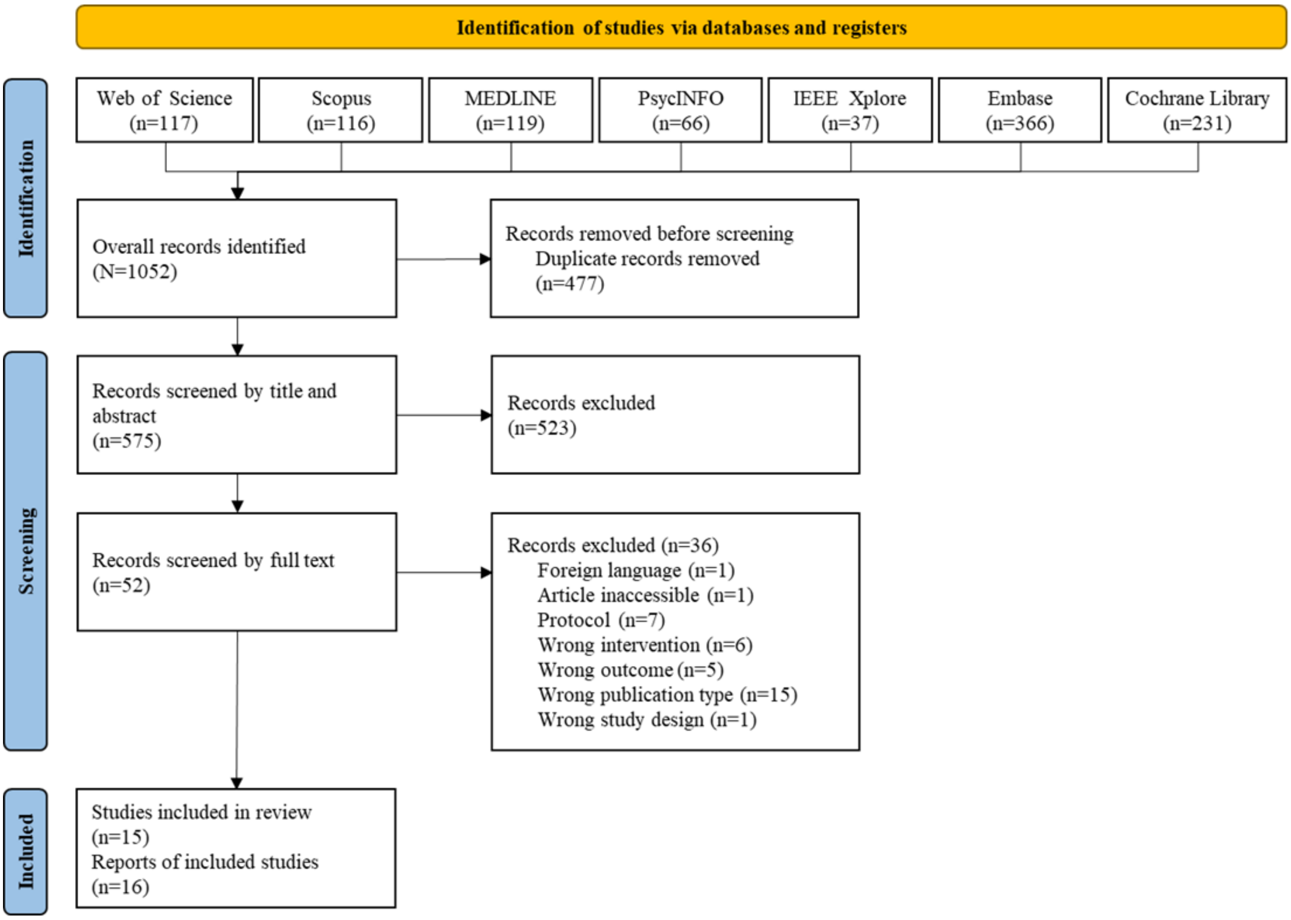


Figure 2 illustrates an uneven distribution of research evidence on IMT interventions for depression management across different regions and income levels. European countries dominate the research landscape (10/16, 63%) [34,35,37-39,43-47], with Poland alone contributing more than half of the reports (9/16, 56%) [34,35,37-39,43-46] between 2021 and 2023. However, Poland showed a decreasing trend; of the 9 reports in this review, 5 (56%) were published in 2021, while only 1 (11%) was published in 2023. China maintains a steady presence in Asia with a study each in 2022 [48] and 2023 [49], amounting to 13% (2/16) of the total. By contrast, the United States [41], Australia [40], Brazil [42], and Iran (the only lower–middle-income country contributing to this research field) [36] contributed only 1 (6%) study each to the total of 16 studies. The trend analysis indicates a fluctuating global interest in the field, with a concentration of research in Europe (10/16, 63%) [34,35,37-39,43-47], intermittent contributions from other regions, and a spike in publications during 2022 (7/16, 44%) [35,39,41,42,44,47,48].

Figure 2. Trends of publication by geographic location of the reports.

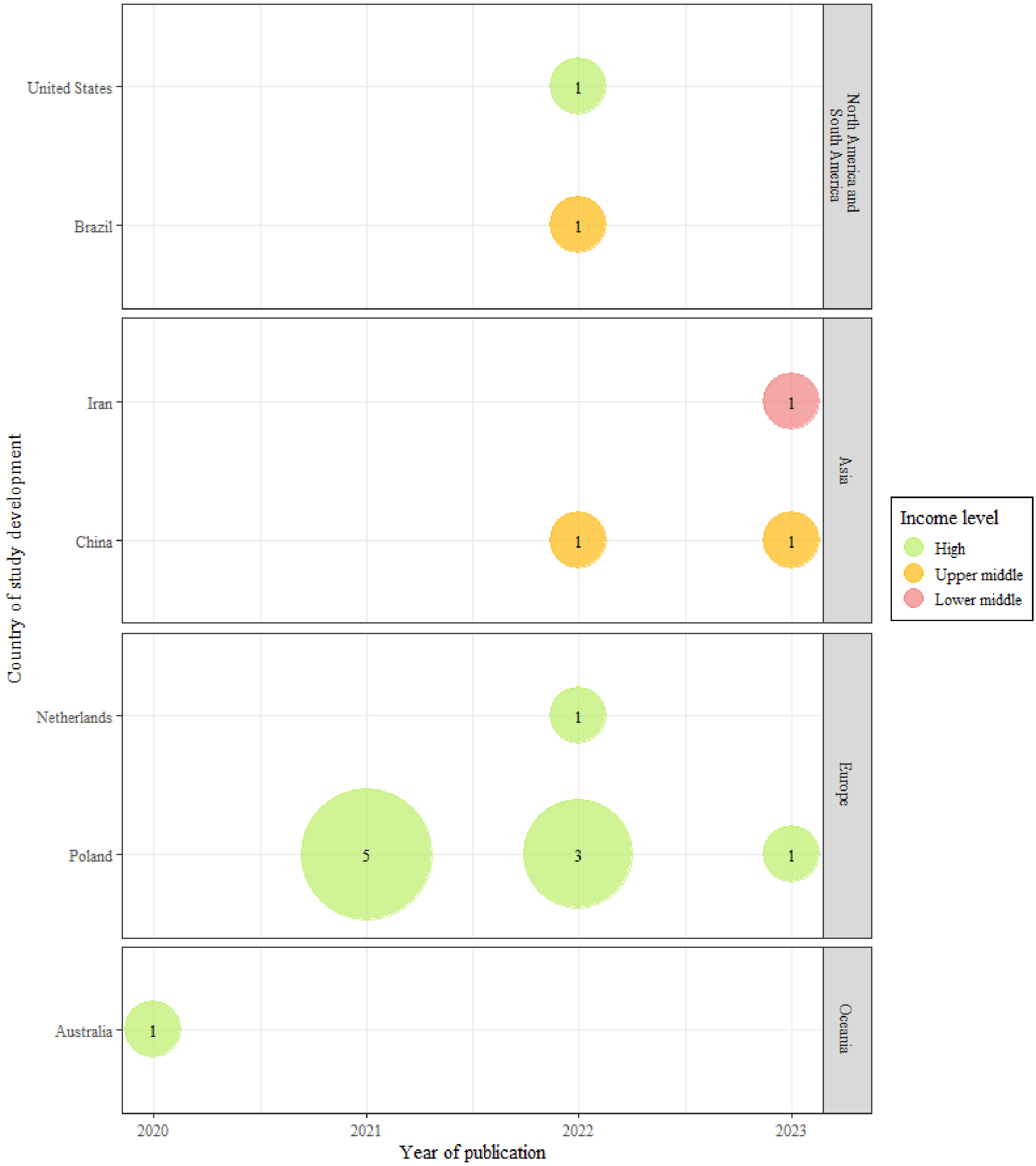


Table 1 illustrates the general characteristics of the selected reports. The majority (15/16, 94%) [34-40,42-49] adopted a parallel group design, while a single study (1/16, 6%) used a crossover trial design [40]. Among the 16 reports, only 1 (6%) featured 3 arms (2 control groups alongside 1 VR intervention group) [41]. Regarding the research timeline, most of the studies (9/16, 56%) were conducted between 2020 and 2021 [36,38,39,41-44,47,48]. However, for 3 (19%) of the 16 studies, the specific study period was not reported [35,37,45].

Table 1. Characteristics of included studies.

Reports	Year of study development	Country of study development	Study design	Participant characteristics	Specific depressive conditions studied and the measurement scale used	Sample size	VR[a] group intervention description	Non-VR group intervention description	Effect
Cieślik et al [34], 2023	2022	Poland	Parallel group 2-arm randomized controlled trial+masked outcome assessor+2 time point measures (before and after the intervention)	Sex: female (60/60, 100%); age (y): mean 68.2 (SD 5.5)	Depression among older adults measured using the GDS-30[b]	60 (randomized; intervention: 30; control: 30)	General fitness training (40 min of low-intensity general fitness exercises)+Virtual Therapeutic Garden (20 min of intense visual, auditory, and kinesthetic stimuli through immersion in a garden with the therapist's voice guiding the patient)	General fitness training (20 min)+relaxation session (10 min) and psychoeducation (10 min)	Intervention: baseline=mean 13.10 (SD 4.26); after treatment=mean 7.33 (SD 3.88); Cohen *d*=1.86, 95% CI 1.26 to 2.45; control: baseline=mean 13.27 (SD 3.80); after treatment=mean 11.57 (SD 5.49); Cohen *d*=0.42, 95% CI 0.04 to 0.79; test used: ANCOVA[c] (between groups); *P*<.001
Czech et al [35], 2022	Unclear	Poland	Parallel group 2-arm randomized controlled trial+2 time point measures (before and after the intervention)	Sex: female (16/16, 100%); age (y): intervention=mean 50.6 (SD 12.6); control=mean 59.6 (SD 7.9); other: participants with breast cancer	Depressive symptoms measured by Beck Depression Inventory	16 (randomized; intervention: 9; control: 7)	Virtual Therapeutic Garden: intense visual, auditory, and kinesthetic stimuli through garden immersion with the therapist's voice guiding the patient	Standard of care (not specified)	Intervention: baseline=mean 13.33 (SD 5.57); after treatment=mean 8.11 (SD 6.17); control: baseline=mean 9.00 (SD 7.07); after treatment=mean 7.00 (SD 5.51); test used: 1-way repeated measures ANOVA (for time and groups); *P*=.04
Farahimanesh et al [36], 2023	2021	Iran	Parallel group 2-arm randomized controlled trial+2 time point measures (before and after the intervention)	Sex: intervention=female 18/30 (60%); control=female 15/30 (50%); age (y): intervention=mean 49.1 (SD 10.9); control=mean 49.7 (SD 10.4); other: at least 2 months of social distancing measures related to the COVID-19 pandemic	Depressive symptoms measured by the Depression Anxiety Stress Scale-21	60 (randomized; intervention: 30; control: 30)	COVID Feel Good: a daily intervention with 7 thematic modules, each with two integrated parts: (1) watching a 10-min 360-degree VR video titled Secret Garden+listening to a relaxation induction narrative and (2) social tasks with a different purpose for each day of the wk	No treatment	Intervention: baseline=mean 6.6 (SD 3.1); group 1 after intervention=mean 6 (SD 2.86); group 2 after intervention=mean 5.63 (SD 2.95); control: baseline=mean 6.93 (SD 2.78); group 1 after treatment=mean 6.93 (SD 2.38); group 2 after treatment=mean 6.90 (SD 2.34); test used: ANOVA (for time and groups); *P*=.002

Reports	Year of study development	Country of study development	Study design	Participant characteristics	Specific depressive conditions studied and the measurement scale used	Sample size	VR[a] group intervention description	Non-VR group intervention description	Effect
Jóźwik et al [38], 2021	Unclear	Poland	Parallel group 2-arm randomized controlled trial+2 time point measures (before and after the intervention)	Sex: female (26/26, 100%); age (y): mean 65.4 (SD 8.0); intervention=mean 65.6 (SD 10.1); control=mean 65.2 (SD 6.5); other: participants with ischemic heart disease	Depressive symptoms measured by the HADS[d]	52 (randomized; intervention: initial=26, at conclusion=17; control: 26)	Interval training on a cycle ergometer (40 min)+general fitness exercises (40 min)+Virtual Therapeutic Garden (20 min of intense visual, auditory, and kinesthetic stimuli through immersion in a garden with the therapist's voice guiding the patient)	Interval training on a cycle ergometer (40 min)+general fitness exercises (40 min)+Schultz autogenic training guided by therapist and CD recording	Intervention: baseline=mean 6.14 (SD 3.77); after treatment=mean 4.86 (SD 3.48); mean difference=−1.29 (95% CI −2.12 to −0.46); control: baseline=mean 6.35 (SD 3.91); after treatment=mean 6.53 (SD 3.86); change=0.18 (95% CI −0.16 to 0.52); test used: 2-way repeated measures ANOVA (for time and groups); *P*=.01
Jóźwik et al [37], 2021	2020	Poland	Parallel group 2-arm randomized controlled trial+masked outcome assessor+2time point measures (before and after the intervention)	Sex: intervention=female 17/28 (61%); control=female 25/49 (51%); age (y): intervention=mean 66 (SD 9.7); control=mean 63.9 (SD 6.9); other: participants with coronary artery disease+cardiac rehabilitation phase II	Depressive symptoms measured by the HADS	100 (randomized; intervention: initial=50, at conclusion=28; control: initial=50; at conclusion=49)	Interval training on a cycle ergometer (40 min)+general fitness exercises (40 min)+Virtual Therapeutic Garden (20 min of intense visual, auditory, and kinesthetic stimuli through immersion in a garden with the therapist's voice guiding the patient)	Interval training on a cycle ergometer (40 min)+general fitness exercises (40 min)+Schultz autogenic training via CD recording (20 min)	Intervention: baseline=mean 6.41 (SD 4.21); after treatment=mean 5.06 (SD 3.88); control: baseline=mean 7.35 (SD 3.80); after treatment=mean 7.27 (SD 4.00); test used: *t* test for independent trials; *P*=.07 for posttreatment measurements

Reports	Year of study development	Country of study development	Study design	Participant characteristics	Specific depressive conditions studied and the measurement scale used	Sample size	VR[a] group intervention description	Non-VR group intervention description	Effect
Kiper et al [39], 2022	2020	Poland	Parallel group 2-arm randomized controlled trial+2 time point measures (before and after the intervention)	Sex: intervention=female 17/30 (57%); control=female 13/30 (43%); age (y): intervention=mean 65.5 (SD 6.7); control=mean 65.6 (SD 4.9); other: participants with a history of ischemic stroke and older adult depression diagnosed using the GDS-30 (>10)	Older adult depression measured using the GDS-30; depressive symptoms measured using the HADS (depression subscale)	60 (randomized; intervention: initial=30, at conclusion=22; control: initial=30, at conclusion=17)	First 3 wks: functional rehabilitation (60 min)+Virtual Therapeutic Garden (20 min of intense visual, auditory, and kinesthetic stimuli through immersion in a garden with the therapist's voice guiding the patient); next 3 weeks: functional rehabilitation (60 min)	First 3 wks: functional rehabilitation (60 min)+Schultz autogenic training via CD recording (20 min); next 3 wks: functional rehabilitation (60 min)	GDS-30—intervention: baseline vs group 1 after treatment=mean difference −6.33 (95% CI −4.42 to −8.24); baseline vs group 2 after treatment=mean difference −6.60 (95% CI −4.69 to −8.51); control: baseline vs group 1 after treatment=mean difference −3.40 (95% CI −1.49 to −5.31); baseline vs group 2 after treatment=mean difference −3.17 (95% CI −1.26 to −5.08); test used: 2-way repeated measures ANOVA (for time and groups); *P*<.01; HADS (depression scale)—intervention: baseline vs group 1 after treatment=mean difference –1.50 (95% CI –0.32 to –3.32); baseline vs group 2 after treatment=mean difference –2.05 (95% CI –0.23 to –3.87); control: baseline vs group 1 after treatment=mean difference 0.79 (95% CI –1.08 to –2.66); baseline vs group 2 after treatment=mean difference 0.73 (95% CI –1.13 to –2.61); test used: 2-way repeated measures ANOVA (for time and groups); *P*<.31

Reports	Year of study development	Country of study development	Study design	Participant characteristics	Specific depressive conditions studied and the measurement scale used	Sample size	VR[a] group intervention description	Non-VR group intervention description	Effect
Lakhani et al [40], 2020	2019	Australia	Crossover group 2-arm randomized controlled trial+3 time point measures (1 before and 2 after the intervention)	Sex: group 1=female 0/10 (0%); group 2=female 8/14 (57%); age (y): group 1=mean 56.2 (SD 20.7); group 2=mean 48.0 (SD 16.2); other: participants with spinal cord injury	Depressive symptoms measured using the Patient Health Questionnaire-8	24 (randomized; group 1: initial=10, at conclusion=6; group 2: initial=14, at conclusion=10)	Standard of care (tailored regular rehabilitation involving occupational therapy and physiotherapy)+VR session of environment exposure using the National Geographic app and the YouTube VR app	Standard of care (tailored regular rehabilitation involving occupational therapy and physiotherapy)	Group 1: baseline=mean 5.83 (SD 4.71); group after intervention treatment=mean 3.33 (SD 3.44); control group after treatment=mean 4.17 (SD 5.04); group 2: baseline=mean 5.50 (SD 2.99); control group after treatment=mean 6.30 (SD 1.95); intervention group after treatment=mean 5.90 (SD 3.28); test used: paired *t* test (between groups); *P*=.04 (between baseline and after the control treatment) and *P*=.47 (between after the control treatment and after the intervention treatment); test used: 1-way ANOVA (within groups); *P*=.09 (group 1) and *P*=.83 (group 2)
Paul et al [41], 2022	2020-2021	United States	Parallel group 3-arm randomized controlled trial+4 time point measures (1 before and 3 after the intervention)	Sex: initial=female 7/13 (54%); complete study=female 4/10 (10%); age (y): initial=mean 35.4 (SD 12.3); complete study=mean 34.6 (SD 11.50); other: participants with major depressive disorder diagnosed using the Patient Health Questionnaire-8 (score>10)	Depressive symptoms measured using the Patient Health Questionnaire-9	13 (randomized; intervention: initial=5, at conclusion=3; control 1: 4 and control 2: 4)	Behavioral activation therapy via teleconference platform+VR activities (360-degree YouTube VR videos) after teleconference session (during the wk)	Control 1: verbally asked questions for depression screening using the Patient Health Questionnaire-9 via telephone call; control 2: behavioral activation therapy via teleconference platform+in-person activities after teleconference session (during the wk)	Intervention: baseline vs posttreatment 3=mean difference −5.67; control 1: baseline vs posttreatment 3=mean difference −3.00; control 2: baseline vs posttreatment 3=mean difference −0.25 (no statistical test was used)

Reports	Year of study development	Country of study development	Study design	Participant characteristics	Specific depressive conditions studied and the measurement scale used	Sample size	VR[a] group intervention description	Non-VR group intervention description	Effect
Rodrigues et al [42], 2022	2020-2021	Brazil	Parallel group 2-arm randomized controlled trial+2 time point measures (before and after the intervention)	Sex: intervention=female 11/22 (50%); control=female 11/22 (50%); age (y): intervention=mean 48.9 (SD 13.9); control=mean 48.5 (SD 16.9); other: participants with COVID-19 infection	Depressive symptoms measured by the HADS	44 (randomized; intervention: 22; control: 22)	Usual therapy (activities to guide the hospitalization process, coping with the hospitalization process, energy conservation, daily living activity training, cognitive rehabilitation, internet-based call or visit, positioning, mobility joint, functional mobility, kinesiotherapy, assistive technology, sensory stimulation tailored to patient needs)+VR therapy (360-degree videos with images of landscapes and mindfulness techniques)	Usual therapy (activities to guide the hospitalization process, coping with the hospitalization process, energy conservation, daily living activity training, cognitive rehabilitation, internet-based call or visit, positioning, mobility joint, functional mobility, kinesiotherapy, assistive technology, sensory stimulation tailored to patient needs)+VR control (videos with advertisements not related to relaxation and well-being content)	Intervention: baseline=mean 9.83 (SD 4.31); after treatment=mean 7.17 (2.79); Cohen *d*=0.73; control: baseline=mean 15.00 (SD 10.31); after treatment=mean 13.00 (SD 9.49); Cohen *d*=0.20; test used: Wilcoxon test (within groups); *P*=.04 (intervention) and *P*=.08 (control); test used: Mann-Whitney *U* test (between groups); *P*>.05
Rutkowski et al [43], 2021	2020	Poland	Parallel group 2-arm randomized controlled trial+2 time point measures (before and after the intervention)	Sex: intervention=female 21/25 (84%); control=female 20/25 (80%); age (y): intervention=mean 64.4 (SD 5.7); control=mean 67.6 (SD 9.4); other: participants with chronic obstructive pulmonary disease and depression or anxiety diagnosed using the Hamilton Anxiety and Depression Scale (score>8)	Depressive symptoms measured using the HADS	50 (randomized; intervention: 25; control: 25)	Traditional pulmonary rehabilitation program (15-30 min)+Virtual Therapeutic Garden (20 min of intense visual, auditory, and kinesthetic stimuli through immersion in a garden with the therapist's voice guiding the patient)	Traditional pulmonary rehabilitation program (15-30 min)+Schultz autogenic training session (20 min)	Intervention: baseline=mean 7.96 (SD 2.76); after treatment=mean 6.04 (SD 3.21); control: baseline=mean 6.64 (SD 2.80); after treatment=mean 7.08 (SD 3.56); test used: repeated measures ANOVA (within groups); *P*=.001 (intervention) and *P*=.45 (control)

Reports	Year of study development	Country of study development	Study design	Participant characteristics	Specific depressive conditions studied and the measurement scale used	Sample size	VR[a] group intervention description	Non-VR group intervention description	Effect
Rutkowski et al [44], 2022	2021	Poland	Parallel group 2-arm randomized controlled trial+2 time point measures (before and after the intervention)	Sex: female 20/32 (69%); age (y): mean 57.8 (SD 4.9); other: participants with COVID-19 infection	Depressive symptoms measured using the HADS	32 (randomized; intervention: 16; control: 16)	Pulmonary rehabilitation program+VR-based exercise capacity training: bicycle ergometer+Virtual Park VR experience (bicycle trip in an island simulation synced with ergometer)+VR-based relaxation (Virtual Therapeutic Garden: intense visual, auditory, and kinesthetic stimuli through immersion in a garden with the therapist's voice guiding the patient)	Pulmonary rehabilitation program+traditional exercise capacity training: bicycle ergometer exercise+Schultz autogenic training	Intervention: baseline=mean 6.9 (SD 3.9); after treatment=mean 4.7 (SD 3.5); control: baseline=7.64 (4.5); after treatment=6.6 (4.8); test used: paired *t* test (within groups); *P*=.008 (intervention) and *P*=.02 (control)
Szczepańska-Gieracha et al [45], 2021	Unclear	Poland	Parallel group 2-arm randomized controlled trial+2 time point measures (before and after the intervention)	Sex: female 20/32 (63%); intervention=female 9/15 (60%); control=female 11/17 (65%); age (y): mean 68.9 (SD 6.3); intervention=mean 69.5 (SD 7.5); control=mean 68.4 (SD 5.0); other: participants with coronary artery disease	Depressive symptoms measured using the HADS	34 (randomized; intervention: initial=17, at conclusion=15; control: 17)	Cardiac rehabilitation (1.5 h): cardiological training individually prescribed based on an exertion test and heart rate reserve+Virtual Therapeutic Garden (20 min of intense visual, auditory, and kinesthetic stimuli through immersion in a garden with the therapist's voice guiding the patient)	Cardiac rehabilitation (1.5 h): cardiological training individually prescribed based on an exertion test and heart rate reserve+Schultz autogenic training (20 min) delivered by a therapist and CD recording	Intervention: baseline=mean 9.00 (SD 2.39); after treatment=mean 6.93 (SD 3.01); Cohen *d*=0.89; *P*=.003; control: baseline=mean 9.24 (SD 2.41); after treatment=mean 9.35 (SD 2.50); Cohen *d*=−0.15; test used: paired *t* test (within groups); *P*=.003 (intervention) and *P*=.54 (control)

Reports	Year of study development	Country of study development	Study design	Participant characteristics	Specific depressive conditions studied and the measurement scale used	Sample size	VR[a] group intervention description	Non-VR group intervention description	Effect
Szczepańska-Gieracha et al [46], 2021	2019	Poland	Parallel group 2-arm randomized controlled trial+masked outcome assessor+3time point measures (1 before and 2 after the intervention)	Sex: female 23/23 (100%); age (y): mean 70.7 (SD 13.7); intervention=mean 70.2 (SD 4.9); control=mean 71.2 (SD 4.4); other: female older adult with severe depression (GDS-30 score>10) and nonrespondent to treatment program	Older adult depression measured using the GDS-30	25 (randomized; intervention: initial=13, at conclusion=11; control: 12)	Support group meetings: general fitness training (40 min) and relaxation exercises, as well as health-promoting education and psychoeducation (20 min)+Virtual Therapeutic Garden (intense visual, auditory, and kinesthetic stimuli through immersion in a garden with the therapist's voice guiding the patient)	Support group meetings: general fitness training (40 min) and relaxation exercises, as well as health-promoting education and psychoeducation (20 min)	Intervention: baseline=mean 12.27 (SD 4.45); group 1 after treatment=mean 8.27 (SD 3.60); group 2 after treatment=mean 7.27 (SD 2.57); control: baseline=mean 12.25 (4.53); group 1 after treatment=mean 12.75 (4.82); group 2 after treatment=mean 11.83 (2.62); test used: repeated measures ANOVA (within groups); *P*<.001 (intervention) and *P*=.61 (control); test used: repeated measures ANOVA (for time and groups); *P*<.001

Reports	Year of study development	Country of study development	Study design	Participant characteristics	Specific depressive conditions studied and the measurement scale used	Sample size	VR[a] group intervention description	Non-VR group intervention description	Effect
Vlake et al [47], 2022	2020-2021	Netherlands	Parallel group 2-arm randomized controlled trial, open label+3 time point measures (1 before and 2 after the intervention)	Sex: intervention=female 10/45 (22%); control=female 16/44 (64%); age (y): intervention=median 61 (IQR 54-65); control=median 59 (IQR 51-65); other: participants included survivors of COVID-19 infection who had been discharged from the ICU[e]	Depressive symptoms measured using the HADS	89 (randomized; intervention: initial=45; 1-month assessment=45 and 3-month assessment=39; control: initial=44; 1-month assessment=44 and 3-month assessment=38)	Intensivist consultation (60 min): revision of treatment, screening for postintensive care syndrome–related impairment, and referral to appropriate health provider+intensive care unit VR (14 min of an informational VR video aimed to immerse patients in the ICU environment and provide voice-over explanations regarding different facets of the ICU and ICU treatment, which consisted of 6 scenes covering topics such as ICU equipment, procedures, and COVID-19)	Intensivist consultation (60 min): treatment revision, postintensive care syndrome–related impairment screening, and referral to the appropriate health provider	Intervention: baseline=18%; 1 month after treatment=24%; 3 months after treatment=23%; control: baseline=33%; 1 month after treatment=41%; 3 months after treatment=29%; test used=logistic regression (for time points); *P*=.57 (baseline vs first time point 1 month after the treatment) and *P*=.51 (baseline vs second time point 3 months after the treatment)
Zhang et al [48], 2022	2021	China	Parallel group 2-arm randomized controlled trial+2 time point measures (before and after the intervention)	Sex: intervention=female 38/38 (100%); control=female 39/39 (100%); age (y): intervention=mean 52.3 (SD 7.7); control=mean 51.0 (SD 7.9); other: participants with a history of breast cancer and 2 courses of chemotherapy completed	Depressive symptoms measured using the Self-Rating Depression Scale	90 (randomized; intervention: initial=45, at conclusion=38; control: initial=45, at conclusion=39)	Care as usual+VR-CALM[f] intervention (30 min): immersion in calming and beautiful virtual environments (such as the seaside or Butterfly Valley) while receiving CALM therapy delivered by a trained therapist	Care as usual	Intervention: baseline=mean 51.320 (SD 11.552); after treatment=mean 46.630 (SD 9.824); control: baseline=mean 48.640 (SD 4.934); after treatment=mean 50.210 (SD 3.806); test used: paired *t* test (within groups); *P*≤.001 (intervention) and *P*=.14 (control)

Reports	Year of study development	Country of study development	Study design	Participant characteristics	Specific depressive conditions studied and the measurement scale used	Sample size	VR[a] group intervention description	Non-VR group intervention description	Effect
Zhang et al [49], 2023	2021-2022	China	Parallel group 2-arm randomized controlled trial+2 time point measures (before and after the intervention)	Sex: intervention=female 17/30 (57%); control=female 16/30 (53%); age (y): intervention: mean 33.5 (SD 11.1); control: mean 35.3 (SD 10.6); other: participants with leukemia undergoing chemotherapy	Depressive symptoms measured using the Center for Epidemiological Studies Depression Scale	63 (randomized; intervention: initial=32, at conclusion=30; control: initial=31, at conclusion=30)	Usual care: addressing patients' physiological needs and routine psychological care+VR meditation (20 min of 360-degree videos composed of images of landscapes [beach or forest], background music, and meditation guidance)	Usual care: addressing patients' physiological needs and routine psychological care	Intervention: baseline=mean 14.23 (SD 8.11); after treatment=mean 11.13 (SD 6.01); control: baseline=mean 14.03 (SD 7.68); after treatment=mean 14.10 (SD 7.18); test used: paired *t* test (within groups); *P*<.001 (intervention) and *P*=.93 (control); test used: independent 2-tailed *t* test (between groups); *P*=.19 (after treatment)

[a]VR: virtual reality.

[b]GDS-30: Geriatric Depression Scale-30.

[c]ANCOVA: analysis of covariance.

[d]HADS: Hospital Anxiety and Depression Scale.

[e]ICU: intensive care unit.

[f]VR-CALM: Managing Cancer and Living Meaningfully based on VR.

Participant demographics showed that, of the 16 reports, 5 (31%) exclusively involved female participants [34,35,37,46,48]; furthermore, 7 (44%) included participants whose mean or median age was >60 years [34,37-39,43,45,46], while only 2 (13%) included participants whose mean age was <45 years [41,49]. Comorbid conditions were noted as inclusion criteria in 15 (94%) of the 16 studies. Among these 15 studies, cancer [35,48,49], heart disease [37,38,45], and COVID-19 infection [42,44,47] were reported in 3 (20%) studies each. Other conditions such as chronic obstructive pulmonary disease [43], social isolation [36], stroke [39], and spinal cord injury [40] were also mentioned.

Of the 16 reviewed reports, 14 (88%) assessed depressive symptoms [34-38,40-45,47-49]. The most widely used instrument was the Hospital Anxiety and Depression Scale (8/16, 50%) [37-39,42-45,47]. Other questionnaires used to measure depressive conditions included the Beck Depression Inventory [35], the Patient Health Questionnaire-8 [40], the Patient Health Questionnaire-9 [41], the Center for Epidemiological Studies Depression Scale [49], the Depression Anxiety Stress Scale-21 [36], and the Self-Rating Depression Scale [48]. In addition to depressive symptoms, older adult depression was explicitly evaluated in 3 (19%) of the studies [34,39,46], using the Geriatric Depression Scale-30.

Regarding the effects of IMTs on depression, only 1 (6%) of the 16 studies reported no improvement in depression scales [47]. Among the 15 studies that demonstrated a positive effect, 11 (73%) reported statistically significant results [34-37,39,43-46,48,49]; 1 (7%) reported nonsignificant results [38]; 2 (13%) reported mixed results [40,42], indicating variability in statistical significance across different measures or outcomes; and 1 (7%) did not perform any statistical analysis [41].

Technical Aspects of IMT Interventions

The reports included in this review all feature VR interventions, showcasing a broad spectrum of technical characteristics (Table 2). The VR interventions varied, with a majority of the studies (9/16, 56%) implementing the *Virtual Therapeutic Garden* intervention [34,35,37-39,43-46] and 13% (2/16) conducting VR exposure sessions using content from the National Geographic app and the YouTube VR app [40,41]. Unique interventions included *COVID Feel Good* [36], *ICU-VR* [47], *VR-CALM* [48], general VR therapy [42], and individually tailored VR meditation programs.

XSL•FO
RenderX

Table 2. Technical characteristics of the immersive technology interventions.

Reports	VR[a] intervention	Device characteristics	Amount of time used and frequency of use	Intervention duration	Setting of use	Usability assessment	VR design framework used	Degree of guidance
Cieślik et al [34], 2023	Virtual Therapeutic Garden	Type: headset+controllers; specific type: VR TierOne device (VR HTC Vive goggles+HTC Vive controllers)	20 min/session, twice a wk	4 wk	Unclear	Not mentioned	Not mentioned	Self-administered therapy
Czech et al [35], 2022	Virtual Therapeutic Garden	Type: headset+controllers; specific type: VR TierOne device	15 min/session, daily	2 wk (8 sessions)	Hospital (it is not clear whether participants were ambulatory or inpatients)	Not mentioned	Not mentioned	Unclear
Farahimanesh et al [36], 2023	COVID Feel Good	Type: unclear; the study mentions that a head-mounted display or cardboard headset could be used, but there is no information about the VR device type used; specific type: unclear	20 min/session, daily	7 days	Unclear	Not mentioned	Not mentioned	Unclear
Jóźwik et al [37], 2021	Virtual Therapeutic Garden	Type: headset+controllers; specific type: VR TierOne device	Unclear time per session and frequency	Unclear	Hospital (it is not clear whether participants were ambulatory or inpatients)	Not mentioned	Not mentioned	Unclear
Jóźwik et al [38], 2021	Virtual Therapeutic Garden	Type: headset+controllers; specific type: VR TierOne device (HTC Vive PRO VR goggles+unspecified controllers)	20 min/session, 3 times a wk	3 wk (8 sessions)	Hospital (ambulatory)	Not mentioned	Not mentioned	Unclear
Kiper et al [39], 2022	Virtual Therapeutic Garden	Type: headset+controllers; specific type: VR TierOne device (HTC Vive PRO VR goggles+unspecified controllers)	20 min/session; frequency unclear	3 wk (10 sessions)	Hospital (inpatient)	Not mentioned	Methodology of VR clinical trials in health care by an international working group	Unclear
Lakhani et al [40], 2020	VR session of environment exposure using the National Geographic App and the YouTube VR app	Type: headset; specific type: Oculus Go	20 min/session, daily	1 wk (3 sessions)	Hospital (ambulatory)	Not mentioned	Not mentioned	Unclear
Paul et al [41], 2022	VR session of environment exposure using YouTube VR videos	Type: headset; specific type: Limbix	Time per session and frequency unclear	3 wk	Home based	Not mentioned	Not mentioned	Self-administered therapy

Reports	VR[a] intervention	Device characteristics	Amount of time used and frequency of use	Intervention duration	Setting of use	Usability assessment	VR design framework used	Degree of guidance
Rodrigues et al [42], 2022	VR therapy (360-degree videos with images of landscapes and mindfulness techniques)	Type: smartphone VR headset; specific type: Oculos Realidade Virtual 3D Gamer Warrior JS080 (adaptable with smartphone)	10 min/session, once	1 session	Hospital (inpatient)	Not mentioned	Not mentioned	Predominantly self-help
Rutkowski et al [43], 2021	Virtual Therapeutic Garden	Type: headset+controllers; specific type: VR TierOne device	20 min/session, 5 times a wk	2 wk	Hospital (it is not clear whether participants were ambulatory or inpatients)	Not mentioned	Not mentioned	Unclear
Rutkowski et al [44], 2022	Virtual Therapeutic Garden	Type: headset; specific type: VR TierOne device (VR goggles+controllers)	Unclear time per session, 5 times a wk	3 wk	Hospital (inpatient)	Not mentioned	Not mentioned	Unclear
Szczepańska-Gieracha et al [45], 2021	Virtual Therapeutic Garden	Type: headset; specific type: VR TierOne device (HTC Vive PRO VR goggles+unspecified controllers)	20 min/session, twice a wk	4 wk (8 sessions)	Hospital (inpatient)	Not mentioned	Methodology of VR clinical trials in health care by an international working group	Predominantly self-help
Szczepańska-Gieracha et al [46], 2021	Virtual Therapeutic Garden	Type: headset+controllers; specific type: VR TierOne device	20 min/session, twice a wk	4 wk	Unclear	Not mentioned	Methodology of VR clinical trials in health care by an international working group	Unclear
Vlake et al [47], 2022	ICU-VR	Type: headset+headphones; specific type: Oculus Go+unspecified headphones	14 min/session, once	1 session	Hospital (ambulatory)	Not mentioned	Not mentioned	Predominantly self-help
Zhang et al [48], 2022	VR-CALM	Type: headset+controllers; specific type: unclear	30 min/session; frequency unclear	3 mo (6 sessions)	Hospital (inpatient)	Not mentioned	Not mentioned	Minimal-contact therapy
Zhang et al [49], 2023	Tailored VR meditation	Type: headset; specific type: PRO 6 DOF (Beijing Iqiyi Intelligent Technology)	20 min/session, daily	14 d	Hospital (inpatient)	Not mentioned	Not mentioned	Predominantly self-help

[a]VR: virtual reality.

The immersion devices predominantly used were stand-alone VR headsets (14/16, 88%) [34,35,37-41,43-49]. Of the remaining 2 studies, 1 (50%) used VR headset adapters for smartphones [42], whereas 1 (50%) did not specify the type of device used [36]. The types of specific VR devices varied, with the VR TierOne device (9/16, 56%) [34,35,37-39,43-46] and Oculus Go (2/16, 13%) being the most commonly reported [40,47].

The duration of VR interventions ranged from a single session to daily sessions over up to 8 weeks. However, most of the interventions (10/16, 63%) were administered over a period of 1 day [42,47] to 3 months [48], with 3 weeks (4/16, 25%) being the most common intervention period [38,39,41,44]. The frequency of sessions varied, with daily sessions reported in 4 (25%) of the 16 reports [35,36,40,49] and less frequent sessions noted in the others (12/16, 75%). The majority of the interventions (9/16, 56%) were to be used for 20 minutes per session [34,36,38-40,43,45,46,49]; 3 (19%) of the 16 reports did not specify the session length [37,41,44].

The settings for the interventions were predominantly hospital based (11/16, 69%), with a mix of inpatient, ambulatory, or unclear settings. Only 2 (13%) of the 16 reports specified the intervention as self-administered therapy, with the remaining studies (14/16, 88%) not clearly reporting the degree of guidance provided. None of the reports mentioned usability evaluation as part of their methods or results.

Regarding the VR design framework, only 2 (13%) of the 16 reports mentioned using the *methodology of VR clinical trials in health care by an international working group*. The rest of the reports (14/16, 88%) did not mention any design framework used.

Therapeutic Approaches

The analysis of therapeutic approaches used in the IMT interventions reveals a varied landscape of techniques (Table 3). Ericksonian psychotherapy was used in 9 (56%) of the 16 reports, indicating its recognized role in VR-based interventions. Mindfulness-based cognitive therapy was the second leading therapeutic approach featured, used in 13% (2/16) of the reports, while cognitive behavioral therapy and behavioral activation were mentioned in only 6% (1/16) of the reports each. Other therapeutic methods included personal psychotherapy based on Managing Cancer and Living Meaningfully (1/16, 6%) [48] and the Roy Adaptation Model (1/16, 6%) [49].

Table 3. Therapeutic approaches used in the immersive technology interventions.

Reports	Ericksonian psychotherapy	Mindfulness-based cognitive therapy	Cognitive behavioral therapy	Behavioral activation	Others
Cieślik et al [34], 2023	✓				
Czech et al [35], 2022	✓				
Farahimanesh et al [36], 2023			✓		
Jóźwik et al [37], 2021	✓				
Jóźwik et al [38], 2021	✓				
Kiper et al [39], 2022	✓				
Lakhani et al [40], 2020					
Paul et al [41], 2022				✓	
Rodrigues et al [42], 2022		✓[a]			
Rutkowski et al [43], 2021	✓				
Rutkowski et al [44], 2022	✓				
Szczepańska-Gieracha et al [45], 2021	✓				
Szczepańska-Gieracha et al [46], 2021	✓	✓			
Vlake et al [47], 2022					
Zhang et al [48], 2022					✓
Zhang et al [49], 2023					✓

[a]It is unclear whether all patients in the virtual reality group received a mindfulness-based intervention or just 360-degree videos.

Implementation Characteristics

Regarding the stages of implementation, 2 (13%) of the 16 articles were at a *preliminary* stage (pilot or feasibility study). By contrast, 6 (38%) of the 16 studies had progressed to the *implementation* stage (referring to prior feasibility results); the remaining articles (8/16, 50%) did not clearly delineate their implementation phase. No article reported an implementation framework used, nor were barriers or enablers of implementation identified.

Most of the reports (15/16, 94%) declared no conflicts of interest; an exception was found in 1 (6%) of the 16 articles, where a potential conflict was disclosed due to an author's corporate affiliation, emphasizing the importance of transparency in research affiliations and possible biases. Funding sources varied, with half of the studies (8/16, 50%) reporting no external funding, suggesting that a significant portion of research in this area is conducted independently of external financial support. Notably, individual studies were supported by grants from prominent institutions such as the National Science Foundation of China as well as national research grants, demonstrating the global investment in health VR intervention research. All included reports reported ethical oversight, ranging from institutional review boards to national ethics committees.

Discussion

Principal Findings

Our scoping review, conducted in accordance with the framework formulated by Arksey and O'Malley [26], identified a total of 16 peer-reviewed articles focusing on the use of IMTs in treating depression. In comparison, other reviews using similar methodologies and research questions have reported varying numbers of studies; for instance, Fodor et al [23] found 24 studies examining the effects of VR interventions on depressive outcomes, whereas Zeng et al [24] included only 5 studies in their review. The larger number of studies in the first review can be attributed to the inclusion of several articles that measured depression as a secondary outcome, which is a common approach in the literature on VR interventions. By contrast, our scoping review strictly included reports that explicitly measured depression as a primary outcome. This was done to specifically target papers that focused on designing and developing interventions addressing this mental disorder.

Regarding the second review, the limited number of studies included can be primarily attributed to the authors' focus on exercise interventions using VR. However, this explanation might not be complete. During our full-text review phase, we identified numerous studies assessing VR interventions related to exercise among the excluded studies (36/52, 69%). Another possible explanation for the limited number of studies is the year of the review's development: 2018. Bibliometric studies on VR in health care [50] and specifically in depression [51] have shown a significant upward trend since that year, likely linked to the increased availability of VR technologies [52]. This aligns with our findings, where we observed a clear upward trend over the years.

Secondary Questions

From Which Regions or Countries Does the Evidence Come?

Research on IMT interventions for depression care was predominantly conducted in Europe (10/16, 63%), with Poland contributing the most studies (9/16, 56%), indicating a robust regional focus within the field. Our results align with what has been previously observed in the literature, where it is noted that most articles published on VR in depression originate from Europe [51]. This trend denotes a divergence in the topics related to VR in health because most articles on VR in health, in general, have been published in the United States [50].

These figures suggest that European high-income countries exhibit a more consistent and robust research trajectory related to IMT-based treatments for depression. By contrast, other high-income countries such as the United States and Australia and upper–middle-income countries such as Brazil show sporadic participation. Nevertheless, it is essential to approach these figures cautiously, recognizing that the frequency of publications within specific years might not accurately reflect ongoing research interest or the immediacy of research outputs, given the cyclical nature of research funding, project development, and publication processes.

Which Technical Aspects of IMTs Have Been Reported in the Evidence?

In our review, no study reported using AR; instead, stand-alone VR headsets emerged as the primary technology, underscoring a trend toward self-contained IMT devices in treating depression. This observation aligns with existing evidence; for example, a systematic review concentrating on the mental health applications of AR did not reveal any applications of this particular IMT modality in either the treatment or the assessment of depression [53].

In terms of duration, most of the VR interventions in our review (10/16, 63%) ranged from 2 to 8 weeks, encompassing 1 to 10 sessions. This aligns with the range of 1 to 16 sessions reported in similar VR studies [23]. In addition, we observed that the predominant setting for these interventions was hospital based (11/16, 69%), with only 6% (1/16) being delivered in home settings. This finding aligns with existing evidence in mindfulness VR interventions, where only 1 of 15 studies was home based [54]. Such a trend indicates a current focus on clinical settings for VR intervention delivery, suggesting potential areas for expansion into more accessible home-based environments.

Only 2 (13%) of the 16 studies reported using a specific IMT design framework, pointing to a potential area for standardization in future research. Conceptual and methodological frameworks are pivotal because they provide a structured approach, align the study's methodology with its objectives, and facilitate the integration of technology to achieve therapeutic goals [55]. Their application in IMT research is essential for producing reliable and applicable results, particularly in the intricate mental health field [56]. This underuse highlights the need for a more structured and theoretically informed approach in future research, which could enhance the quality, applicability, and standardization of IMT interventions for treating depression.

What Therapeutic Approaches Were Used?

Ericksonian psychotherapy was the most common therapeutic approach incorporated into the VR interventions (9/16, 56%). The Ericksonian approach to psychotherapy and hypnosis is based on three key assumptions: (1) the belief in an altered state of consciousness and the existence of specific markers indicating this altered state, (2) the superiority of indirect over direct suggestion in therapy, and (3) the view that a patient's hypnotizability is a function of the hypnotist's skill. However, empirical support for the validity of these critical assumptions is limited [57]. Notably, most studies using this approach originated from Poland (9/16, 56%), indicating a geographic concentration of the evidence. Therefore, there is a need to evaluate this intervention in diverse settings to validate its efficacy more broadly. Despite the geographic concentration of Ericksonian therapy within VR interventions, the use of hypnosis and mindfulness techniques can be advantageous in both face-to-face psychotherapy and virtual contexts. These techniques can alleviate life problems and symptoms associated with mental disorders, including depression [58].

Mindfulness-based cognitive therapy has shown effectiveness in reducing depressive symptoms and elucidating the active mechanisms during mindfulness [59,60]. By contrast, cognitive behavioral therapy and behavioral activation are considered therapies with solid evidence for reducing depression [61]. Thus, behavioral therapies may possess a more robust theoretical basis than other treatment models in IMT interventions for depression care, suggesting a potential direction for future research and application.

What Are the Barriers and Facilitators to Implementing IMT Interventions for Depression Treatment?

The included studies do not provide evidence on barriers and facilitators to implementation. One possible reason is that we did not include qualitative studies in our scoping review (qualitative research focuses on these types of outcomes). However, a framework for implementing digital mental health interventions identified the key elements: access to the intervention, cost-effectiveness, and user satisfaction, in addition to the evaluation of the effectiveness of the intervention [6].

The primary facilitator for the implementation of IMT interventions for depressive symptoms described in the literature is the availability of evidence supporting the efficacy of the treatment [23,24]. There is also evidence of VR's acceptability, feasibility, and user satisfaction in mental health settings [62]. However, the cost of VR equipment and the cost of training health professionals may be barriers to access in low- and middle-income countries. In addition, we found no evidence of the cost-effectiveness of IMT interventions on mental health outcomes within the health care system. Therefore, cost-effectiveness and cost-utility studies compared with usual care or other psychological interventions must be developed to provide sufficient evidence to evaluate the implementation of IMT interventions within public health systems.

What Outcomes Have Been Evaluated in Studies Examining the Impact of IMT Interventions on Addressing Depression?

In general, all studies used a psychometric scale to assess the impact of the intervention on depressive symptoms, and the instruments used have evidence of reliability and validity; therefore, we considered the results to be adequately assessed. However, there was a high degree of heterogeneity in the instruments used. Some studies used scales focused on hospital settings (eg, the Hospital Anxiety and Depression Scale), others used scales designed for geriatric use (eg, the Geriatric Depression Scale-30), and still others used specialized instruments developed for population use (eg, the Patient Health Questionnaire). Although all instruments assess depressive symptoms, they may assess different forms of the presence of depressive symptoms. Older adults should be considered to have manifestations of depression that are clinically different from those of adults with depression [63]. Therefore, it is essential to consider the setting in which each study was conducted when comparing results.

Strengths and Limitations

To our knowledge, this study is one of the first to comprehensively identify the current state of research regarding the use of IMT interventions specifically focused on depression. Our work fills essential gaps in existing literature by mapping the current evidence and providing insightful recommendations for future research development. Additional strengths of this study include providing valuable insights into the geographic distribution of research efforts and the range of therapeutic approaches used. This contributes significantly to a deeper understanding of the field and highlights areas where further research is needed.

This study has significant limitations that deserve consideration. First, our selection process, involving screening by title and abstract followed by full-text review, may have introduced selection bias if relevant studies were inadvertently excluded due to inadequate information in titles or abstracts. To mitigate this risk, we used a thorough screening process with multiple reviewers for each study, aiming to reduce selection bias.

Second, our search was limited to articles published in English, potentially leading to language bias by excluding relevant studies in other languages. While future reviews could include studies in multiple languages for a broader range of evidence, this limitation did not significantly narrow the scope of our review because a substantial portion of the evidence in this field is published in English.

Third, the studies included were restricted to RCTs. While RCTs are considered the gold standard in clinical research because they provide high-quality evidence on the efficacy of interventions, this restriction may have limited the comprehensiveness of our review. Specifically, valuable exploratory, observational, and qualitative studies that could provide insights into the implementation, user experience, and broader contextual factors related to IMTs in depression care were excluded. Future reviews could consider including a wider range of study designs to provide a more holistic view of the field, thereby enhancing our understanding of the efficacy of these technologies and their practical application, barriers to implementation, and patient perspectives.

Fourth, we did not consider the high cost of IMT equipment, the training of health professionals, and other economic aspects in the study extraction process, which could be significant barriers, especially in low- and middle-income countries. This oversight underscores the necessity for cost-effectiveness and cost-utility studies to assess the feasibility of IMT interventions in diverse health care settings.

Fifth, the considerable variability in the psychometric scales used across the studies could impact result comparability. We recognize this heterogeneity and recommend that future research consider setting and population-specific scales to improve comparability.

Finally, our scoping review did not include a formal quality appraisal of the included studies. However, it is important to note that our review focused exclusively on RCTs published in peer-reviewed journals. This focus on RCTs, combined with the peer-review process, increased the likelihood that high-quality studies were included. While this approach does not guarantee the quality of each study, it does suggest that the evidence base we have mapped is likely to be more rigorous and high-quality research compared to broader inclusion criteria.

Conclusions

Our scoping review on the use of IMTs for treating depression identified 16 peer-reviewed articles predominantly focused on stand-alone VR headsets. Most of the research was concentrated in Europe (10/16, 63%), specifically Poland (9/16, 56%), suggesting a need for more geographically diverse studies. Furthermore, the therapeutic approaches in these studies largely centered around Ericksonian psychotherapy; however, given the limited empirical support for the fundamental assumptions of Ericksonian psychotherapy and the geographic bias, there is a clear need for exploring a variety of therapeutic approaches in IMT interventions for depression care.

A notable gap in the literature is the absence of AR approaches for depression treatment in the studies reviewed. This points toward an opportunity for future research in this area. In addition, while VR shows promise in mental health settings, concerns about the cost and accessibility, particularly in low- and middle-income countries, highlight the need for more research into the cost-effectiveness of these interventions.

In summary, while the use of IMTs in treating depression shows promise, our review indicates the need for more diverse, inclusive, and comprehensive research. Future studies should address the identified gaps, particularly in AR, cost-effectiveness, and geographic diversity, to fully harness the potential of IMT interventions in depression care.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[DOCX File , 108 KB - mental_v11i1e56056_app1.docx]

Multimedia Appendix 2
Search strategies by database consulted.
[DOCX File , 26 KB - mental_v11i1e56056_app2.docx]

Multimedia Appendix 3
List of articles excluded from the full-text review (36/52, 69%).
[DOCX File , 14 KB - mental_v11i1e56056_app3.docx]

References

1. Depressive disorder (depression). World Health Organization. 2023 Mar 31. URL: https://www.who.int/news-room/fact-sheets/detail/depression [accessed 2023-04-18]
2. GBD results. Institute for Health Metrics and Evaluation. URL: http://ghdx.healthdata.org/gbd-results-tool [accessed 2024-04-15]
3. Thornicroft G, Chatterji S, Evans-Lacko S, Gruber M, Sampson N, Aguilar-Gaxiola S, et al. Undertreatment of people with major depressive disorder in 21 countries. Br J Psychiatry 2017 Feb;210(2):119-124 [FREE Full text] [doi: 10.1192/bjp.bp.116.188078] [Medline: 27908899]
4. Mental health Atlas 2020. World Health Organization. 2021 Oct 08. URL: https://www.who.int/publications-detail-redirect/9789240036703 [accessed 2024-02-22]
5. Demyttenaere K, Bruffaerts R, Posada-Villa J, Gasquet I, Kovess V, Lepine JP, et al. Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. JAMA 2004 Jun 02;291(21):2581-2590. [doi: 10.1001/jama.291.21.2581] [Medline: 15173149]
6. Villarreal-Zegarra D, Alarcon-Ruiz CA, Melendez-Torres GJ, Torres-Puente R, Navarro-Flores A, Cavero V, et al. Development of a framework for the implementation of synchronous digital mental health: realist synthesis of systematic reviews. JMIR Ment Health 2022 Mar 29;9(3):e34760 [FREE Full text] [doi: 10.2196/34760] [Medline: 35348469]
7. Samuel S. Immersive technologies in the healthcare space. In: Mishra S, Tripathy HK, Mallick P, Shaalan K, editors. Augmented Intelligence in Healthcare: A Pragmatic and Integrated Analysis. Singapore, Singapore: Springer; Apr 20, 2022.
8. Sukhato K, Lotrakul M, Dellow A, Ittasakul P, Thakkinstian A, Anothaisintawee T. Efficacy of home-based non-pharmacological interventions for treating depression: a systematic review and network meta-analysis of randomised controlled trials. BMJ Open 2017 Jul 12;7(7):e014499 [FREE Full text] [doi: 10.1136/bmjopen-2016-014499] [Medline: 28706086]
9. Platt LM, Whitburn AI, Platt-Koch AG, Koch RL. Nonpharmacological alternatives to benzodiazepine drugs for the treatment of anxiety in outpatient populations: a literature review. J Psychosoc Nurs Ment Health Serv 2016 Aug 01;54(8):35-42. [doi: 10.3928/02793695-20160725-07] [Medline: 27479478]
10. Suh A, Prophet J. The state of immersive technology research: a literature analysis. Comput Hum Behav 2018 Sep;86:77-90. [doi: 10.1016/j.chb.2018.04.019]
11. Pasco D. The potential of using virtual reality technology in physical activity settings. Quest 2013;65(4):429-441. [doi: 10.1080/00336297.2013.795906]
12. Mouatt B, Smith AE, Mellow ML, Parfitt G, Smith RT, Stanton TR. The use of virtual reality to influence motivation, affect, enjoyment, and engagement during exercise: a scoping review. Front Virtual Real 2020 Dec 23;1 [FREE Full text] [doi: 10.3389/frvir.2020.564664]
13. Jahn FS, Skovbye M, Obenhausen K, Jespersen AE, Miskowiak KW. Cognitive training with fully immersive virtual reality in patients with neurological and psychiatric disorders: a systematic review of randomized controlled trials. Psychiatry Res 2021 Jun;300:113928 [FREE Full text] [doi: 10.1016/j.psychres.2021.113928] [Medline: 33857847]

14. Dixon LB, Holoshitz Y, Nossel I. Treatment engagement of individuals experiencing mental illness: review and update. World Psychiatry 2016 Feb;15(1):13-20 [FREE Full text] [doi: 10.1002/wps.20306] [Medline: 26833597]
15. Hasan S, Alhaj H, Hassoulas A. The efficacy and therapeutic alliance of augmented reality exposure therapy in treating adults with phobic disorders: systematic review. JMIR Ment Health 2023 Nov 30;10:e51318 [FREE Full text] [doi: 10.2196/51318] [Medline: 38032710]
16. Jingili N, Oyelere SS, Nyström MB, Anyshchenko L. A systematic review on the efficacy of virtual reality and gamification interventions for managing anxiety and depression. Front Digit Health 2023 Nov 7;5:1239435 [FREE Full text] [doi: 10.3389/fdgth.2023.1239435] [Medline: 38026832]
17. Rosa A, Pujia AM, Docimo R, Arcuri C. Managing dental phobia in children with the use of virtual reality: a systematic review of the current literature. Children (Basel) 2023 Oct 30;10(11):1763 [FREE Full text] [doi: 10.3390/children10111763] [Medline: 38002854]
18. Schröder D, Wrona KJ, Müller F, Heinemann S, Fischer F, Dockweiler C. Impact of virtual reality applications in the treatment of anxiety disorders: a systematic review and meta-analysis of randomized-controlled trials. J Behav Ther Exp Psychiatry 2023 Dec;81:101893 [FREE Full text] [doi: 10.1016/j.jbtep.2023.101893] [Medline: 37453405]
19. Heo S, Park J. Effects of virtual reality-based graded exposure therapy on PTSD symptoms: a systematic review and meta-analysis. Int J Environ Res Public Health 2022 Nov 29;19(23):15911 [FREE Full text] [doi: 10.3390/ijerph192315911] [Medline: 36497989]
20. Wiebe A, Kannen K, Selaskowski B, Mehren A, Thöne AK, Pramme L, et al. Virtual reality in the diagnostic and therapy for mental disorders: a systematic review. Clin Psychol Rev 2022 Dec;98:102213 [FREE Full text] [doi: 10.1016/j.cpr.2022.102213] [Medline: 36356351]
21. Mittmann G, Zehetner V, Hoehl S, Schrank B, Barnard A, Woodcock K. Using augmented reality toward improving social skills: scoping review. JMIR Serious Games 2023 Sep 20;11:e42117 [FREE Full text] [doi: 10.2196/42117] [Medline: 37728971]
22. Shahmoradi L, Rezayi S. Cognitive rehabilitation in people with autism spectrum disorder: a systematic review of emerging virtual reality-based approaches. J Neuroeng Rehabil 2022 Aug 18;19(1):91 [FREE Full text] [doi: 10.1186/s12984-022-01069-5] [Medline: 35982460]
23. Fodor LA, Cote CD, Cuijpers P, Szamoskozi , David D, Cristea IA. The effectiveness of virtual reality based interventions for symptoms of anxiety and depression: a meta-analysis. Sci Rep 2018 Jul 09;8(1):10323 [FREE Full text] [doi: 10.1038/s41598-018-28113-6] [Medline: 29985400]
24. Zeng N, Pope Z, Lee JE, Gao Z. Virtual reality exercise for anxiety and depression: a preliminary review of current research in an emerging field. J Clin Med 2018 Mar 04;7(3):42 [FREE Full text] [doi: 10.3390/jcm7030042] [Medline: 29510528]
25. Baghaei N, Chitale V, Hlasnik A, Stemmet L, Liang HN, Porter R. Virtual reality for supporting the treatment of depression and anxiety: scoping review. JMIR Ment Health 2021 Sep 23;8(9):e29681 [FREE Full text] [doi: 10.2196/29681] [Medline: 34554097]
26. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol 2005;8(1):19-32. [doi: 10.1080/1364557032000119616]
27. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci 2010 Sep 20;5:69 [FREE Full text] [doi: 10.1186/1748-5908-5-69] [Medline: 20854677]
28. Daudt HM, van Mossel C, Scott SJ. Enhancing the scoping study methodology: a large, inter-professional team's experience with Arksey and O'Malley's framework. BMC Med Res Methodol 2013 Mar 23;13:48 [FREE Full text] [doi: 10.1186/1471-2288-13-48] [Medline: 23522333]
29. Aromataris E, Lockwood C, Porritt K, Pilla B, Jordan Z. JBI Manual for Evidence Synthesis. Adelaide, Australia: Joanna Briggs Institute; 2024.
30. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
31. Reategui-Rivera CM, Finkelstein J, De la Cruz K, Diaz-Sanchez P, Villarreal-Zegarra D. A scoping review to mapping the evidence and gaps related to the use of virtual and augmented reality to address depressive disorders. Open Science Framework Home. 2023 Oct 19. URL: https://osf.io/px9sb/ [accessed 2024-04-15]
32. Hopewell S, Dutton S, Yu LM, Chan AW, Altman DG. The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in PubMed. BMJ 2010 Mar 23;340:c723 [FREE Full text] [doi: 10.1136/bmj.c723] [Medline: 20332510]
33. Rodgers M, Sowden A, Petticrew M, Arai L, Roberts H, Britten N, et al. Testing methodological guidance on the conduct of narrative synthesis in systematic reviews: effectiveness of interventions to promote smoke alarm ownership and function. Evaluation 2009;15(1):49-73. [doi: 10.1177/1356389008097871]
34. Cieślik B, Juszko K, Kiper P, Szczepańska-Gieracha J. Immersive virtual reality as support for the mental health of elderly women: a randomized controlled trial. Virtual Real 2023 May 07 (forthcoming) [FREE Full text] [doi: 10.1007/s10055-023-00797-w] [Medline: 37360811]

35. Czech O, Siewierska K, Krzywińska A, Skórniak J, Maciejczyk A, Matkowski R, et al. Virtual therapy complementary prehabilitation of women diagnosed with breast cancer-a pilot study. Int J Environ Res Public Health 2022 Dec 30;20(1):722 [FREE Full text] [doi: 10.3390/ijerph20010722] [Medline: 36613047]
36. Farahimanesh S, Serino S, Tuena C, Di Lernia D, Wiederhold BK, Bernardelli L, et al. Effectiveness of a virtual-reality-based self-help intervention for lowering the psychological burden during the COVID-19 pandemic: results from a randomized controlled trial in Iran. J Clin Med 2023 Mar 02;12(5):2006 [FREE Full text] [doi: 10.3390/jcm12052006] [Medline: 36902793]
37. Jóźwik S, Cieślik B, Gajda R, Szczepańska-Gieracha J. The use of virtual therapy in cardiac rehabilitation of female patients with heart disease. Medicina (Kaunas) 2021 Jul 28;57(8):768 [FREE Full text] [doi: 10.3390/medicina57080768] [Medline: 34440974]
38. Jóźwik S, Cieślik B, Gajda R, Szczepańska-Gieracha J. Evaluation of the impact of virtual reality-enhanced cardiac rehabilitation on depressive and anxiety symptoms in patients with coronary artery disease: a randomised controlled trial. J Clin Med 2021 May 16;10(10):2148 [FREE Full text] [doi: 10.3390/jcm10102148] [Medline: 34065625]
39. Kiper P, Przysiężna E, Cieślik B, Broniec-Siekaniec K, Kucińska A, Szczygieł J, et al. Effects of immersive virtual therapy as a method supporting recovery of depressive symptoms in post-stroke rehabilitation: randomized controlled trial. Clin Interv Aging 2022 Nov 23;17:1673-1685 [FREE Full text] [doi: 10.2147/CIA.S375754] [Medline: 36447623]
40. Lakhani A, Martin K, Gray L, Mallison J, Grimbeek P, Hollins I, et al. What is the impact of engaging with natural environments delivered via virtual reality on the psycho-emotional health of people with spinal cord injury receiving rehabilitation in hospital? Findings from a pilot randomized controlled trial. Arch Phys Med Rehabil 2020 Sep;101(9):1532-1540. [doi: 10.1016/j.apmr.2020.05.013] [Medline: 32502564]
41. Paul M, Bullock K, Bailenson J. Virtual reality behavioral activation for adults with major depressive disorder: feasibility randomized controlled trial. JMIR Ment Health 2022 May 06;9(5):e35526 [FREE Full text] [doi: 10.2196/35526] [Medline: 35404830]
42. Rodrigues IM, Lima AG, Santos AE, Santos AC, Nascimento LS, Serra MV, et al. A single session of virtual reality improved tiredness, shortness of breath, anxiety, depression and well-being in hospitalized individuals with COVID-19: a randomized clinical trial. J Pers Med 2022 May 19;12(5):829 [FREE Full text] [doi: 10.3390/jpm12050829] [Medline: 35629250]
43. Rutkowski S, Szczegielniak J, Szczepańska-Gieracha J. Evaluation of the efficacy of immersive virtual reality therapy as a method supporting pulmonary rehabilitation: a randomized controlled trial. J Clin Med 2021 Jan 18;10(2):352 [FREE Full text] [doi: 10.3390/jcm10020352] [Medline: 33477733]
44. Rutkowski S, Bogacz K, Czech O, Rutkowska A, Szczegielniak J. Effectiveness of an inpatient virtual reality-based pulmonary rehabilitation program among COVID-19 patients on symptoms of anxiety, depression and quality of life: preliminary results from a randomized controlled trial. Int J Environ Res Public Health 2022 Dec 17;19(24):16980 [FREE Full text] [doi: 10.3390/ijerph192416980] [Medline: 36554860]
45. Szczepańska-Gieracha J, Jóźwik S, Cieślik B, Mazurek J, Gajda R. Immersive virtual reality therapy as a support for cardiac rehabilitation: a pilot randomized-controlled trial. Cyberpsychol Behav Soc Netw 2021 Aug;24(8):543-549 [FREE Full text] [doi: 10.1089/cyber.2020.0297] [Medline: 33577375]
46. Szczepańska-Gieracha J, Cieślik B, Serweta A, Klajs K. Virtual therapeutic garden: a promising method supporting the treatment of depressive symptoms in late-life: a randomized pilot study. J Clin Med 2021 May 01;10(9):1942 [FREE Full text] [doi: 10.3390/jcm10091942] [Medline: 34062721]
47. Vlake JH, van Bommel J, Wils EJ, Bienvenu J, Hellemons ME, Korevaar TI, et al. Intensive care unit-specific virtual reality for critically ill patients with COVID-19: multicenter randomized controlled trial. J Med Internet Res 2022 Jan 31;24(1):e32368 [FREE Full text] [doi: 10.2196/32368] [Medline: 34978530]
48. Zhang X, Yao S, Wang M, Yin X, Bi Z, Jing Y, et al. The impact of VR-CALM intervention based on VR on psychological distress and symptom management in breast cancer survivors. J Oncol 2022 Jun 7;2022:1012813-1012810 [FREE Full text] [doi: 10.1155/2022/1012813] [Medline: 35712124]
49. Zhang B, Jin X, Kuang X, Shen B, Qiu D, Peng J, et al. Effects of a virtual reality-based meditation intervention on anxiety and depression among patients with acute leukemia during induction chemotherapy: a randomized controlled trial. Cancer Nurs 2023 Jan 23 (forthcoming). [doi: 10.1097/NCC.0000000000001206] [Medline: 36693237]
50. Pawassar CM, Tiberius V. Virtual reality in health care: bibliometric analysis. JMIR Serious Games 2021 Dec 01;9(4):e32721 [FREE Full text] [doi: 10.2196/32721] [Medline: 34855606]
51. Jingili N, Oyelere SS, Ojwang F, Agbo FJ, Nyström MB. Virtual reality for addressing depression and anxiety: a bibliometric analysis. Int J Environ Res Public Health 2023 Apr 24;20(9):5621 [FREE Full text] [doi: 10.3390/ijerph20095621] [Medline: 37174141]
52. Castelvecchi D. Low-cost headsets boost virtual reality's lab appeal. Nature 2016 May 12;533(7602):153-154. [doi: 10.1038/533153a] [Medline: 27172022]
53. Bakır CN, Abbas SO, Sever E, Özcan Morey A, Aslan Genç H, Mutluer T. Use of augmented reality in mental health-related conditions: a systematic review. Digit Health 2023;9:20552076231203649 [FREE Full text] [doi: 10.1177/20552076231203649] [Medline: 37791140]

54. Riches S, Jeyarajaguru P, Taylor L, Fialho C, Little J, Ahmed L, et al. Virtual reality relaxation for people with mental health conditions: a systematic review. Soc Psychiatry Psychiatr Epidemiol 2023 Jul;58(7):989-1007 [FREE Full text] [doi: 10.1007/s00127-022-02417-5] [Medline: 36658261]
55. Ravitch S, Carl N. Chapter 2: conceptual frameworks in research. In: Qualitative Research Bridging the Conceptual, Theoretical, and Methodological. Thousand Oaks, CA: SAGE Publications, Inc; 2020.
56. Birckhead B, Khalil C, Liu X, Conovitz S, Rizzo A, Danovitch I, et al. Recommendations for methodology of virtual reality clinical trials in health care by an international working group: iterative study. JMIR Ment Health 2019 Jan 31;6(1):e11973 [FREE Full text] [doi: 10.2196/11973] [Medline: 30702436]
57. Matthews WJ. Ericksonian approaches to hypnosis and therapy: where are we now? Int J Clin Exp Hypn 2000 Oct;48(4):418-26; discussion 433. [doi: 10.1080/00207140008410370] [Medline: 11011501]
58. Green J, Laurence J, Lynn S. Hypnosis and psychotherapy: from Mesmer to mindfulness. Psychol Conscious Theory Res Pract 2014 Jun;1(2):199-212. [doi: 10.1037/cns0000015]
59. Musa ZA, Kim Lam S, Binti Mamat @ Mukhtar F, Kwong Yan S, Tajudeen Olalekan O, Kim Geok S. Effectiveness of mindfulness-based cognitive therapy on the management of depressive disorder: systematic review. Int J Afr Nurs Sci 2020;12:100200. [doi: 10.1016/j.ijans.2020.100200]
60. Nakamura H, Tawatsuji Y, Fang S, Matsui T. Explanation of emotion regulation mechanism of mindfulness using a brain function model. Neural Netw 2021 Jun;138:198-214 [FREE Full text] [doi: 10.1016/j.neunet.2021.01.029] [Medline: 33684653]
61. Barth J, Munder T, Gerger H, Nüesch E, Trelle S, Znoj H, et al. Comparative efficacy of seven psychotherapeutic interventions for patients with depression: a network meta-analysis. PLoS Med 2013;10(5):e1001454 [FREE Full text] [doi: 10.1371/journal.pmed.1001454] [Medline: 23723742]
62. Humbert A, Kohls E, Baldofski S, Epple C, Rummel-Kluge C. Acceptability, feasibility, and user satisfaction of a virtual reality relaxation intervention in a psychiatric outpatient setting during the COVID-19 pandemic. Front Psychiatry 2023;14:1271702 [FREE Full text] [doi: 10.3389/fpsyt.2023.1271702] [Medline: 37953932]
63. Pocklington C. Depression in older adults. Brit J Med Pract 2017 Mar;10(1):a1007 [FREE Full text]

Abbreviations

AR: augmented reality
IMT: immersive technology
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RCT: randomized clinical trial
VR: virtual reality

Edited by J Torous; submitted 09.01.24; peer-reviewed by JD Carrier, M Sadiq, K Woodcock, N Jingili; comments to author 06.02.24; revised version received 27.02.24; accepted 16.03.24; published 25.04.24.

Please cite as:
Reategui-Rivera CM, Villarreal-Zegarra D, De La Cruz-Torralva K, Díaz-Sánchez P, Finkelstein J
Immersive Technologies for Depression Care: Scoping Review
JMIR Ment Health 2024;11:e56056
URL: https://mental.jmir.org/2024/1/e56056
doi:10.2196/56056
PMID:38663004

Original Paper

Virtual Reality Exposure Therapy for Reducing School Anxiety in Adolescents: Pilot Study

Gesa Beele[1*], MSc; Paula Liesong[1*], MSc; Sabine Bojanowski[1], PhD; Kristian Hildebrand[2], Prof Dr; Malte Weingart[2], BSc; Julia Asbrand[3], Prof Dr; Christoph U Correll[1,4,5], Prof Dr Med; Nexhmedin Morina[6], Prof Dr; Peter J Uhlhaas[1,7], Prof Dr

[1]Department of Child and Adolescent Psychiatry, Charité-Universitätsmedizin Berlin, Berlin, Germany
[2]Berlin University of Applied Sciences and Technology, Berlin, Germany
[3]Department of Clinical Psychology for Childhood and Adolescence, University of Jena, Jena, Germany
[4]Department of Psychiatry and Molecular Medicine, Donald and Barbara Zucker School of Medicine at Hostra/Northwell, Hempstead, NY, United States
[5]Department of Psychiatry, The Zucker Hillside Hospital, Northwell Health, Glen Oaks, NY, United States
[6]Institute of Psychology, University of Münster, Münster, Germany
[7]Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, United Kingdom
[*]these authors contributed equally

Corresponding Author:
Peter J Uhlhaas, Prof Dr
Department of Child and Adolescent Psychiatry
Charité-Universitätsmedizin Berlin
Augustenburgerplatz 1
Berlin, 13353
Germany
Phone: 49 30450 ext 516193
Email: peter.uhlhaas@charite.de

Abstract

Background: Virtual reality exposure therapy (VRET) is a promising treatment approach for anxiety disorders. However, while its efficacy has been demonstrated in adults, research on the efficacy of VRET in the treatment of adolescents with anxiety disorders is largely lacking.

Objective: A pilot study was carried out to test whether exposure to a virtual reality (VR) school environment elicits state anxiety and autonomic arousal in adolescents with school anxiety (diagnoses covering social anxiety disorder or specific phobia involving school contexts). In addition, we examined whether repeated VR exposure led to a reduction in this fear response, trait school anxiety, and social anxiety symptoms. Moreover, the relationship of presence, the subjective sense of "being there," during VR exposure with anxiety measures and treatment response was examined.

Methods: In a pilot study, 10 adolescents with school anxiety (age range 14 to 17 years) participated in five VRET sessions. Self-reported state anxiety, heart rate, and presence during exposure, as well as trait school anxiety and social anxiety before and after treatment, were measured.

Results: The VR scenario induced state anxiety and autonomic arousal. After VRET, a significant reduction in state anxiety (η^2=0.74) and social anxiety symptoms (d=0.82) as well as a trend toward a decrease in trait school anxiety were observed, while autonomic arousal did not change. In addition, presence during VR exposure was associated with state anxiety and treatment response.

Conclusions: Our findings indicate the feasibility and potential effectiveness of VRET as a treatment method for symptoms of school and social anxiety in adolescents.

(JMIR Ment Health 2024;11:e56235) doi:10.2196/56235

KEYWORDS

virtual reality exposure therapy; VRET; school anxiety; social anxiety; adolescents; virtual reality; VR; autonomic arousal; exposure therapy; posttreatment; digital health; simulation

Introduction

Anxiety disorders are the most common mental disorders in adolescence, with approximately 25% of adolescents meeting diagnostic criteria within 12 months [1]. Anxiety disorders frequently manifest in educational contexts [2-4]. School anxiety is a syndrome that summarizes fears of school-related situations such as fear of failure in performance situations, or fear of social situations [5,6]. It repeatedly leads to avoidance behavior including school refusal [7-9]. Prevalence for anxiety-based school refusal ranges from 1% to 4% of students over a 12-month period [7,10], although these figures must be interpreted cautiously due to differing conceptualizations and limited research [11]. School anxiety is not classified as a disorder in diagnostic manuals [12,13]. However, school anxiety and absenteeism are associated with an increased risk of poor academic performance and impaired social functioning [6,14-17].

School anxiety can occur within different anxiety disorders, such as specific phobias involving school contexts, and particularly in social anxiety disorder (SAD) [6,18]. These disorders overlap with school anxiety in terms of fears, symptoms, and consequences [5,11,19]. SAD entails anxiety in different social interaction and performance situations, most often including the school setting [20,21].

The most common intervention for school anxiety, SAD, and specific phobias is cognitive behavioral therapy (CBT), which has been demonstrated to promote attendance at school and reduction of anxiety symptoms [11,22-25]. An important component of CBT is exposure in vivo, which is associated with large effect sizes in the treatment of anxiety disorders [26,27]. However, despite its effectiveness, exposure in vivo is not consistently used for children and adolescents with anxiety disorders [28,29]. Reasons include lack of resources and inconvenient scheduling of sessions [30,31], especially in school settings. In addition, exposure to school-related stimuli requires a high level of cooperation with the school [32].

Virtual reality exposure therapy (VRET) may overcome some of these challenges. VRET involves exposure to feared objects and situations using interactive virtual environments via head-mounted displays and headphones, thus enabling greater control and an individual adaptation of the exposure situation [33]. VRET has been rated by therapists to be more practical in the treatment of SAD than exposure in vivo [34]. Moreover, youth who are unwilling to participate in exposure in vivo may be more likely to choose VRET because it could be more acceptable [35,36].

Several studies have provided evidence for the efficacy of VRET in adults with anxiety disorders, showing comparable effect sizes in symptom reduction as in vivo exposure [33,36]. The effects appear to generalize to daily life and are maintained [37-39]. Although specific phobias have been most frequently studied to date, a growing body of research also shows promising results for SAD, reporting reductions in state and trait anxiety [34,40-42]. However, evidence for the effectiveness of VRET in the treatment of children and adolescents is largely lacking [43].

Preliminary research suggests that VRET is acceptable to adolescents and is associated with reductions in anxiety symptoms in youth (eg, [43-46]). Kahlon et al [47] reported a reduction in public speaking anxiety in adolescents after a speech task in a virtual classroom which was maintained at follow-up. To date, only one study examined VRET for school anxiety [48], indicating reductions in school-related fears. However, this study recruited participants in schools without a formal clinical diagnosis and combined VRET with other methods such as relaxation training.

The principles underlying VRET are not yet fully understood. However, it is assumed that the working mechanisms of VRET mirror those of exposure in vivo [49]. Accordingly, previous accounts of VRET refer to the “emotional processing theory” [50] that proposes that repeated confrontation with anxiety-relevant stimuli gradually reduces emotional and physiological reactivity by altering the cognitive representation of these stimuli. The “inhibitory learning model” [51] emphasizes the relevance of expectancy violations, that is, the mismatch between expected and actual outcomes during exposure. It postulates that through repeated exposure to anxiety-relevant stimuli, patients learn that the feared consequences do not occur or are less severe than expected. However, as many feared outcomes cannot occur in VRET, and therefore, cannot be tested, the effectiveness of VRET can probably not be solely attributed to the violation of expectancies [49,52,53]. Therefore, it is suggested that multiple mechanisms may contribute to anxiety reduction in VRET [49]. The reduction of fear (habituation) and expectancy violations will often occur in an equivalent manner and only slightly influence the design of the actual exposure session [54].

One important factor regarding the mechanisms of VRET itself might be presence, which refers to the subjective interpretation of the virtual environment as if it were real, feeling engaged, and connected to it [55,56]. Presence is considered a precondition for evoking anxiety through virtual scenarios, as it indicates that the virtual environment feels realistic and attracts the focus of attention [57-59]. It may therefore also be relevant for therapeutic efficacy [55]. However, the findings regarding the relationship between presence and treatment outcome are mixed [47,55,59,60].

This study examined VRET to reduce school anxiety in adolescents with a diagnosis of SAD or specific phobia involving school contexts. We hypothesized that a virtual school scenario would elicit self-reported state anxiety and autonomic arousal (heart rate [HR]) in adolescents with school anxiety. Moreover, we expected state anxiety, autonomic arousal, and trait school, as well as social anxiety, would be reduced after VRET. Moreover, we assumed that there is a positive association between a sense of presence during the virtual reality (VR) sessions and anxiety measures, as well as treatment response.

Methods

Participants

A total of 13 adolescents aged 12 to 18 years were recruited from in and outpatient services from the Clinic for Psychiatry, Psychosomatics and Psychotherapy of Children and Adolescence, Charité Universitätsmedizin Berlin. Inclusion criteria were (1) school anxiety assessed by the current psychotherapist and (2) a diagnosis of SAD (F40.1) or specific phobia involving school contexts (F40.2) according to the *ICD-10* (*International Statistical Classification of Diseases, Tenth Revision*) [13]. Exclusion criteria were acute suicidality, motion sickness, and visual impairments.

Three participants did not complete the assessments due to a suicide attempt not related to the study (n=1), a delay caused by coronavirus disease, and one participant did not wish to complete the study. The sample characteristics (n=10) are depicted in Table 1. All participants had a diagnosis of SAD.

Table 1. Baseline demographic and clinical characteristics.

Characteristics	Value (N=10)
Age	
Mean (SD)	15.7 (0.9)
Minimum	14
Maximum	17
Class	
Mean (SD)	10
Minimum	9
Maximum	11
School anxiety[a]	
Test anxiety	
Mean (SD)	12.9 (2.4)
Minimum	15
Maximum	8
General anxiety	
Mean (SD)	12.5 (1.7)
Minimum	15
Maximum	9
School reluctance	
Mean (SD)	7.9 (1.6)
Minimum	10
Maximum	5
Social anxiety[b]	
Mean (SD)	17.7 (2.3)
Minimum	13
Maximum	21
Relative frequency, n (%)	
Sex (female)	6 (60)
School type (grammar school)	7 (70)
Diagnosis (SAD[c])	10 (100)
Comorbid disorders[d]	9 (90)

[a]School anxiety was measured with the Anxiety Questionnaire for Pupils.
[b]Social anxiety was measured with the Diagnostic System for Mental Disorders in Children and Adolescents in German (DISYPS-III-SBB-ANG).
[c]SAD: social anxiety disorder (F40.1).
[d]Comorbid disorders: F32.1/2, F40.2, F41.0, F42.2, F50.0/8, F81.0.

Ethical Considerations

The study was approved by the ethics committee of the Charité Universitätsmedizin (EA2/254/21). Both written and oral informed consent was given by participants and their caregivers. All participants received an expense allowance of 10 € (approximately US $11) per hour for their participation in the study.

Psychological Assessments

To examine school-related trait anxiety, the Angstfragebogen für Schüler (AFS; English: Anxiety Questionnaire for Pupils) [61] was used. This self-report questionnaire consists of 50 items, which are to be answered on a dichotomous scale (correct, not correct), that are categorized into four subscales: (1) test anxiety: fear of exam situations or performance failure (eg, “I am always afraid that I will get bad grades during exams”), (2) general (manifest) anxiety: general anxiety symptoms and reduced self-confidence (eg, “I worry too much”), (3) school reluctance: inner resistance and loss of motivation toward educational matters (eg, “I am often bad-tempered in class”), and (4) social desirability which assesses the response behavior of the participants (eg, “I have never lied”). The latter subscale was not included in this study, as it does not measure school anxiety and was therefore not considered relevant for the purpose of the study. The internal consistencies of the reported scales (α=.73 to α=.89) and the test-retest reliability after one month (rtt=0.71 to rtt=0.76) are high. Validity and sensitivity to change were supported by correlations with construct-related measures and various studies [61].

The anxiety disorders section of the Diagnostic System for Mental Disorders in Children and Adolescents in German (DISYPS-III-SBB-ANG) [62] was used for the assessment of social anxiety symptoms. Responses range from 0 (not at all) to 3 (particularly). The authors have reported a good internal consistency of the subscale (α=.87) but not test-retest reliability. The validity of the scale was supported by examinations of convergent and divergent validity [62].

In addition, the Subjective Units of Distress (SUD) [63] was used as a measure of subjective state anxiety. It was used as a verbal self-report during each phase of the scenario (“On a scale from 0 (relaxed or no fear) to 10 (biggest fear ever), how high is your fear right now?”). Correlations with autonomic stress parameters [64] and an established state anxiety questionnaire [65,66] support the validity of the SUD.

To assess the presence in virtual environments, the Igroup Presence Questionnaire (IPQ) [56] was administered. The IPQ is a self-report questionnaire that includes 14 items which are rated on a 7-point Likert scale ranging from 0 (not at all) to 6 (very much). It comprises three subscales (spatial presence, involvement, and experienced realism) and one general item about global presence. The IPQ has demonstrated good psychometric properties with a high internal consistency of the total score (α=.85 to α=.87) [56,67].

Autonomic Measures

Autonomic parameters were continuously assessed throughout the sessions using the eSense Pulse system (Mindfield Biosystems), consisting of a chest belt containing a 1-channel-electrocardiogram with 500 Hz sampling. After the measurement, the autonomic data were divided according to the different phases of the VR sessions. For inspecting and preprocessing the interbeat intervals data, the program ARTiiFACT [68] was used, with artifacts detected by absolute median deviation, manually checked, and replaced by cubic spline interpolation. Mean HR, influenced by the sympathetic and parasympathetic nervous system, was calculated for each phase of the first and the final session.

Virtual Reality Equipment

Participants wore a head-mounted display (VR glasses) and headphones (VIVE Pro). Participants further used two controllers for walking and jumping, as well as grabbing and carrying objects. The VR scenario was implemented using the VR simulation software Game Engine Unity 2020 (Unity Technologies).

Procedure

Participants completed 5 largely standardized exposure sessions within a one-month period, which were conducted by two graduate master-level students of clinical psychology in an office at the Department of Child and Adolescent Psychiatry, Charité (for study protocols, see Multimedia Appendix 1). Before the first and after the final session, participants completed the AFS and DISYPS. The first session began with psychoeducation about the rationale of exposure and the relevance of safety and avoidance behaviors. In the following sessions, the rationale was repeated and the adolescents’ negative expectations of potential outcomes in the exposure were verbally assessed. During exposure, safety and avoidance behaviors were prevented when distraction or avoidance were observed. The HR was continuously measured.

The VR scenario (Figure 1) consisted of the following phases which were controlled by the experimenter: (1) at a street in front of the school grounds, participants started to get used to the VR (“acclimatization”). (2) They could freely explore the schoolyard and school building (“exploration”). (3) Participants went to the classroom and sat at a table while typical school sounds were presented (“class”). (4) The ringing of a school bell announced the start of class and the participants were asked by the teacher to come to the front of the class and complete three tasks which were motivated by the Trier Social Stress Test [69,70]: (a) introducing themselves (“introduction”), (b) solve small arithmetic problems verbally (“math”), and (c) talk about a recent personal experience (“event”). Through prerecorded sentences, questions, and comments (in a neutral tone), the virtual teacher could interact with the participants. From the second session onwards, the virtual classmates responded to the participants by looking at them, laughing, or applauding. During each phase, the participants were asked to rate their SUD. From the phase “event” onwards, the assessment was repeated at regular intervals and a mean value was calculated afterward. Finally, the 3 tasks (“introduction,” “math,” and “event”) were repeated in a randomized order until self-reported state anxiety had decreased by about 40% [71]. Following these tasks, the participants sat down (“end”) until the end of the session.

Following each VR session, the experimenter discussed with the adolescents the occurrence of their negative expectations. Finally, participants completed the IPQ. Each session lasted about 60 minutes in total.

Figure 1. Virtual reality scenario. The upper pictures are from the experimenter's perspective, and the lower pictures are from the participants' perspective. Panels (A; schoolyard) and (B; school corridor) are examples of the phase "exploration." The lower pictures are related to the phases "introduction," "math," and "event.".

Statistical Analysis

Statistical analyses were performed using RStudio (version 2023.06.1; Posit PBC). SUD and HR as dependent variables within and between sessions were each analyzed using two-factor repeated measures ANOVAs (rmANOVA), with repeated measures factors session (first, final) and phase ("acclimatization," "exploration," "class," "introduction," "math," "event," "end"). Pre- and postreductions in trait school anxiety and social anxiety symptoms as dependent variables were analyzed using 1-tailed, paired-samples *t* tests of the AFS subscales and the DISYPS' social anxiety symptoms. Reductions in trait school anxiety were assessed for clinical significance with the reliable change index (RCI) [72], which was calculated using the SD and the test-retest reliability of the subscales from the normative sample of the AFS [61]. Clinical relevance was defined as the response criterion of a symptom reduction of 30% [73]. The 1-tailed Pearson product-moment correlations were calculated to examine the relationships between presence and state anxiety (IPQ total score and mean SUD and HR during the first session), as well as treatment response (mean IPQ total score across the five sessions and the post- and predifference scores of the AFS subscales).

Both uncorrected and corrected *P* values are reported. If the normal distribution of the data could not be confirmed with the Shapiro-Wilk test, nonparametric alternatives are reported. For rmANOVAs, Greenhouse-Geisser corrections were applied for variables that failed Mauchly's test of sphericity.

Results

State Anxiety and Autonomic Arousal

The rmANOVA for SUD (Table 2) resulted in a significant main effect of phase. Post hoc analyses (Multimedia Appendix 2) revealed that SUDs during the tasks inside the class, "class," "introduction," "math," and "event" were significantly higher than during "acclimatization" and "end" (*t* values 4.3; *P* values .001; $Ps_{corr.}$ .02. The main effect of the session and the session × phase interaction were also significant. Post hoc analyses showed that SUD was lower in all phases of the final session than the first session (*t* values 2.32; *P* values .046, $Ps_{corr.}$ .046). In addition, there was a steeper increase and decrease in SUD in the first session than in the final session (Figure 2A).

Table 2. Results of repeated measures ANOVAs for SUD[a] and autonomic arousal (HR[b]) in the first and the final session.

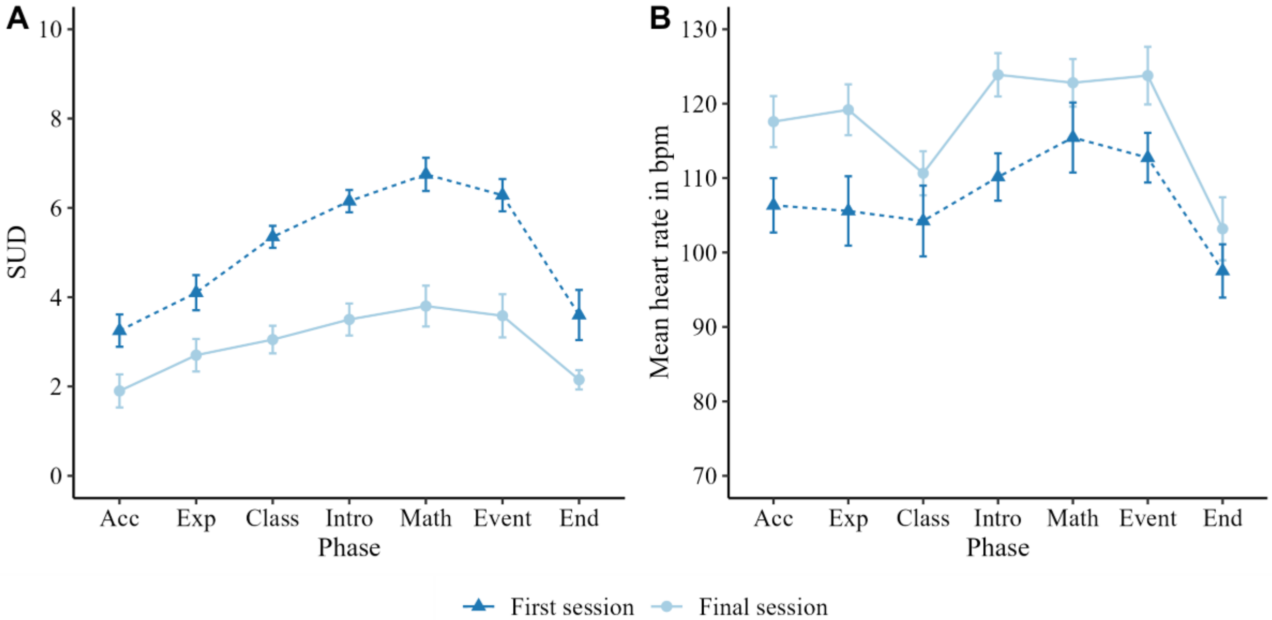
Variable and effect	*F* test (*df*)[c]	*P* value	Partial η^2
SUD (n=10)			
Phase	15.21 (6, 54)	<.001	0.63
Session	25.49 (1, 9)	<.001	0.74
Session × phase	4.56 (6, 54)	.006	0.34
HR (n=4)			
Phase	10.84 (6, 18)	.03	0.78
Session	3.00 (1, 3)	.18	0.50
Session × phase	1.71 (6, 18)	.25	0.36

[a]SUD: Subjective Units of Distress (state anxiety).

[b]HR: heart rate.

[c]All *df*s for phase and session × phase were Greenhouse-Geisser-adjusted.

Figure 2. Mean trajectory of SUD (A) and HR (B) in the first and the final session. Note that n=10 for SUD and n=4 for HR. Acc: acclimatization; Exp: exploration; HR: heart rate; Intro: introduction; SUD: Subjective Units of Distress (state anxiety); Means and standard errors.



Due to technical connection problems between the pulse sensor and the PC, complete autonomic measurements were only available from 4 participants. The rmANOVA of HR revealed a significant main effect of phase (Table 2 and Figure 2B). Post hoc analyses (Multimedia Appendix 2) showed that HR was significantly higher during "introduction" and "math" than during "acclimatization," "exploration," and "end" (*t* values 3.76; *P* values .007; Ps_{corr}. .09). Neither the main effect of the session nor the session × phase interaction was significant, however. The participants' individual trajectories corresponded with the reported effects (Multimedia Appendix 3).

Trait School Anxiety and Social Anxiety

For trait school anxiety (AFS), the analyses revealed a significant reduction from pre to post-VRET in the subscale school reluctance and a trend toward significance in test anxiety, but not in general anxiety (Table 3). For social anxiety (DISYPS), the 1-tailed paired samples *t* test revealed a significant reduction from pre to post VRET (Table 3).

Table 3. Reductions in trait school anxiety and social anxiety from pre to post-VRET[a].

Variable	Pre	Post	Test statistic[b]	*P* value	*P* corr.[c]	ES[d]
AFS[e]: trait school anxiety						
Test anxiety, mean (SD)	12.9 (2.4)	12.2 (2.8)	t_9=–1.56	.08	.15	*d*=–0.26
General anxiety, median (IQR)	12.5 (12.0-13.75)	12.0 (9.5-12.75)	V=4	.10	.15	*r*=0.49
School reluctance, mean (SD)	7.9 (1.6)	7.0 (2.6)	t_9=–1.87	.05	.14	*d*=–0.34
DISYPS[f]: trait social anxiety, mean (SD)	17.7 (2.3)	15.5 (3.0)	t_9=–2.66	.01	—[g]	*d*=–0.82

[a]VRET: virtual reality exposure therapy.

[b]Analyses were 1-tailed paired-samples *t* tests except for general anxiety, for which the Wilcoxon signed rank test was used.

[c]Bonferroni-Holm correction was applied for corrected *P* values.

[d]ES: effect size.

[e]AFS: Anxiety Questionnaire for Pupils.

[f]DISYPS: Diagnostic System for Mental Disorders According to ICD-10 (International Statistical Classification of Diseases, Tenth Revision) and DSM-5 (Diagnostic and Statistical Manual of Mental Disorders [Fifth Edition]) in Children and Adolescents.

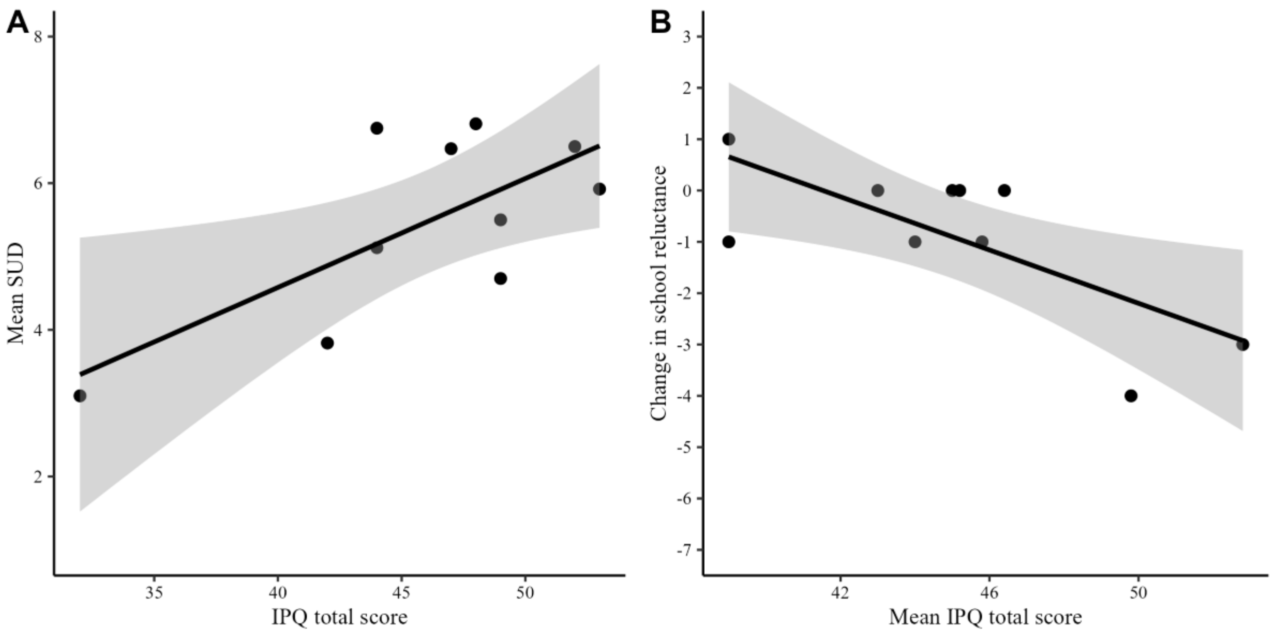
[g]Not applicable.

Regarding the clinical significance of the change in trait school anxiety, 7 of 10 participants did not show clinically significant improvements. One participant showed clinically significant improvements on the subscales general anxiety and school reluctance. Two participants showed a reduction >30% on at least one subscale, but no reliable change (RCI<|1.96|). The individual percentage change and the RCI of all participants can be found in Multimedia Appendix 4.

Presence and Anxiety

A significant correlation between the IPQ total score and mean SUD during the first session was observed (*r*=0.70; *P*=.01; 95% CI 0.24-1.00), suggesting that higher levels of presence were associated with higher levels of state anxiety (Figure 3A), but not with HR (*r*=–0.13; *P*=.58; 95% CI –0.86 to 1.00).

Figure 3. Association between presence and Subjective Units of Distress (state anxiety) in the first session and between mean presence across sessions and pre to post change in school reluctance (N=10). The graph represents the regression line with standard error. (B) Change was calculated as a Post-pre-difference , so that negative values represent a reduction from pre to post-VRET. IPQ: Igroup Presence Questionnaire; SUD: Subjective Units of Distress; VRET: virtual reality exposure therapy.



In addition, there was a negative correlation between the mean IPQ score across all five sessions and the post- and predifference scores in school reluctance (*r*=–0.73; *P*=.009; $P_{corr.}$=.03; 95% CI –1.00 to –0.29), indicating that higher levels of presence were associated with greater reductions in school reluctance (Figure 3B). However, no correlation was found between mean presence across sessions and the post- and predifferences in test anxiety (*r*=–0.19; *P*=.29; $P_{corr.}$=.42; 95% CI –1.00 to 0.40) and general anxiety (ρ=–0.29; *P*=.21; $P_{corr.}$=.42; 95% CI –0.78 to 0.42).

Discussion

Principal Findings

This study examined a novel VRET intervention to reduce school anxiety in adolescents. While there is consistent evidence for the efficacy of VRET for the treatment of anxiety disorders in adult populations [33,40,49], it is currently unclear whether VRET may be also effective in adolescents. The VR scenario elicited state anxiety consistently across participants, particularly in the first session with a large effect size. VRET was associated with a reduction in state anxiety and social anxiety symptoms, as well as a trend toward a decrease in trait school anxiety, suggesting the feasibility and potential effectiveness of VRET in the treatment of school and social anxiety symptoms in adolescents. Autonomic arousal did not change. Presence during VR exposure was associated with state anxiety and treatment response in a subscale of the AFS measuring school reluctance.

The strongest state anxiety ratings were reported during the tasks inside the classroom, suggesting that anxiety was not only triggered by the general laboratory and VR setting but specifically by anxiety-relevant stimuli and tasks within the scenario. This finding aligns with previous research demonstrating that virtual social-evaluative situations can evoke anxiety in both adults [40,74-77] and youth [44].

The effects of VRET on HR were less clear potentially due to the small sample size (n=4). However, HR was elevated inside the classroom, especially in the first session. In addition, although not statistically tested, throughout the whole exposure, HR was higher than in adolescents at rest reported by other studies [78,79] suggesting that the VR scenario also elicited autonomic arousal consistent with previous findings [47,74,76,77,80,81]. Autonomic arousal may have been less specifically linked to the anxiety-relevant tasks than self-reported anxiety, as youth with SAD are discussed to show blunted autonomic reactivity [79,82-84].

VRET was associated with a reduction in state anxiety across all participants (η^2=0.74), potentially caused by mechanisms of inhibitory learning and/or habituation in the specific situation. Moreover, we also observed a pre- to post reduction in social anxiety and a trend for a decrease in trait school anxiety. These findings are consistent with previous research on VRET for SAD and public speaking anxiety in adults [40,41,85,86], as well as with studies examining VRET in youth on public speaking and school anxiety [47,48]. These findings suggest that the VRET approach may be used to reduce state and potentially also for trait anxiety in adolescents.

However, despite large effects on state anxiety, the effect sizes and clinical response rates related to trait school anxiety in this study were below the typical outcomes of CBT and exposure therapy for SAD and school anxiety [23,42,87,88]. Moreover, most participants did not show reliable changes in trait school anxiety.

Several factors may account for these findings. First, exposure therapy for SAD and school refusal is mostly not examined as a stand-alone treatment, but rather in combination with other cognitive-behavioral methods such as cognitive restructuring [23,40,87,89]. In SAD and school anxiety, the cognitive component may play a greater role than in specific phobias, partly because of interpretation biases, indicating the need for cognitive interventions alongside exposure therapy [5,90,91]. Second, our VRET involved comparatively few sessions [87,89], which may be relevant since the number of treatment sessions may be a moderator of pre- to postanxiety reduction in CBT interventions [88]. Additionally, we included only one scenario, although multiple contexts and various stimuli are recommended to enhance therapeutic effects [51,89,92,93]. Furthermore, additional factors should be addressed, such as bullying and peer relationships, as these are considered potential contributors to the maintenance of school anxiety and school refusal [6,7,16]. Thus, incorporating social skills training might enhance treatment outcomes [88]. Finally, the majority of participants had comorbid depression, which may have negatively affected the treatment response [94,95].

VRET in our study was not associated with a reduction in autonomic arousal. These data are consistent with a previous study on CBT for children with SAD [82]. Accordingly, it is possible that autonomic changes require longer time scales to occur following psychological interventions [82,96], especially in anxiety disorders that are more complex than specific phobias. Moreover, the task demands of the VR scenario may have contributed to the absence of HR reductions [82] as social-evaluative tasks and cognitive load can elicit autonomic arousal even in healthy individuals [97-99].

The analyses of presence yielded a strong positive correlation with state anxiety in the first session. While this finding is consistent with previous research, the causality and direction of this effect, currently discussed to be bidirectional, should be further explored [55,59,80,100]. In addition, our results indicate a relationship between presence and treatment response, in particular school reluctance, aligning with prior research on VRET for public speaking anxiety [55,60]. It should be noted, however, that no correlation was found with the other subscales of trait school anxiety. Given the heterogeneous literature [47,59], this emphasizes the need for further research on presence and its relevance to VRET's therapeutic success.

Limitations

The sample size of this pilot study was small (n=10). In addition, the results derive from uncorrected *P* values and should be interpreted with caution. Moreover, cognitive symptoms and autonomic responses in social-evaluative situations can also depend on factors such as age and type of sample [98,101]. Hence, the findings may not generalize to other groups with school anxiety, especially children and individuals without SAD. Moreover, the pilot study did not include a control or comparison group. Therefore, we cannot separate the effects of VRET from other factors, such as spontaneous remission, the concurrent therapy also targeting SAD, school visits, and nonspecific therapeutic factors due to contact with the experimenter.

Another limitation pertains to the presence of the experimenter during VR exposure, which could have contributed to the anxiety response. In addition, VRET was not conducted by trained therapists, which might have reduced its effectiveness.

Nonetheless, previous studies also employing nonexpert therapists have observed reductions in anxiety as well [46,102] with lay therapists potentially being a highly sought-after resource to bridge the current gap between demand and mental health care. Furthermore, while the ability to move may have increased the presence during the VR exposure, it limits the autonomic results, as HR can be influenced by movement and position changes [103].

Conclusions

In summary, this pilot study offers preliminary evidence regarding the feasibility and potential effectiveness of a newly developed VRET for adolescents with school anxiety related to SAD, suggesting that VRET may constitute an effective treatment component for symptoms of anxiety in adolescence. The VR scenario successfully elicited self-reported state anxiety and autonomic arousal. Importantly, reductions in self-reported state anxiety and trait social anxiety, as well as a trend for a decrease in trait school anxiety were observed after five exposure sessions. Incorporation of a greater number of virtual scenarios and cognitive elements into VRET may further improve treatment outcomes. Future studies and RCTs should address the effectiveness of VRET for adolescents with school anxiety in large samples that include a waiting list group, active psychological placebo, or exposure in vivo to better examine and compare the efficacy of the VRET approach.

Conflicts of Interest

CU Correll has been a consultant and/or advisor to or has received honoraria from: AbbVie, Acadia, Adock Ingram, Alkermes, Allergan, Angelini, Aristo, Biogen, Boehringer-Ingelheim, Bristol-Meyers Squibb, Cardio Diagnostics, Cerevel, CNX Therapeutics, Compass Pathways, Darnitsa, Delpor, Denovo, Eli Lilly, Gedeon Richter, Hikma, Holmusk, IntraCellular Therapies, Jamjoom Pharma, Janssen/J&J, Karuna, LB Pharma, Lundbeck, MedInCell, MedLink, Merck, Mindpax, Mitsubishi Tanabe Pharma, Maplight, Mylan, Neumora Therapeutics, Neurocrine, Neurelis, Newron, Noven, Novo Nordisk, Otsuka, PPD Biotech, Recordati, Relmada, Reviva, Rovi, Sage, Saladax, Sanofi, Seqirus, SK Life Science, Sumitomo Pharma America, Sunovion, Sun Pharma, Supernus, Tabuk, Takeda, Teva, Terran, Tolmar, Vertex, Viatris and Xenon Pharmaceuticals. He provided expert testimony for Janssen, Lundbeck and Otsuka. He served on a Data Safety Monitoring Board for Compass Pathways, Denovo, IntraCellular Therapies, Lundbeck, Relmada, Reviva, Rovi, Supernus, and Teva. He has received grant support from Boehringer-Ingelheim, Janssen and Takeda. He received royalties from UpToDate and is also a stock option holder of Cardio Diagnostics, Kuleon Biosciences, LB Pharma, Medlink, Mindpax, Quantic, Terran.

Multimedia Appendix 1
Study protocols.
[DOCX File , 27 KB - mental_v11i1e56235_app1.docx]

Multimedia Appendix 2
Results of relevant post hoc pairwise comparisons of state anxiety and autonomic parameters.
[DOCX File , 30 KB - mental_v11i1e56235_app2.docx]

Multimedia Appendix 3
Individual trajectories of state anxiety and autonomic parameters.
[DOCX File , 259 KB - mental_v11i1e56235_app3.docx]

Multimedia Appendix 4
Percentage change and RCI (reliable change index) in trait school anxiety and social anxiety from pre to post VRET (virtual reality exposure therapy).
[DOCX File , 20 KB - mental_v11i1e56235_app4.docx]

References

1. Kessler RC, Avenevoli S, Costello EJ, Georgiades K, Green JG, Gruber MJ, et al. Prevalence, persistence, and sociodemographic correlates of DSM-IV disorders in the National Comorbidity Survey Replication Adolescent Supplement. Arch Gen Psychiatry 2012;69(4):372-380 [FREE Full text] [doi: 10.1001/archgenpsychiatry.2011.160] [Medline: 22147808]
2. Hilwerling L, Linnenbank M, Flockert H, Wedemeyer M, Warnking G, Blanke J, et al. School absenteeism—requirements for a successful therapeutic approach. Klinische Padiatrie 2020;232(5):265-268. [doi: 10.1055/a-1198-5684] [Medline: 32702763]
3. Richards HJ, Hadwin JA. An exploration of the relationship between trait anxiety and school attendance in young people. School Mental Health 2011;3(4):236-244. [doi: 10.1007/s12310-011-9054-9]
4. Van Ameringen M, Mancini C, Farvolden P. The impact of anxiety disorders on educational achievement. Journal of anxiety disorders 2003;17(5):561-571. [doi: 10.1016/s0887-6185(02)00228-1] [Medline: 12941366]

5. Melfsen S, Walitza S. Soziale Ängste und Schulangst: Entwicklungsrisiken Erkennen und Behandeln. Geneva: Beltz Verlag; 2013.
6. Petermann F, Petermann U. Schulangst. Monatsschrift Kinderheilkunde 2010;158(4):391-401. [doi: 10.1007/s00112-010-2180-x]
7. Egger HL, Costello JE, Angold A. School refusal and psychiatric disorders: a community study. J Am Acad Child Adolesc Psychiatry 2003;42(7):797-807. [doi: 10.1097/01.CHI.0000046865.56865.79] [Medline: 12819439]
8. Gonzálvez C, Díaz-Herrero Á, Sanmartín R, Vicent M, Fernández-Sogorb A, García-Fernández JM. Testing the functional profiles of school refusal behavior and clarifying their relationship with school anxiety. Front Public Health 2020;8:598915 [FREE Full text] [doi: 10.3389/fpubh.2020.598915] [Medline: 33344402]
9. Ingul JM, Nordahl HM. Anxiety as a risk factor for school absenteeism: what differentiates anxious school attenders from non-attenders? Ann Gener Psychiatry 2013;12(1):25 [FREE Full text] [doi: 10.1186/1744-859X-12-25] [Medline: 23886245]
10. Havik T, Bru E, Ertesvåg SK. Assessing reasons for school non-attendance. Scand J Educ Res 2015;59(3):316-336. [doi: 10.1080/00313831.2014.904424]
11. Elliott JG, Place M. Practitioner review: school refusal: developments in conceptualisation and treatment since 2000. J Child Psychol Psychiatry 2019;60(1):4-15. [doi: 10.1111/jcpp.12848] [Medline: 29197106]
12. American Psychiatric Association (APA). Diagnostic and Statistical Manual of Mental Disorders: DSM-5. United States: American Psychiatric Publishing; 2013.
13. World Health Organization (WHO). Internationale Klassifikationen Psychischer Störungen. United Kingdom: Hogrefe; 2016.
14. Epstein S, Roberts E, Sedgwick R, Polling C, Finning K, Ford T, et al. School absenteeism as a risk factor for self-harm and suicidal ideation in children and adolescents: a systematic review and meta-analysis. Eur Child Adolesc Psychiatry 2020;29(9):1175-1194 [FREE Full text] [doi: 10.1007/s00787-019-01327-3] [Medline: 30989389]
15. Filippello P, Buzzai C, Costa S, Sorrenti L. School refusal and absenteeism: perception of teacher behaviors, psychological basic needs, and academic achievement. Front Psychol 2019;10:1471 [FREE Full text] [doi: 10.3389/fpsyg.2019.01471] [Medline: 31316431]
16. Kearney CA. School absenteeism and school refusal behavior in youth: a contemporary review. Clin Psychol Rev 2008;28(3):451-471. [doi: 10.1016/j.cpr.2007.07.012] [Medline: 17720288]
17. Smerillo NE, Reynolds AJ, Temple JA, Ou SR. Chronic absence, eighth-grade achievement, and high school attainment in the Chicago longitudinal study. J School Psychol 2018;67:163-178 [FREE Full text] [doi: 10.1016/j.jsp.2017.11.001] [Medline: 29571532]
18. Büch H, Döpfner M. Soziale Ängste: Therapieprogramm Für Kinder Und Jugendliche Mit Angst- Und Zwangsstörungen (THAZ). United Kingdom: Hogrefe; 2012.
19. Hugon M, Delgado B, Ingles CJ, Hidalgo MD, Garcia-Fernandez JM, Martinez-Monteagudo MC. Reliability and validity evidence of scores on the French version of the questionnaire about interpersonal difficulties for adolescents. Psychologica Belgica 2015;55(3):159-174 [FREE Full text] [doi: 10.5334/pb.bl] [Medline: 30479422]
20. Beidel DC, Turner SM, Young BJ, Ammerman RT, Sallee FR, Crosby LE. Psychopathology of adolescent social phobia. J Psychopathol Behav Assess 2007;29(1):46-53 [FREE Full text] [doi: 10.1007/s10862-006-9021-1]
21. Weeks M, Coplan RJ, Kingsbury A. The correlates and consequences of early appearing social anxiety in young children. J Anxiety Disord 2009;23(7):965-972. [doi: 10.1016/j.janxdis.2009.06.006] [Medline: 19596545]
22. Heyne D, Sauter FM, Van Widenfelt BM, Vermeiren R, Westenberg PM. School refusal and anxiety in adolescence: non-randomized trial of a developmentally sensitive cognitive behavioral therapy. J Anxiety Disord 2011;25(7):870-878. [doi: 10.1016/j.janxdis.2011.04.006] [Medline: 21602027]
23. Walter D, Hautmann C, Rizk S, Lehmkuhl G, Doepfner M. Short- and long-term effects of inpatient cognitive-behavioral treatment of adolescents with anxious-depressed school absenteeism: a within-subject comparison of changes. Child Fam Behav Ther 2014;36(3):171-190. [doi: 10.1080/07317107.2014.934173] [Medline: 21602027]
24. Asbrand J, Heinrichs N, Schmidtendorf S, Nitschke K, Tuschen-Caffier B. Experience versus report: where are changes seen after exposure-based cognitive-behavioral therapy? A randomized controlled group treatment of childhood social anxiety disorder. Child Psychiatry Hum Dev 2020;51(3):427-441 [FREE Full text] [doi: 10.1007/s10578-019-00954-w] [Medline: 31960175]
25. Wang Z, Whiteside SPH, Sim L, Farah W, Morrow AS, Alsawas M, et al. Comparative effectiveness and safety of cognitive behavioral therapy and pharmacotherapy for childhood anxiety disorders: a systematic review and meta-analysis. JAMA Pediatr 2017;171(11):1049-1056 [FREE Full text] [doi: 10.1001/jamapediatrics.2017.3036] [Medline: 28859190]
26. Higa-McMillan CK, Francis SE, Rith-Najarian L, Chorpita BF. Evidence base update: 50 years of research on treatment for child and adolescent anxiety. J Clin Child Adolesc Psychol 2016;45(2):91-113. [doi: 10.1080/15374416.2015.1046177] [Medline: 26087438]
27. Whiteside SPH, Sim LA, Morrow AS, Farah WH, Hilliker DR, Murad MH, et al. A meta-analysis to guide the enhancement of CBT for childhood anxiety: exposure over anxiety management. Clin Child Fam Psychol Rev 2020;23(1):102-121 [FREE Full text] [doi: 10.1007/s10567-019-00303-2] [Medline: 31628568]

28. Higa-McMillan C, Kotte A, Jackson D, Daleiden EL. Overlapping and non-overlapping practices in usual and evidence-based care for youth anxiety. J Behav Health Serv Res 2017;44(4):684-694. [doi: 10.1007/s11414-016-9502-2] [Medline: 26945583]
29. Whiteside SPH, Deacon BJ, Benito K, Stewart E. Factors associated with practitioners' use of exposure therapy for childhood anxiety disorders. J Anxiety Disord 2016;40:29-36 [FREE Full text] [doi: 10.1016/j.janxdis.2016.04.001] [Medline: 27085463]
30. Pittig A, Kotter R, Hoyer J. The struggle of behavioral therapists with exposure: self-reported practicability, negative beliefs, and therapist distress about exposure-based interventions. Behav Ther 2019;50(2):353-366. [doi: 10.1016/j.beth.2018.07.003] [Medline: 30824251]
31. Reid AM, Bolshakova MI, Guzick AG, Fernandez AG, Striley CW, Geffken GR, et al. Common barriers to the dissemination of exposure therapy for youth with anxiety disorders. Community Ment Health J 2017;53(4):432-437. [doi: 10.1007/s10597-017-0108-9] [Medline: 28181093]
32. Maeda N, Hatada S, Sonoda J, Takayama I. School-based intensive exposure therapy for school refusal behavior. Clin Case Stud 2012;11(4):299-311. [doi: 10.1177/1534650112457456]
33. Freitas JRS, Velosa VHS, Abreu LTN, Jardim RL, Santos JAV, Peres B, et al. Virtual reality exposure treatment in phobias: a systematic review. Psychiatr Q 2021;92(4):1685-1710. [doi: 10.1007/s11126-021-09935-6] [Medline: 34173160]
34. Bouchard S, Dumoulin S, Robillard G, Guitard T, Klinger E, Forget H, et al. Virtual reality compared with exposure in the treatment of social anxiety disorder: a three-arm randomised controlled trial. Br J Psychiatry 2017;210(4):276-283. [doi: 10.1192/bjp.bp.116.184234] [Medline: 27979818]
35. Garcia-Palacios A, Botella C, Hoffman H, Fabregat S. Comparing acceptance and refusal rates of virtual reality exposure vs. in vivo exposure by patients with specific phobias. CyberPsychol Behav 2007;10(5):722-724. [doi: 10.1089/cpb.2007.9962] [Medline: 17927544]
36. Carl E, Stein AT, Levihn-Coon A, Pogue JR, Rothbaum B, Emmelkamp P, et al. Virtual reality exposure therapy for anxiety and related disorders: a meta-analysis of randomized controlled trials. J Anxiety Disord 2019;61:27-36. [doi: 10.1016/j.janxdis.2018.08.003] [Medline: 30287083]
37. Morina N, Ijntema H, Meyerbröker K, Emmelkamp PMG. Can virtual reality exposure therapy gains be generalized to real-life? A meta-analysis of studies applying behavioral assessments. Behav Res Ther 2015;74:18-24. [doi: 10.1016/j.brat.2015.08.010] [Medline: 26355646]
38. Rothbaum BO, Anderson P, Zimand E, Hodges L, Lang D, Wilson J. Virtual reality exposure therapy and standard (in vivo) exposure therapy in the treatment of fear of flying. Behav Ther 2006;37(1):80-90. [doi: 10.1016/j.beth.2005.04.004] [Medline: 16942963]
39. Opriş D, Pintea S, García-Palacios A, Botella C, Szamosközi Ş, David D. Virtual reality exposure therapy in anxiety disorders: a quantitative meta-analysis. Depress Anxiety 2012;29(2):85-93. [doi: 10.1002/da.20910] [Medline: 22065564]
40. Emmelkamp PMG, Meyerbröker K, Morina N. Virtual reality therapy in social anxiety disorder. Curr Psychiatry Rep 2020;22(7):32 [FREE Full text] [doi: 10.1007/s11920-020-01156-1] [Medline: 32405657]
41. Lindner P, Miloff A, Fagernäs S, Andersen J, Sigeman M, Andersson G, et al. Therapist-led and self-led one-session virtual reality exposure therapy for public speaking anxiety with consumer hardware and software: a randomized controlled trial. J Anxiety Disord 2019;61:45-54. [doi: 10.1016/j.janxdis.2018.07.003] [Medline: 30054173]
42. Morina N, Kampmann I, Emmelkamp P, Barbui C, Hoppen TH. Meta-analysis of virtual reality exposure therapy for social anxiety disorder. Psychol Med 2023;53(5):2176-2178 [FREE Full text] [doi: 10.1017/S0033291721001690] [Medline: 34001293]
43. Kothgassner OD, Felnhofer A. Lack of research on efficacy of virtual reality exposure therapy (VRET) for anxiety disorders in children and adolescents : a systematic review. Neuropsychiatr 2021;35(2):68-75 [FREE Full text] [doi: 10.1007/s40211-020-00349-7] [Medline: 32372291]
44. Parrish DE, Oxhandler HK, Duron JF, Swank P, Bordnick P. Feasibility of virtual reality environments for adolescent social anxiety disorder. Res Soc Work Pract 2016;26(7):825-835. [doi: 10.1177/1049731514568897]
45. Sarver NW, Beidel DC, Spitalnick JS. The feasibility and acceptability of virtual environments in the treatment of childhood social anxiety disorder. J Clin Child Adolesc Psychol 2014;43(1):63-73 [FREE Full text] [doi: 10.1080/15374416.2013.843461] [Medline: 24144182]
46. Servera M, Sáez B, Gelabert Mir JM. Feasibility of a virtual reality program to treat children with fear of darkness with nonexpert therapists. RPCNA 2020;7(2):16-21. [doi: 10.21134/rpcna.2020.07.2.2]
47. Kahlon S, Lindner P, Nordgreen T. Virtual reality exposure therapy for adolescents with fear of public speaking: a non-randomized feasibility and pilot study. Child Adolesc Psychiatry Ment Health 2019;13(1):47 [FREE Full text] [doi: 10.1186/s13034-019-0307-y] [Medline: 31890004]
48. Gutiérrez-Maldonado J, Magallón-Neri E, Rus-Calafell M, Peñaloza-Salazar C. Virtual reality exposure therapy for school phobia. Anu Psicol 2009;40(2):223-236 [FREE Full text]
49. Meyerbröker K, Morina N. The use of virtual reality in assessment and treatment of anxiety and related disorders. Clin Psychol Psychother 2021;28(3):466-476 [FREE Full text] [doi: 10.1002/cpp.2623] [Medline: 34097318]
50. Foa EB, Kozak MJ. Emotional processing of fear: exposure to corrective information. Psychol Bull 1986;99(1):20-35 [FREE Full text] [doi: 10.1037//0033-2909.99.1.20]

51. Craske MG, Treanor M, Conway CC, Zbozinek T, Vervliet B. Maximizing exposure therapy: an inhibitory learning approach. Behav Res Ther 2014;58:10-23 [FREE Full text] [doi: 10.1016/j.brat.2014.04.006] [Medline: 24864005]
52. Roesmann K, Leehr EJ, Böhnlein J, Gathmann B, Herrmann MJ, Junghöfer M, et al. Mechanisms of action underlying virtual reality exposure treatment in spider phobia: pivotal role of within-session fear reduction. J Anxiety Disord 2023;100:102790. [doi: 10.1016/j.janxdis.2023.102790] [Medline: 37879242]
53. Scheveneels S, Boddez Y, Van Daele T, Hermans D. Virtually unexpected: no role for expectancy violation in virtual reality exposure for public speaking anxiety. Front Psychol 2019;10:2849 [FREE Full text] [doi: 10.3389/fpsyg.2019.02849] [Medline: 31920878]
54. Pittig A, Stevens S, Vervliet B, Treanor M, Conway CC, Zbozinek T, et al. Optimierung expositionsbasierter therapie. Psychotherapeut (Berl) 2015;60(5):401-418. [doi: 10.1007/s00278-015-0042-7]
55. Price M, Mehta N, Tone EB, Anderson PL. Does engagement with exposure yield better outcomes? Components of presence as a predictor of treatment response for virtual reality exposure therapy for social phobia. J Anxiety Disord 2011;25(6):763-770 [FREE Full text] [doi: 10.1016/j.janxdis.2011.03.004] [Medline: 21515027]
56. Schubert T, Friedmann F, Regenbrecht H. The experience of presence: factor analytic insights. Presence: Teleoperators Virtual Environ 2001;10(3):266-281. [doi: 10.1162/105474601300343603]
57. Diemer J, Alpers GW, Peperkorn HM, Shiban Y, Mühlberger A. The impact of perception and presence on emotional reactions: a review of research in virtual reality. Front Psychol 2015;6:26 [FREE Full text] [doi: 10.3389/fpsyg.2015.00026] [Medline: 25688218]
58. Ling Y, Nefs HT, Morina N, Heynderickx I, Brinkman WP. A meta-analysis on the relationship between self-reported presence and anxiety in virtual reality exposure therapy for anxiety disorders. PLoS One 2014;9(5):e96144 [FREE Full text] [doi: 10.1371/journal.pone.0096144] [Medline: 24801324]
59. Price M, Anderson P. The role of presence in virtual reality exposure therapy. J Anxiety Disord 2007;21(5):742-751 [FREE Full text] [doi: 10.1016/j.janxdis.2006.11.002] [Medline: 17145164]
60. Reeves R, Elliott A, Curran D, Dyer K, Hanna D. 360° video virtual reality exposure therapy for public speaking anxiety: a randomized controlled trial. J Anxiety Disord 2021;83:102451. [doi: 10.1016/j.janxdis.2021.102451] [Medline: 34304028]
61. Wieczerkowski W, Nickel H, Janowski A, Fittkau B, Rauer W, Petermann F. Angstfragebogen Für Schüler. Germany: Hogrefe; 2016.
62. Döpfner M, Görtz-Dorten A. Diagnostik-System Für Psychische Störungen Nach ICD-10 Und DSM-5 Für Kinder Und Jugendliche - III. Germany: Hogrefe; 2017.
63. Wolpe J. The Practice of Behavior Therapy. United Kingdom: Pergamon Press; 1990.
64. Thyer BA, Papsdorf JD, Davis R, Vallecorsa S. Autonomic correlates of the subjective anxiety scale. J Behav Ther Exp Psychiatry 1984;15(1):3-7 [FREE Full text] [doi: 10.1016/0005-7916(84)90115-0] [Medline: 6470155]
65. Kaplan DM, Smith T, Coons JA. A validity study of the subjective unit of discomfort (SUD) score. Meas Eval Couns Dev 1995;27(4):195-199 [FREE Full text]
66. Kim D, Bae H, Chon Park Y. Validity of the subjective units of disturbance scale in EMDR. J EMDR Prac Res 2008;2(1):57-62. [doi: 10.1891/1933-3196.2.1.57]
67. Igroup presence questionnaire (IPQ) factor analysis. Igroup Project Consortium. URL: http://www.igroup.org/pq/ipq/factor.php [accessed 2024-08-27]
68. Kaufmann T, Sütterlin S, Schulz SM, Vögele C. ARTiiFACT: a tool for heart rate artifact processing and heart rate variability analysis. Behav Res Methods 2011;43(4):1161-1170. [doi: 10.3758/s13428-011-0107-7] [Medline: 21573720]
69. Buske-Kirschbaum A, Jobst S, Wustmans A, Kirschbaum C, Rauh W, Hellhammer D. Attenuated free cortisol response to psychosocial stress in children with atopic dermatitis. Psychosom Med 1997;59(4):419-426. [doi: 10.1097/00006842-199707000-00012] [Medline: 9251162]
70. Zimmer P, Buttlar B, Halbeisen G, Walther E, Domes G. Virtually stressed? A refined virtual reality adaptation of the trier social stress test (TSST) induces robust endocrine responses. Psychoneuroendocrinology 2019;101:186-192. [doi: 10.1016/j.psyneuen.2018.11.010] [Medline: 30469086]
71. Hagena S, Gebauer M. Therapie-Tools Angststörungen. Germany: Beltz; 2014.
72. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. In: Kazdin AE, editor. Methodological Issues & Strategies in Clinical Research. New York, US: American Psychological Association; 1992:631-642.
73. Leehr EJ, Roesmann K, Böhnlein J, Dannlowski U, Gathmann B, Herrmann MJ, et al. Clinical predictors of treatment response towards exposure therapy in virtuo in spider phobia: a machine learning and external cross-validation approach. J Anxiety Disord 2021;83:102448. [doi: 10.1016/j.janxdis.2021.102448] [Medline: 34298236]
74. Hartanto D, Kampmann IL, Morina N, Emmelkamp PGM, Neerincx MA, Brinkman WP. Controlling social stress in virtual reality environments. PLoS One 2014;9(3):e92804 [FREE Full text] [doi: 10.1371/journal.pone.0092804] [Medline: 24671006]
75. Kishimoto T, Ding X. The influences of virtual social feedback on social anxiety disorders. Behav Cogn Psychother 2019;47(6):726-735. [doi: 10.1017/S1352465819000377] [Medline: 31106720]

76. Owens ME, Beidel DC. Can virtual reality effectively elicit distress associated with social anxiety disorder? J Psychopathol Behav Assess 2015;37(2):296-305. [doi: 10.1007/s10862-014-9454-x]
77. Shiban Y, Diemer J, Brandl S, Zack R, Mühlberger A, Wüst S. Trier social stress test in vivo and in virtual reality: dissociation of response domains. Int J Psychophysiol 2016;110:47-55. [doi: 10.1016/j.ijpsycho.2016.10.008] [Medline: 27742258]
78. Rabbia F, Grosso T, Genova GC, Conterno A, De Vito B, Mulatero P, et al. Assessing resting heart rate in adolescents: determinants and correlates. J Hum Hypertens 2002;16(5):327-332. [doi: 10.1038/sj.jhh.1001398] [Medline: 12082493]
79. Sharma VK, Subramanian SK, Arunachalam V, Rajendran R. Heart rate variability in adolescents—normative data stratified by sex and physical activity. J Clin Diagn Res 2015;9(10):CC08-CC13 [FREE Full text] [doi: 10.7860/JCDR/2015/15373.6662] [Medline: 26557514]
80. Diemer J, Mühlberger A, Pauli P, Zwanzger P. Virtual reality exposure in anxiety disorders: impact on psychophysiological reactivity. World J Biol Psychiatry 2014;15(6):427-442. [doi: 10.3109/15622975.2014.892632] [Medline: 24666248]
81. Felnhofer A, Hlavacs H, Beutl L, Kryspin-Exner I, Kothgassner OD. Physical presence, social presence, and anxiety in participants with social anxiety disorder during virtual cue exposure. Cyberpsychol Behav Soc Netw 2019;22(1):46-50. [doi: 10.1089/cyber.2018.0221] [Medline: 30407091]
82. Asbrand J, Vögele C, Heinrichs N, Nitschke K, Tuschen-Caffier B. Autonomic dysregulation in child social anxiety disorder: an experimental design using CBT treatment. Appl Psychophysiol Biofeedback 2022;47(3):199-212 [FREE Full text] [doi: 10.1007/s10484-022-09548-0] [Medline: 35641719]
83. Krämer M, Seefeldt WL, Heinrichs N, Tuschen-Caffier B, Schmitz J, Wolf OT, et al. Subjective, autonomic, and endocrine reactivity during social stress in children with social phobia. J Abnorm Child Psychol 2012;40(1):95-104. [doi: 10.1007/s10802-011-9548-9] [Medline: 21800016]
84. Siess J, Blechert J, Schmitz J. Psychophysiological arousal and biased perception of bodily anxiety symptoms in socially anxious children and adolescents: a systematic review. Eur Child Adolesc Psychiatry 2014;23(3):127-142. [doi: 10.1007/s00787-013-0443-5] [Medline: 23812865]
85. Horigome T, Kurokawa S, Sawada K, Kudo S, Shiga K, Mimura M, et al. Virtual reality exposure therapy for social anxiety disorder: a systematic review and meta-analysis. Psychol Med 2020;50(15):2487-2497. [doi: 10.1017/S0033291720003785] [Medline: 33070784]
86. Premkumar P, Heym N, Brown DJ, Battersby S, Sumich A, Huntington B, et al. The effectiveness of self-guided virtual-reality exposure therapy for public-speaking anxiety. Front Psychiatry 2021;12:694610 [FREE Full text] [doi: 10.3389/fpsyt.2021.694610] [Medline: 34489755]
87. In-Albon T, Schneider S. Psychotherapy of childhood anxiety disorders: a meta-analysis. Psychother Psychosom 2007;76(1):15-24. [doi: 10.1159/000096361] [Medline: 17170560]
88. Scaini S, Belotti R, Ogliari A, Battaglia M. A comprehensive meta-analysis of cognitive-behavioral interventions for social anxiety disorder in children and adolescents. J Anxiety Disord 2016;42:105-112. [doi: 10.1016/j.janxdis.2016.05.008] [Medline: 27399932]
89. Wechsler TF, Kümpers F, Mühlberger A. Inferiority or even superiority of virtual reality exposure therapy in phobias?—A systematic review and quantitative meta-analysis on randomized controlled trials specifically comparing the efficacy of virtual reality exposure to gold standard exposure in agoraphobia, specific phobia, and social phobia. Front Psychol 2019;10:1758 [FREE Full text] [doi: 10.3389/fpsyg.2019.01758] [Medline: 31551840]
90. Peris TS, Compton SN, Kendall PC, Birmaher B, Sherrill J, March J, et al. Trajectories of change in youth anxiety during cognitive-behavior therapy. J Consult Clin Psychol 2015;83(2):239-252 [FREE Full text] [doi: 10.1037/a0038402] [Medline: 25486372]
91. Leigh E, Clark DM. Understanding social anxiety disorder in adolescents and improving treatment outcomes: applying the cognitive model of clark and wells (1995). Clin Child Fam Psychol Rev 2018;21(3):388-414 [FREE Full text] [doi: 10.1007/s10567-018-0258-5] [Medline: 29654442]
92. Traub J, In-Albon T. Therapie-tools angststörungen im kindes- und jugendalter. In: Therapietools. Germany: Beltz; 2017.
93. Shiban Y, Pauli P, Mühlberger A. Effect of multiple context exposure on renewal in spider phobia. Behav Res Ther 2013;51(2):68-74. [doi: 10.1016/j.brat.2012.10.007] [Medline: 23261707]
94. Ginsburg GS, Kendall PC, Sakolsky D, Compton SN, Piacentini J, Albano AM, et al. Remission after acute treatment in children and adolescents with anxiety disorders: findings from the CAMS. J Consult Clin Psychol 2011;79(6):806-813 [FREE Full text] [doi: 10.1037/a0025933] [Medline: 22122292]
95. Lundkvist-Houndoumadi I, Thastum M. Anxious children and adolescents non-responding to CBT: clinical predictors and families' experiences of therapy. Clin Psychol Psychother 2017;24(1):82-93. [doi: 10.1002/cpp.1982] [Medline: 26514088]
96. Mumm JLM, Pyrkosch L, Plag J, Nagel P, Petzold MB, Bischoff S, et al. Heart rate variability in patients with agoraphobia with or without panic disorder remains stable during CBT but increases following in-vivo exposure. J Anxiety Disord 2019;64:16-23. [doi: 10.1016/j.janxdis.2019.03.001] [Medline: 30875662]
97. Bernardi L, Wdowczyk-Szulc J, Valenti C, Castoldi S, Passino C, Spadacini G, et al. Effects of controlled breathing, mental activity and mental stress with or without verbalization on heart rate variability. J Am Coll Cardiol 2000;35(6):1462-1469 [FREE Full text] [doi: 10.1016/s0735-1097(00)00595-7] [Medline: 10807448]

98. Seddon JA, Rodriguez VJ, Provencher Y, Raftery-Helmer J, Hersh J, Labelle PR, et al. Meta-analysis of the effectiveness of the trier social stress test in eliciting physiological stress responses in children and adolescents. Psychoneuroendocrinology 2020;116:104582. [doi: 10.1016/j.psyneuen.2020.104582] [Medline: 32305745]
99. Solhjoo S, Haigney MC, McBee E, van Merrienboer JJG, Schuwirth L, Artino AR, et al. Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load. Sci Rep 2019;9(1):14668 [FREE Full text] [doi: 10.1038/s41598-019-50280-3] [Medline: 31604964]
100. Gromer D, Reinke M, Christner I, Pauli P. Causal interactive links between presence and fear in virtual reality height exposure. Front Psychol 2019;10:141 [FREE Full text] [doi: 10.3389/fpsyg.2019.00141] [Medline: 30761054]
101. Alfano CA, Beidel DC, Turner SM. Cognitive correlates of social phobia among children and adolescents. J Abnorm Child Psychol 2006;34(2):189-201. [doi: 10.1007/s10802-005-9012-9] [Medline: 16514553]
102. Wallach HS, Safir MP, Bar-Zvi M. Virtual reality cognitive behavior therapy for public speaking anxiety: a randomized clinical trial. Behav Modif 2009;33(3):314-338. [doi: 10.1177/0145445509331926] [Medline: 19321811]
103. Wilhelm FH, Pfaltz MC, Grossman P, Roth W. Distinguishing emotional from physical activation in ambulatory psychophysiological monitoring. Biomed Sci Instrum 2006;42:458-463. [Medline: 16817651]

Abbreviations

AFS: Angstfragebogen für Schüler (English: Anxiety Questionnaire for Pupils)
CBT: cognitive behavioral therapy
DISYPS-III (SBB-ANG): Diagnostic System for Mental Disorders in Children and Adolescents in German (anxiety disorder section, self-report)
HR: heart rate
ICD-10: International Statistical Classification of Diseases, Tenth Revision
IPQ: Igroup Presence Questionnaire
RCI: reliable change index
SAD: social anxiety disorder
SUD: subjective units of distress
VR: virtual reality
VRET: virtual reality exposure therapy

Edited by J Torous; submitted 10.01.24; peer-reviewed by R Pot-Kolder, P Vagos; comments to author 25.04.24; revised version received 11.07.24; accepted 30.07.24; published 05.11.24.

Please cite as:
Beele G, Liesong P, Bojanowski S, Hildebrand K, Weingart M, Asbrand J, Correll CU, Morina N, Uhlhaas PJ
Virtual Reality Exposure Therapy for Reducing School Anxiety in Adolescents: Pilot Study
JMIR Ment Health 2024;11:e56235
URL: https://mental.jmir.org/2024/1/e56235
doi:10.2196/56235
PMID:

Original Paper

The Effect of Explicit Suicide Language in Engagement With a Suicide Prevention Search Page Help-Seeking Prompt: Nonrandomized Trial

Sandersan Onie[1,2], PhD; Susanne Oliver Armstrong[1], PhD; Natasha Josifovski[1], PhD; Patrick Berlinquette[3], BA; Nicola Livingstone[1], BDM; Sarah Holland[1], BCom; Coco Finemore[1], BDS; Nyree Gale[1], MCP; Emma Elder[1], MEd; George Laggis[1], BCom; Cassandra Heffernan[1], BN; Adam Theobald[1], MPsych; Michelle Torok[1], PhD; Fiona Shand[1], PhD; Mark Larsen[1], DPhil

[1]Black Dog Institute, University of New South Wales, Sydney, Australia
[2]Faculty of Psychology, Universitas Airlangga, Surabaya, Indonesia
[3]BERLIN SEM, New York, NY, United States

Corresponding Author:
Sandersan Onie, PhD
Black Dog Institute
University of New South Wales
Hospital Road
Sydney, 2031
Australia
Phone: 61 432359134
Email: s.onie@blackdog.org.au

Abstract

Background: Given that signage, messaging, and advertisements (ads) are the gateway to many interventions in suicide prevention, it is important that we understand what type of messaging works best for whom.

Objective: We investigated whether explicitly mentioning suicide increases engagement using internet ads by investigating engagement with campaigns with different categories of keywords searched, which may reflect different cognitive states.

Methods: We ran a 2-arm study Australia-wide, with or without ads featuring explicit suicide wording. We analyzed whether there were differences in engagement for campaigns with explicit and nonexplicit ads for low-risk (distressed but not explicitly suicidal), high-risk (explicitly suicidal), and help-seeking for suicide keywords.

Results: Our analyses revealed that having explicit wording has opposite effects, depending on the search terms used: explicit wording reduced the engagement rate for individuals searching for low-risk keywords but increased engagement for those using high-risk keywords.

Conclusions: The findings suggest that individuals who are aware of their suicidality respond better to campaigns that explicitly use the word "suicide." We found that individuals who search for low-risk keywords also respond to explicit ads, suggesting that some individuals who are experiencing suicidality search for low-risk keywords.

(JMIR Ment Health 2024;11:e50283) doi:10.2196/50283

KEYWORDS

suicide; suicide prevention; Google; Google Ads; internet search; explicit wording; mental health; suicidal; advertisement; advertisements; messaging; prevention signage; campaign; campaigns; distress; engagement; prompt; prompts; information seeking; help seeking; searching; search

Introduction

Over the past decade, there has been a growth in the different types of interventions for people contemplating suicide. For example, apps can help people keep themselves safe during a suicidal crisis [1]; phone booths are installed at frequently used locations, which give rapid access to a suicide hotline [2]; or an online banner containing a suicide hotline number may appear if individuals search for suicide-related terms using a search engine [3]. Despite these help-seeking pathways, recent research has shown that less than half of individuals contemplating suicide seek professional help before a suicide attempt [4]. Thus, in parallel to developing new interventions, it may be important to understand how best to promote help seeking and access to help for individuals contemplating suicide for existing services and products.

Help-seeking prompts may function as a gateway to these interventions and come in various forms. In this case, a help-seeking prompt refers to any media encouraging service use. For example, a person may prompt an individual to see their general practitioner (GP) for their mental health, a description on an app store may prompt them to download a suicide prevention app, signage at frequently used locations may point them toward a phone booth, or text on an internet search page may encourage them to call a hotline. All these help-seeking prompts play an important role in being the first point of engagement for the individual, introducing the intervention and promoting its use. If the advertisement (ad) cannot encourage the individual to engage with the intervention, it cannot fulfil its role.

One way we can investigate how different types of prompt messaging affect engagement is by using internet search ads. Previous studies have shown that individuals may search for suicide-related terms on the internet before a suicide attempt [5], that search volumes for a particular region correspond to the suicide rate for that region [6], and that internet searches are used to seek help [7]. For example, an individual may use the internet to find the closest crisis center, a local psychiatrist, or an app to help manage their suicidal thoughts. Searching for "suicide help" yields over 1 billion results on Google, and previous studies have shown a high frequency of searches, with over 120,000 searches over a 19-day study period for suicide- and distress-related keywords in Australia alone [7]. Thus, the search page may be an ideal place to intervene and investigate what messaging is most effective for individuals contemplating suicide.

Internet search ads are triggered when keywords from a prepared list are used, presenting an ad at the top of the search results that links to a relevant page or intervention. Internet search ads also offer the ability to assess how different types of messaging perform with people in different cognitive states, reflected by their search terms. For example, individuals searching for keywords associated with suicide but not explicitly communicating suicidality (eg, loneliness, hopelessness) may engage differently with a particular ad wording than individuals searching for keywords explicitly indicating suicidality or seeking help. This allows us to assess what messaging is best for different risk or distress levels.

The reach and effectiveness of these ads are measured with the impressions (how many times an ad is shown), clicks (the number of clicks on the ad), and conversions (specific behaviors performed on the linked website). Engagement is specifically measured with the click rate (the proportion of individuals who saw the ad and then clicked on it), the conversion rate (the proportion of individuals who clicked on the ad and then engaged with the website), and the total conversion rate (the proportion of individuals who saw the ad and then engaged with the website) [7].

A previous study has highlighted that one major discussion area regarding communicating with an individual contemplating suicide is the explicit use of the word "suicide" in the ad [8]. In one component of this study, lived experience advisors indicated that the use of the word "suicide" in ads might alienate some individuals who may indeed be experiencing thoughts of suicide but may not recognize, acknowledge, or identify their thoughts as being of suicide. The advisors elaborated that by not using the word "suicide," we may be able to reach individuals at a precrisis phase for early intervention.

Conversely, other lived experience advisors from the same study communicated that it is imperative to be clear on the subject matter by using the word "suicide." These advisors communicated that by being explicit, we can overcome the stigma associated with the word and the individual contemplating suicide may be more likely to engage with the service as it is specific to their needs or current situation. This is further supported by contemporary suicide first aid programs, which encourage the explicit use of the word "suicide" for the same reason [9]. Furthermore, all lived experience advisors highlighted the importance of understanding what wording is effective for different suicide risk levels to maximize engagement. For example, nonexplicit wording may be particularly effective for individuals contemplating suicide but not in a suicidal crisis, as they may not identify their thoughts being that of suicide—and vice versa.

In this study, we sought to compare engagement with two internet ad campaigns, one with explicit suicide wording in its ad and the other with nonexplicit suicide wording. We also investigated whether the pattern of engagement differed by the type of keyword searched (low risk, high risk, help seeking, or means specific). In addition, we examined engagement by gender, age, and time of day.

First, we hypothesized that the campaign with explicit wording related to suicide would perform better for individuals searching for high-risk, help-seeking, or means-specific keywords. Given that these individuals are explicitly experiencing suicidal ideation and can identify and communicate it, they may respond better to an ad that is explicit in what issue it is addressing. Second, we hypothesized that for individuals searching for low risk-keywords, nonexplicit wording would have higher engagement as the explicit wording may alienate individuals who do not identify as having suicidal thoughts. Third, given that the key manipulation is in the ad's wording and not the linked web pages, we hypothesized that we should see this increased engagement in the click rate but not the conversion rate (and, as such, also see increased engagement in the total conversion rate).

Methods

Study Design

The study used a 2-arm quasi-experimental design (explicit vs nonexplicit wording) with 4 initial pathways (individuals searching for different types of keywords: low risk, high risk, help seeking, and means specific; these categories are described in more detail later).

Participants

Individuals over the age of 18 years and currently residing in Australia were included in the study. Google infers the age and gender of individuals through multiple sources of information, primarily the age and gender inputted when creating a Google account, as well as past browsing history (eg, websites visited and engagements) collected through website cookies.

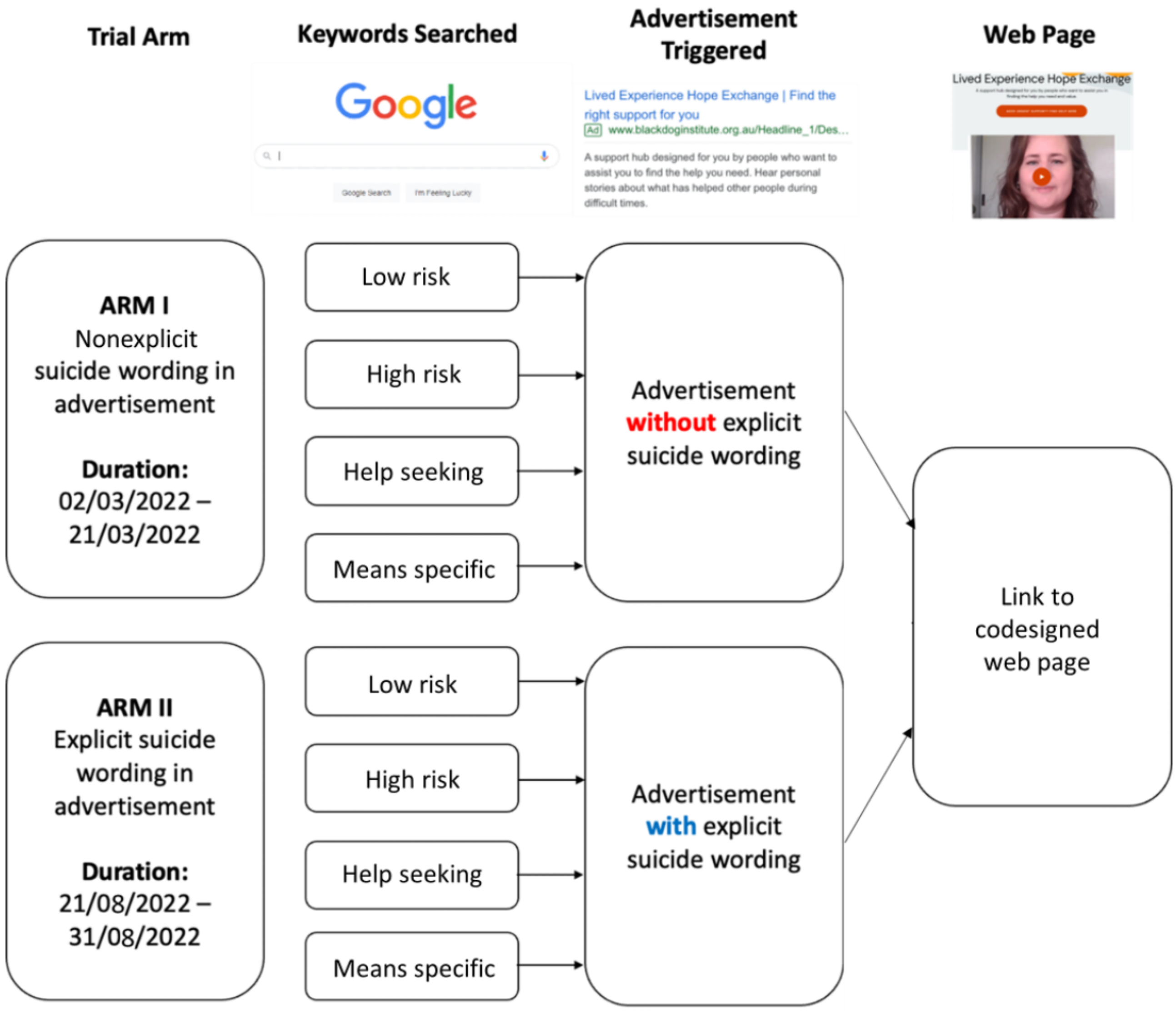
Ethical Considerations

The University of New South Wales Human Research Ethics Committee approved this study (HC210827).

Intervention

The first arm used ads without explicit suicide wording and ran from March 2 to March 21, 2022. The second arm used ads with explicit suicide wording and ran from August 21 to August 31, 2022. The first arm was run as part of another larger study pertaining to the effectiveness of Google Ads campaigns in reaching individuals thinking of suicide and has been reported elsewhere [7]. The 2 arms of the trial were run sequentially, not concurrently. Thus, there was no randomization present. Full details of the keyword generation, ad and landing page codesign process, and content of the landing pages and linked pages can be found elsewhere [8]. A schematic of the campaign can be found in Figure 1.

Figure 1. Schematic of Google Ads campaign.



Keywords

Together with lived experience advisors, researchers, and a Google Ads agent, we generated 4 lists of keywords: low-risk keywords, which included keywords people are likely to search for when in distress or situations associated with suicide, without explicitly mentioning suicide (eg, "feeling so alone," "debt"); high-risk keywords, which included keywords explicitly communicating suicidal ideation or help seeking (eg, "I want to die"); help-seeking keywords, which included keywords explicitly searching for help for suicidal thoughts (eg, "suicide help"); and means-specific keywords, which were related to searching or using specific means [10]. When keywords are entered into Google Ads, the ad is also triggered when semantically similar search terms are entered; thus, the total number of keywords that trigger the ad is substantially larger than what is inputted.

Advertisements

The ads were codesigned alongside a group of lived experience advisors and investigators. The codesign process yielded 6 ads, 3 (50%) explicit suicide wording and 3 (50%) without explicit suicide wording, while controlling as much as possible for other content. Each ad was as closely matched as possible to its counterpart; that is, the first ad in the explicit condition had the same messaging and content as the first ad in the nonexplicit condition, save for explicit suicide wording in the first sentence. We developed more than 1 ad for each condition to allow our findings to be more generalizable and to reflect larger themes found in our codesign process [8] rather than specific wording.

All the ads across both conditions were controlled for the number of characters and number of words, with a range of 38-43 words and 215-234 characters. The absolute difference in the character and word count between each ad and its counterpart was between 1 and 5 characters and 1 and 3 words, respectively.

The codesigned text for the nonexplicit ads is specified next. All character counts include spaces.

> *Lived Experience Hope Exchange. Find the right support for you. A support hub designed for you by people who want to assist you to find the help you need. Hear personal stories about what has helped other people during difficult times. (237 characters, 41 words)*
>
> *Looking for some support? Designed with Lived Experience. Our Hope Exchange has been designed by people who may understand how you're feeling. There are lots of ways to seek help. We want to find the right one for you. (220 characters, 39 words)*
>
> *Find the right support for you. Lived Experience Hope Exchange. We want to help you to find the help that you need and value during challenging times. Hear stories and advice from people who may have felt the way you're feeling now. (234 characters, 42 words)*

The matched text for the explicit wording ads was as follows:

> *Are you feeling suicidal? Lived Experience Hope Exchange. Support designed for you. Designed by people who want to assist you to find the help you need. Hear personal stories about what has helped other people during difficult times. (235 characters, 38 words)*
>
> *Help for suicidal thoughts. Looking for some support? Designed with Lived Experience. Designed by people who may understand how you're feeling. There are lots of ways to seek help. We want to find the right one for you. (221 characters, 38 words)*
>
> *Dealing with suicidal thoughts. Find the right support for you. Lived Experience Hope Exchange. We want to help you to find the help that you need and value. Hear stories and advice from people who may have felt the way you're feeling now. (241 characters, 43 words)*

In each campaign condition, ads were shown to users independently of which category of keywords were searched. When an ad was triggered, 1 of 3 ads in that condition would be randomly shown, resulting in equal presentations across the study.

Landing Page

In collaboration with lived experience advisors, we codesigned a series of landing pages containing lived experience stories, calming and distracting activities, and links to support services and hotlines with descriptions of what the individual will likely experience when engaging in these services. Details of the pages can be found elsewhere [8].

Outcomes

Data on impressions, clicks, the click rate (clicks/impressions), conversions, the conversion rate (conversions/clicks), the cost per click, and the cost per conversion were extracted from Google Ads in a deidentified, aggregated form. The total conversion rate was manually calculated (conversions/impressions). Currently, the total volume of searches for each category is not available through Google Ads.

The primary outcome was the click rate (engagement with the ad) as our manipulation was on the search page rather than on the landing page. Our secondary outcomes were the total conversion rate (total engagement with the campaign, that is, all things being equal, the conversion rate per impression) and the conversion rate (engagement with the landing page).

Conversions contained behaviors the investigators, the lived experience advisors, and the collaborative team considered positive. Triggering any of these conditions was considered a conversion, including:

- Clicking the Get Help button to see available support services
- Downloading any file pertaining to the modules to help de-escalate a crisis or for self-help for suicidality
- Clicking on a link to call a support service
- Spending more than 2 minutes on the website that was designed to promote help seeking and de-escalate crises, as an indication that the individual was engaging with content on the website

Statistical Analyses

In the main analysis, each combination of outcome metric (click rate, conversion rate, and total conversation rate) and keyword type (high risk, low risk, and help seeking) was considered separately. The outcome metric rates associated with the explicit and nonexplicit wording were compared using an incidence rate ratio (IRR) from the *rateratio* function in the *fmsb* package in R (R Foundation for Statistical Computing), which calculates the exact mid-p double-sided *P* value and calculates the CI using the exact Poisson method [11,12].

Interaction terms were considered if at least 1 significant difference between explicit and nonexplicit wording was identified for an outcome measure. In this case, a keyword category with a significant difference due to wording was compared to the other keyword categories. To test the interactions, we first calculated a difference term between the explicit and nonexplicit conditions for a single keyword type. For example, we calculated the difference term for the click rate and low-risk keywords as follows:

This difference term was compared with the other keyword difference terms using the *ratedifference* function in the *fmsb* package in R, which uses a chi-squared test to test for a significant incidence rate difference (IRD) [12]. Thus, this analysis assessed interaction by assessing a difference of a difference.

Where a significant difference of differences (ie, an interaction) was identified in the outcomes, post hoc tests were conducted to determine whether there were underlying differences in the relevant outcome metrics for the explicit or nonexplicit wording conditions or both.

Results

Campaign Metrics

A total of 153,768 impressions, 7263 clicks, and 1657 conversions were achieved during the study periods. The engagement metrics, reported by trial condition, age, and gender, are reported in Table 1.

Due to the exceptionally low numbers in the means-specific group (n=11, 0.01%, impressions; n=1, 0.01%, click; and 0 conversions), these campaigns were excluded from subsequent analyses.

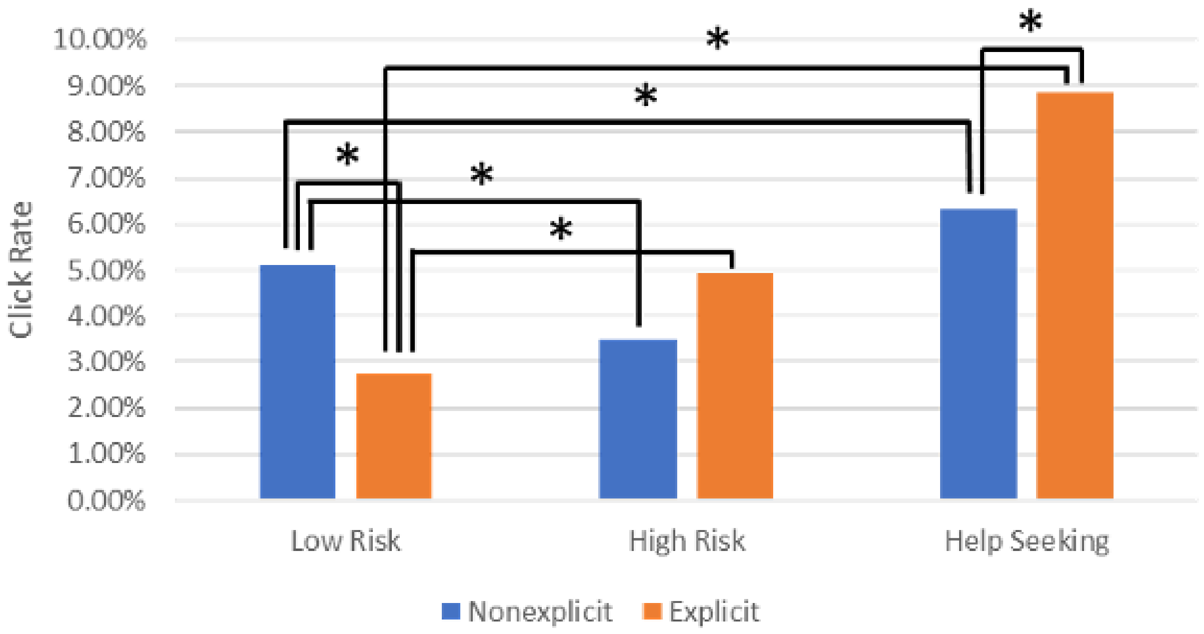
Table 1. Engagement metrics by trial condition, age, and gender.

Trial condition, gender, and age (years)	Impressions (N=153,768), n (%)	Clicks (N=7263), n (%)	Click rate, %	Conversions (N=1657), n (%)	Conversion rate, %	Total conversion rate, %
Nonexplicit trial, gender male						
18-24	6191 (4.03)	308 (4.24)	4.97	39 (2.35)	12.66	0.63
25-34	6955 (4.52)	261 (3.59)	3.75	52 (3.14)	19.92	0.75
35-44	6770 (4.40)	313 (4.31)	4.62	77 (4.65)	24.60	1.14
45-54	6975 (4.54)	398 (5.48)	5.71	122 (7.36)	30.65	1.75
55-64	4644 (3.02)	257 (3.54)	5.53	93 (5.61)	36.19	2.00
≥65	2695 (1.75)	169 (2.33)	6.27	36 (2.17)	21.30	1.34
All ages	34,230 (22.26)	1706 (23.49)	4.98	419 (25.29)	24.56	1.22
Nonexplicit trial, gender female						
18-24	13,243 (8.61)	796 (10.96)	6.01	89 (5.37)	11.18	0.67
25-34	17,448 (11.35)	708 (9.75)	4.06	135 (8.15)	19.07	0.77
35-44	17,996 (11.70)	869 (11.96)	4.83	180 (10.86)	20.71	1.00
45-54	18,963 (12.33)	1013 (13.95)	5.34	311 (18.77)	30.70	1.64
55-64	12,675 (8.24)	737 (10.15)	5.81	187 (11.29)	25.37	1.48
≥65	6326 (4.11)	398 (5.48)	6.29	98 (5.91)	24.62	1.55
All ages	86,651 (56.35)	4521 (62.25)	5.22	1000 (60.35)	22.12	1.15
Explicit trial, gender male						
18-24	1825 (1.19)	62 (0.85)	3.40	4 (0.24)	6.45	0.22
25-34	2281 (1.48)	62 (0.85)	2.72	21 (1.27)	33.87	0.92
35-44	2052 (1.33)	64 (0.88)	3.12	13 (0.78)	20.31	0.63
45-54	1941 (1.26)	76 (1.05)	3.92	19 (1.15)	25.00	0.98
55-64	1351 (0.88)	54 (0.74)	4.00	15 (0.90)	27.78	1.11
≥65	841 (0.55)	46 (0.63)	5.47	8 (0.48)	17.39	0.95
All ages	10,291 (6.69)	364 (5.01)	3.54	80 (4.83)	21.98	0.78
Explicit trial, gender female						
18-24	3436 (2.23)	104 (1.43)	3.03	16 (0.97)	15.38	0.47
25-34	4474 (2.91)	108 (1.49)	2.41	21 (1.27)	19.44	0.47
35-44	4489 (2.92)	140 (1.93)	3.12	30 (1.81)	21.43	0.67
45-54	4824 (3.14)	136 (1.87)	2.82	37 (2.23)	27.21	0.77
55-64	3381 (2.20)	97 (1.34)	2.87	28 (1.69)	28.87	0.83
≥65	1992 (1.30)	87 (1.20)	4.37	26 (1.57)	29.89	1.31
All ages	22,596 (14.69)	672 (9.25)	2.97	158 (9.54)	23.51	0.70
Combined trial						
All ages, male	44,521 (28.95)	2070 (28.50)	4.65	499 (30.11)	24.11	1.12
All ages, female	109,247 (71.05)	5193 (71.50)	4.75	1158 (69.89)	22.30	1.06

Click Rate

The click rate using explicit and nonexplicit wording for each keyword category is shown in Figure 2. A significant difference between explicit and nonexplicit wording for low-risk keywords was found (IRR=1.848, 95% CI 1.718-1.987, *P*<.001), in which there was a higher click rate for nonexplicit versus explicit keywords (5.11% vs 2.77%). For help-seeking keywords, there was a significantly higher click rate for explicit versus nonexplicit keywords (3.50% vs 4.93%; IRR=0.715, 95% CI 0.599-0.854, *P*<.001). A similar pattern was observed for high-risk keywords, with a higher click rate for explicit versus nonexplicit keywords, although this was only marginally nonsignificant (6.33% vs 8.85%; IRR=0.711, 95% CI 0.503-1.005, *P*=.052).

Figure 2. Click rate by keyword type and condition. *P<.05.



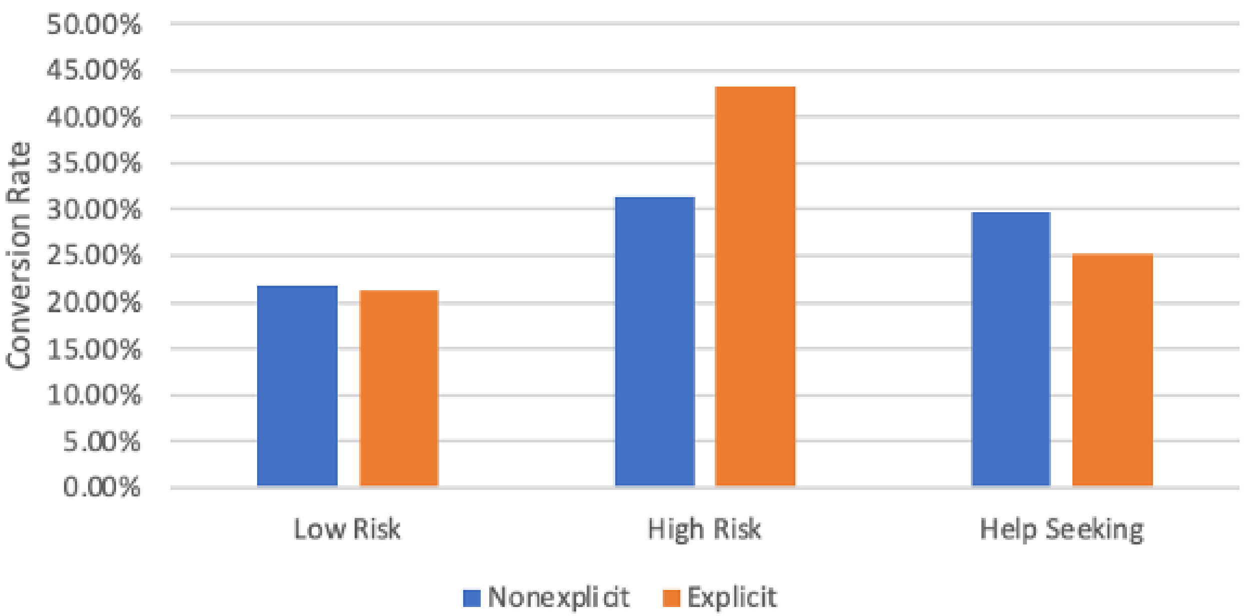
Next, we investigated the presence of 2-way interactions between the 3 groups (low risk and high risk, low risk and help seeking, high risk and help seeking). All 2-way interactions were significant (high risk and help seeking: IRD=0.0109, 95% CI 0.000146-0.0217, *P*=.047; high risk and low risk: IRD=0.0301, 95% CI 0.0301-0.0453, *P*<.001; low risk and help seeking: IRD=0.0472, 95% CI 0.0405-0.0541, *P*<.001).

Post hoc tests revealed a significant difference in the click rate between the low- and high-risk keywords in the explicit condition (2.77% vs 4.93%; IRR=0.5612, 95% CI 0.4144-0.7784, *P*<.001) and the nonexplicit condition (5.11% vs 3.50%; IRR=1.459, 95% CI 1.216-1.766, *P*<.001). Significant differences were also found in the click rate between the low-risk and help-seeking keywords in the explicit condition (2.77% vs 8.85%; IRR=0.312, 95% CI 0.263-0.373, *P*<.001) and the nonexplicit condition (5.11% vs 6.33%; IRR=0.807, 95% CI 0.740-0.881, *P*<.001).

Conversion Rate

A graphical representation of the conversion rate data is shown in Figure 3. There were no significant differences between the nonexplicit and explicit conditions (low risk: 21.86% vs. 21.17%; IRR=1.0326, 95% CI 0.882-1.209, *P*=.69; high risk: 31.36% vs 43.18%; IRR=0.726, 95% CI 0.418-1.263, *P*=.25; help seeking: 29.74% vs 25.16%; IRR=1.182, 95% CI 0.835-1.674, *P*=.35). Thus, interaction effects were not explored.

Figure 3. Conversion rate by keyword type and condition.



Total Conversion Rate

A graphical representation of the total conversion rate can be seen in Figure 4. There was a significant difference in the total conversion rate for low-risk keywords (IRR=1.908, 95% CI 1.630-2.234, *P*<.001), in which the nonexplicit wording had a higher rate (1.12% vs 0.59%), and high-risk keywords (IRR=0.846, 95% CI 0.297-0.896, *P*=.02), in which the explicit wording had a higher rate (1.10% vs 2.13%); however, there was not enough evidence to suggest a difference between explicit and nonexplicit wording (1.88% vs 2.23%) when an individual was searching for help-seeking keywords (IRR=0.846, 95% CI 0.597-1.197, *P*=.34).

Possible interactions between the 3 groups (low risk and high risk, low risk and help seeking, high risk and help seeking) were explored. A significant interaction was identified between low-risk and high-risk keywords (IRD=0.00875, 95% CI 0.00613-0.0114, *P*<.001), as well as between low-risk and help-seeking keywords (IRD=0.0156, 95% CI 0.00901-0.0222, *P*<.001). However, there was no significant interaction between high-risk and help-seeking keywords (IRD=–0.00685, 95% CI –0.0141 to 0.000345, *P*=.06).

Post hoc tests revealed a significant difference in the total conversion rate between the low- and high-risk keywords in the explicit condition (0.59% vs 2.13%; IRR=0.275, 95% CI 0.171-0.442, *P*<.001) but not in the nonexplicit condition (1.12% vs 1.10%; IRR=1.017, 95% CI 0.733-1.411, *P*=.92).

A summary of the analysis outcomes can be seen in Table 2.

Figure 4. Total conversion rate by keyword type and condition. **P*<.05.

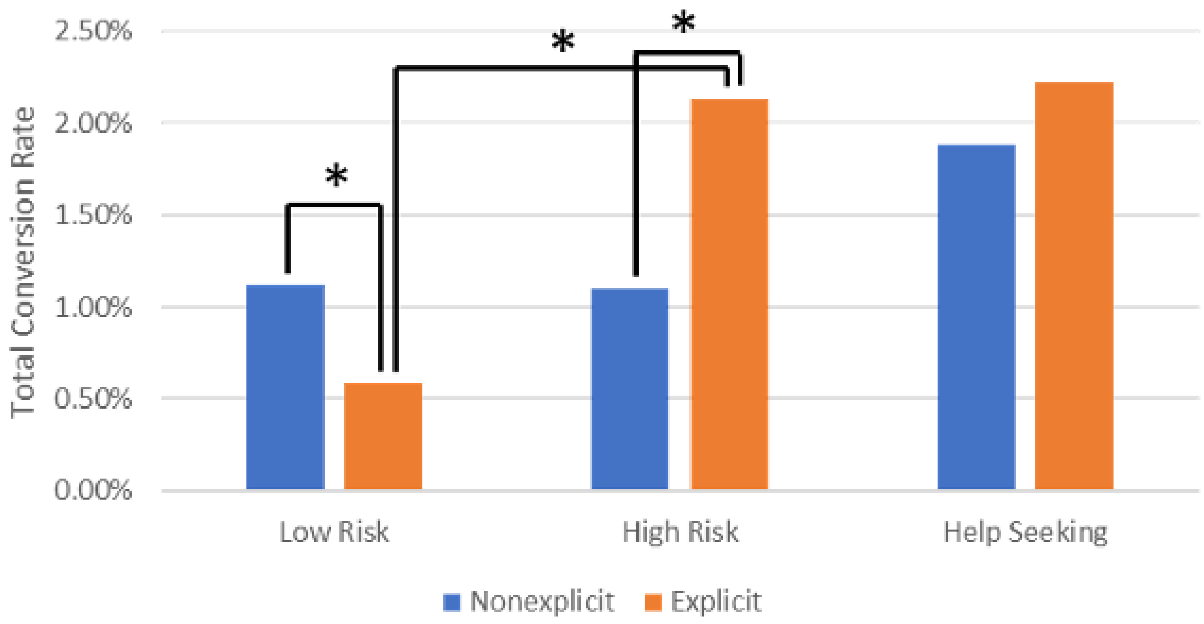


Table 2. Summary of analysis outcomes.

Metric and keyword type	Significant comparisons	Significant interactions
Click rate		
Low risk	Nonexplicit > explicit	Low risk × high risk Low risk × help seeking High risk × help seeking
High risk	Explicit > nonexplicit[a]	Low risk × high risk Low risk × help seeking High risk × help seeking
Help seeking	Explicit > nonexplicit	Low risk × high risk Low risk × help seeking High risk × help seeking
Conversion rate		
Low risk	N/S[b]	—[c]
High risk	N/S	—
Help seeking	N/S	—
Total conversion rate		
Low risk	Nonexplicit > explicit	Low risk × high risk Low risk × help seeking
High risk	Explicit > nonexplicit	Low risk × high risk Low risk × help seeking
Help seeking	N/S	Low risk × high risk Low risk × help seeking

[a]Marginal significance (*P*<.06).

[b]N/S: not significant.

[c]Not available. This was used when interaction analyses were not conducted due to nonsignificant comparisons.

Discussion

Principal Findings

In this study, we compared the impact of explicit and nonexplicit suicide wording in an online ad campaign and webpage targeting those searching for suicide- and distress-related keywords. Analysis of the click rate revealed that for low-risk keywords, nonexplicit wording had a higher click rate; for high-risk keywords, there was marginal evidence that explicit wording had a higher click rate; and for help-seeking keywords, explicit wording had a higher click rate. For the conversion rate, there was no evidence of any differences between conditions. Analysis of the total conversion rate revealed that for low-risk keywords, nonexplicit wording had a higher total conversion rate; for high-risk keywords, explicit wording had a higher total conversion rate; and for help-seeking keywords, there was no evidence of a difference.

Further analysis of the click rate revealed 2-way interactions between low- and high-risk keywords, low-risk and help-seeking keywords, and high-risk and help-seeking keywords, suggesting that the effect of explicit wording in the ad differed among these groups. Further exploratory analysis revealed significant differences between low-risk and high-risk keywords, as well as between low-risk and help-seeking keywords, in the explicit and nonexplicit conditions. These findings further support the suggestion that the effect of wording differentially impacts engagement in keyword groups, rather than an interaction emerging due to the manipulation only affecting one group but not the other. Together, these findings suggest that ads with explicit suicide language are less likely to be clicked on than those with nonexplicit language when individuals are searching for low-risk keywords. The reverse was observed when searching for high-risk or help-seeking keywords (although the former did not reach significance), where ads with explicit language were more likely to be clicked on. This pattern of findings may partly contribute to the pattern of findings in the total conversion rate, in which analysis revealed 2-way interactions between low- and high-risk keywords and between low-risk and help-seeking keywords, suggesting that the effect of explicit wording in the ad differed among these groups. Exploratory analysis revealed that the interaction between low- and high-risk keywords may be driven by a higher total conversion rate when explicit wording is used with high-risk versus low-risk keywords, whereas there was no apparent difference between high-risk and low-risk keywords when nonexplicit wording was used. These findings suggest that having explicit wording has opposite effects, depending on the search terms used: explicit wording reduces the total conversion rate for individuals searching for low-risk keywords but increases the total conversion rate for those using high-risk keywords. There is no evidence to suggest that explicit or nonexplicit wording affects the total conversion rate when help-seeking keywords are used. These findings support both recommendations from lived experience advisors, both for and against the use of explicit wording, as perhaps both are appropriate for individuals in different cognitive states.

Overall, these findings suggest that individuals who search for help-seeking or high-risk suicide keywords respond more to ads and campaigns with explicit suicide wording, demonstrated by the higher click rate and total conversion rate, respectively. Alternatively stated, an ad campaign targeting individuals searching for high-risk keywords is likely to lead to more desirable behaviors if explicit wording is used. This may be because individuals may respond more strongly to a campaign that specifically targets their current situation. Given that the ad explicitly communicates and labels their current issue, the individuals may be more inclined to seek help. As the nonexplicit ad contained no indication that the campaign is for suicidality, the users may not have been certain that the webpage was able to meet their needs.

Furthermore, the high-risk search terms in this study excluded individuals who were explicitly searching for help for their suicidality. Thus, our findings suggest that explicitly naming the issue can improve help seeking for individuals at high risk of suicide but who may not be actively seeking help, as revealed in the total conversion rate. This finding is consistent with current practices in suicide prevention first aid (eg, applied suicide prevention skills training [9], where directly addressing and asking about suicide are strongly encouraged. This study demonstrated that the benefit of explicit wording for individuals in crisis is generalizable beyond suicide first aid and direct face-to-face communication. This finding may have implications for other forms of communication when addressing individuals at high risk of suicide.

The results also showed that individuals searching for low-risk suicide terms respond less to campaigns with explicit suicide wording. Since the low-risk keywords in this study were broad (eg, loneliness), many individuals searching for these terms may not have been suicidal—hence the lower click rate. However, despite being low, we still observed a click rate by individuals searching for low-risk keywords on the explicit campaign compared to the industry standard of 3.17% [13]. This suggests that individuals who may not be searching for explicit keywords may be experiencing suicidality and that targeting low-risk keywords is still beneficial. Conversely, we may have seen an elevated click rate for the low-risk, nonexplicit condition due to people clicking on the ad not realizing it was for individuals experiencing suicidality. Nevertheless, in the low-risk, nonexplicit condition, we still observed a conversion rate of over 20%, relative to the industry standard of 3.75%, suggesting that the landing page was still fulfilling a need. This may be because by using explicit suicide wording, we may alienate individuals who, for various reasons, may not recognize, identify, or acknowledge that their feelings are those of suicide. Thus, when visiting the page, their needs are met. However, one possibility is that some individuals searching for low-risk keywords are unaware that they are suicidal, and using the word "suicide" may help bring awareness to these underlying feelings [8]. Thus, explicit and nonexplicit keywords may have their own benefits; however, our findings suggest that overall, using nonexplicit keywords will reach more people. Although many people searching for low-risk keywords may not be experiencing suicidality, we must ensure that little effort is needed to access suicide-related resources and help, given that there are still individuals searching for low-risk keywords who are experiencing suicidality. Further research is needed to understand how best to tailor the ads to individuals searching for low-risk keywords.

The findings suggest that individuals experiencing suicidality and who could explicitly communicate it have higher engagement patterns with a campaign when the word "suicide" is used in the ad regardless of whether they are explicitly seeking help. However, for individuals experiencing general distress but not searching for suicide-specific terms, using explicit suicide wording leads to lower engagement with the campaign. Thus, in response to the finding that some lived experience advisors advocated for the explicit use of the word "suicide," while others advocated against it, perhaps both are true for individuals in different cognitive states.

We recommend that the development and design of help-seeking prompts for suicide consider at what stage the individual is. If the prompt is intended for individuals with a lower risk of suicide, such as public media campaigns, then the use of the word "suicide" may decrease engagement; however, if the prompt is intended for individuals at high risk of suicide, whether they are or are not actively seeking help, then the use of the word "suicide" is likely to increase engagement. Furthermore, these findings suggest the need to codesign with a range of individuals who have experienced the spectrum of suicidality to understand their needs, the thought process, and how they speak about and internally conceptualize their distress and suicide to formulate different terms that promote help seeking and engagement.

Given that this pattern of finding has been found across 2 modalities (suicide first aid and internet ads), future research should seek to investigate the generalizability of these findings to other help-seeking prompts in suicide prevention, for example, signage at frequently used locations, safety planning app notifications, or the wording on the suicide hotline banner if individuals search for suicide-related terms. Further, future research should further understand what type of ad wording works for whom. For example, previous evidence has suggested that men and women respond differently to tailored ad campaigns [14]. Using Google Ads, we can further investigate what type of ad best engages men and women.

There was a low number of searches and engagement for means-specific keywords in this study. There may be several possibilities for this. First, these numbers may reflect true rates and only a few individuals were searching for these terms. Another possibility is that individuals at the planning stage do not primarily turn to search engines but may use other means of information seeking, as means selection has been found to be influenced by prior familiarity with the means itself [15]. Another possibility is that more individuals were searching for these terms but our current keyword list and Google's function of generating permutations of the keyword list could not capture the range of search terms. This could be rectified by experimenting with keyword setting iterations in Google Ads to capture a wider variety of expressions. Finally, one possibility is that there are cultural differences in search behaviors. For example, a previous study found high incidences of method-related searches in Japan [16], suggesting that cultural differences may exist. Given that a key finding in previous research was that individuals who attempted suicide had searched for means-specific keywords [5], future research should investigate how individuals search for information regarding means and how we can best intervene at this stage.

Future studies may also investigate whether these findings generalize to other psychological or health domains. For example, an individual who is acutely depressed and can identify it may respond better to an ad for therapy explicitly communicating that this is for individuals experiencing depression. Conversely, an individual who may also be experiencing depressive symptoms but is either unfamiliar that they are experiencing depressive symptoms or is from a background where depression is strongly stigmatized may respond better to an ad without explicitly using the word "depression." Future studies should carefully consider conversion actions, as we cannot measure directly whether an individual is processing the information presented to them, and thus, we use proxy measures common to the marketing field (eg, time spent on a page). Given the growing interest in online interventions, standardized methods for measuring engagements should be established.

Limitations

The study has several limitations. One limitation is that due to resourcing constraints, there was a large difference in the number of participants in the 2 conditions. Furthermore, the data were collected at different times of the year, so there may be seasonal or cohort effects; future studies should run campaigns with explicit and nonexplicit wording at the same time. In addition, as research with this type of data is still in its infancy, future research should focus on understanding whether, when, and how search metrics should be normalized against variations in time. Another limitation is that we could only infer cognitive states from search results but did not measure suicidality directly. For example, a person may search for a high-risk keyword as part of a study but not be suicidal. Thus, there is added noise in the data.

Strengths

Our study also has several strengths. The study was run nationwide, allowing us to sample the entire target population rather than just a specific subset. The components of this study, such as the ad wording, landing page, and keywords used to trigger the ad, were codesigned with individuals with lived and living experiences of suicide. Furthermore, by providing data on click rates and conversion rates, we obtained greater mechanistic insight into the findings for the total conversion rate.

Conclusion

Our study demonstrates different engagement levels with an online suicide prevention campaign due to the word "suicide" in the search page ad. Future research should further explore what type of messaging works best for whom, and when paired with the flexibility of the advertising industry, we may be one step closer to ensuring that each person is met with a message that leads to the highest probability of them engaging services, using resources, or seeking help.

Acknowledgments

SO led the conceptualization, methodology, data curation, formal analysis, funding acquisition, project administration, and writing—original draft. SOA coled writing—original draft. ML led supervision and coled funding acquisition, formal analysis, and writing—review and editing. MT and FS coled funding acquisition and writing—review and editing. PB led resources and software. All other authors supported the methodology and investigation. This study was supported by the Australian Government Department of Health–funded National Suicide Prevention Research Fund, managed by Suicide Prevention Australia, National Health and Medical Research Council (NHMRC) Centre of Research Excellence in Suicide Prevention (APP1152952).

Conflicts of Interest

None declared.

References

1. Han J, McGillivray L, Wong QJ, Werner-Seidler A, Wong I, Calear A, et al. A Mobile Health Intervention (LifeBuoy App) to Help Young People Manage Suicidal Thoughts: Protocol for a Mixed-Methods Randomized Controlled Trial. JMIR Res Protoc 2020 Oct 27;9(10):e23655 [FREE Full text] [doi: 10.2196/23655] [Medline: 33107832]
2. Lockley A, Cheung YTD, Cox G, Robinson J, Williamson M, Harris M, et al. Preventing suicide at suicide hotspots: a case study from Australia. Suicide Life Threat Behav 2014 Aug;44(4):392-407. [doi: 10.1111/sltb.12080] [Medline: 25250406]
3. Heussner KM. Google 'suicide' search feature offers lifeline. ABC News. 2010 Apr 10. URL: https://abcnews.go.com/Technology/google-suicide-search-feature-offers-lifeline/story?id=10313064 [accessed 2024-02-23]
4. Stene-Larsen K, Reneflot A. Contact with primary and mental health care prior to suicide: a systematic review of the literature from 2000 to 2017. Scand J Public Health 2019 Feb 05;47(1):9-17. [doi: 10.1177/1403494817746274] [Medline: 29207932]

5. Areán PA, Pratap A, Hsin H, Huppert TK, Hendricks KE, Heagerty PJ, et al. Perceived utility and characterization of personal Google search histories to detect data patterns proximal to a suicide attempt in individuals who previously attempted suicide: pilot cohort study. J Med Internet Res 2021 May 06;23(5):e27918 [FREE Full text] [doi: 10.2196/27918] [Medline: 33955838]
6. Ma-Kellams C, Or F, Baek JH, Kawachi I. Rethinking suicide surveillance: Google search data and self-reported suicidality differentially estimate completed suicide risk. Clin Psychol Sci 2015 Aug 11;4(3):480-484 [FREE Full text] [doi: 10.1177/2167702615593475]
7. Onie S, Berlinquette P, Holland S, Livingstone N, Finemore C, Gale N, et al. Suicide prevention using Google Ads: randomized controlled trial measuring engagement. JMIR Ment Health 2023 Apr 20;10:e42316 [FREE Full text] [doi: 10.2196/42316] [Medline: 37079348]
8. Onie S. Co-design and development of an internet search intervention using Google Adwords for individuals contemplating suicide: the lived experience hope exchange. PsyArXiv Preprint posted online 2023. [doi: 10.31234/osf.io/kzhgw]
9. Gould M, Cross W, Pisani AR, Munfakh JL, Kleinman M. Impact of applied suicide intervention skills training on the National Suicide Prevention Lifeline. Suicide Life Threat Behav 2013 Dec;43(6):676-691 [FREE Full text] [doi: 10.1111/sltb.12049] [Medline: 23889494]
10. Google Ads co-design. Center for Open Science. URL: https://osf.io/d6m45/ [accessed 2024-02-24]
11. Hanley H. Analysis of crude data. In: Rothman KJ, editor. Modern Epidemiology. Boston, MA: Little, Brown & Co; 1986.
12. Sahai H, Khurshid A. Statistics in Epidemiology: Methods, Techniques, and Applications. Boca Raton, FL: CRC Press; 1996.
13. Irvine M. Google Ads benchmarks for your industry [updated!]. WordStream. 2024. URL: https://www.wordstream.com/blog/ws/2016/02/29/google-adwords-industry-benchmarks [accessed 2024-02-23]
14. Lee S, Torok M, Shand F, Chen N, McGillivray L, Burnett A, et al. Performance, cost-effectiveness, and representativeness of Facebook recruitment to suicide prevention research: online survey study. JMIR Ment Health 2020 Oct 22;7(10):e18762 [FREE Full text] [doi: 10.2196/18762] [Medline: 33090115]
15. Milner A, Witt K, Maheen H, LaMontagne AD. Access to means of suicide, occupation and the risk of suicide: a national study over 12 years of coronial data. BMC Psychiatry 2017 Apr 04;17(1):125 [FREE Full text] [doi: 10.1186/s12888-017-1288-0] [Medline: 28376757]
16. Takahashi A, Sueki H, Ito J. Reflection of suicidal ideation in terms searched for by Japanese internet users. Crisis 2023 Jul;44(4):361-364 [FREE Full text] [doi: 10.1027/0227-5910/a000854] [Medline: 35383467]

Abbreviations

ad: advertisement
IRD: incidence rate difference
IRR: incidence rate ratio

Edited by J Torous; submitted 03.07.23; peer-reviewed by H Sueki, H Hsin; comments to author 01.08.23; revised version received 02.10.23; accepted 22.10.23; published 19.03.24.

Please cite as:
Onie S, Armstrong SO, Josifovski N, Berlinquette P, Livingstone N, Holland S, Finemore C, Gale N, Elder E, Laggis G, Heffernan C, Theobald A, Torok M, Shand F, Larsen M
The Effect of Explicit Suicide Language in Engagement With a Suicide Prevention Search Page Help-Seeking Prompt: Nonrandomized Trial
JMIR Ment Health 2024;11:e50283
URL: https://mental.jmir.org/2024/1/e50283
doi:10.2196/50283
PMID:38502162

Original Paper

Web-Based, Human-Guided, or Computer-Guided Transdiagnostic Cognitive Behavioral Therapy in University Students With Anxiety and Depression: Randomized Controlled Trial

Jurrijn Koelen[1*], PhD; Anke Klein[1*], PhD; Nine Wolters[1], PhD; Eline Bol[1], MSc; Lisa De Koning[1], MSc; Samantha Roetink[1], MSc; Jorien Van Blom[1], MSc; Bruno Boutin[2], MSc; Jessica Schaaf[3], PhD; Raoul Grasman[4], PhD; Claudia Maria Van der Heijde[5], PhD; Elske Salemink[6], PhD; Heleen Riper[7], PhD; Eirini Karyotaki[7], PhD; Pim Cuijpers[7], PhD; Silvia Schneider[8], PhD; Ronald Rapee[9], PhD; Peter Vonk[5], MD; Reinout Wiers[1], PhD

[1]Department of Developmental Psychology, University of Amsterdam, Amsterdam, Netherlands
[2]Technical Support Psychology Department, University of Amsterdam, Amsterdam, Netherlands
[3]Cognitive Neuroscience Department, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, Netherlands
[4]Department of Psychological Methods, University of Amsterdam, Amsterdam, Netherlands
[5]Department of Research, Development and Prevention, Student Health Service, University of Amsterdam, Amsterdam, Netherlands
[6]Experimental Psychopathology and Clinical Psychology Lab, Department of Clinical Psychology, Utrecht University, Utrecht, Netherlands
[7]Amsterdam Public Health Research Institute, Department of Clinical Neuro- and Developmental Psychology, Vrije Universiteit, Amsterdam, Netherlands
[8]Mental Health Research and Treatment Center, Department of Clinical Child and Adolescent Psychology, Ruhr-Universität Bochum, Bochum, Germany
[9]Centre for Emotional Health, Department of Psychology, Macquarie University, Sydney, Australia
[*]these authors contributed equally

Corresponding Author:
Reinout Wiers, PhD
Department of Developmental Psychology
University of Amsterdam
P.O.Box 15916
1001 NK
Amsterdam
Netherlands
Phone: 31 20 5256842
Email: r.w.h.j.wiers@uva.nl

Abstract

Background: Internet-based cognitive behavioral interventions (iCBTs) are efficacious treatments for depression and anxiety. However, it is unknown whether adding human guidance is feasible and beneficial within a large educational setting.

Objective: This study aims to potentially demonstrate the superiority of 2 variants of a transdiagnostic iCBT program (human-guided and computer-guided iCBT) over care as usual (CAU) in a large sample of university students and the superiority of human-guided iCBT over computer-guided iCBT.

Methods: A total of 801 students with elevated levels of anxiety, depression, or both from a large university in the Netherlands were recruited as participants and randomized to 1 of 3 conditions: human-guided iCBT, computer-guided iCBT, and CAU. The primary outcome measures were depression (Patient Health Questionnaire) and anxiety (Generalized Anxiety Disorder scale). Secondary outcomes included substance use–related problems (Alcohol Use Disorder Identification Test and Drug Abuse Screening Test—10 items). Linear mixed models were used to estimate the effects of time, treatment group, and their interactions (slopes). The primary research question was whether the 3 conditions differed in improvement over 3 time points (baseline, midtreatment, and after treatment) in terms of depression and anxiety symptoms. Results were analyzed according to the intention-to-treat principle using multiple imputation. Patients were followed exploratively from baseline to 6 and 12 months.

Results: In both short-term and long-term analyses, the slopes for the 3 conditions did not differ significantly in terms of depression and anxiety, although both web-based interventions were marginally more efficacious than CAU over 6 months (*P* values between .02 and .03). All groups showed significant improvement over time (*P*<.001). For the secondary outcomes, only significant improvements over time (across and not between groups) were found for drug use (*P*<.001). Significant differences were found in terms of adherence, indicating that participants in the human-guided condition did more sessions than those in the computer-guided condition (P=.002).

Conclusions: The transdiagnostic iCBT program offers a practical, feasible, and efficacious alternative to usual care to tackle mental health problems in a large university setting. There is no indication that human guidance should be preferred over technological guidance. The potential preference of human support also depends on the scale of implementation and cost-effectiveness, which need to be addressed in future trials.

Trial Registration: International Clinical Trials Registry Platform NL7328/NTR7544; https://trialsearch.who.int/Trial2.aspx?TrialID=NL-OMON26795

(JMIR Ment Health 2024;11:e50503) doi:10.2196/50503

KEYWORDS

internet-based cognitive behavioral intervention; iCBT; university students; transdiagnostic; human guidance; technological guidance

Introduction

Background

University students face various challenges, both academically and personally [1]. During the phase of emerging adulthood (18-25 years of age), difficulties may arise in facing these challenges [2]. Recent large-scale studies among undergraduate students have shown that between 20% and 30% of students experience a mental disorder, most notably depression and anxiety disorders [3-5]. In addition, student life has been associated with excessive alcohol consumption, with some studies indicating that 20% of students are hazardous drinkers [6]. Symptoms of common mental disorder impact the ability to face key challenges of emerging adulthood, such as identity formation and building new, intimate relationships [7,8]. Mental distress may also contribute to students performing poorly academically or even dropping out. This may give rise to concerns about career prospects and finding a place in society [7-9]. These dynamics may lead to a vicious cycle of increasing feelings of failure, which may further exacerbate occurring symptoms. Thus, timely intervention in this group is needed [10,11].

Despite the availability of local services at or around the university, most students do not receive the help they need, and some do not even seek help in the first place [12]. Auerbach et al [4], based on surveys in 21 countries, found that less than 20% of students with a recent or current mental disorder received *minimally adequate treatment* for this disorder, and this rate is even lower in lower middle–income to low-income countries. Even with severe symptoms of a common mental disorder, the 12-month treatment rates did not exceed 45.1%, meaning that half of all students with longer-lasting severe symptoms do not receive help for their problems [12]. Impediments to find help include both factual (eg, lack of skilled therapists, waitlists, lack of time, and financial hurdles) and perceived barriers (eg, skepticism on treatment effectiveness, lack of perceived urgency, and fear of stigmatization) [13-17]. High levels of unmet treatment needs emphasize the urgency to explore alternative possibilities for intervention and potentially a reallocation of resources for this vulnerable group [12].

To address some of these barriers, innovative, scalable, and low-threshold interventions are required. Web-based treatment may meet some of these requirements and has been shown effective for this target group, particularly for depression and anxiety [18-20]. Another potential advantage of web-based interventions is that they require less human investment in terms of time compared with face-to-face clinical care and thus may reduce costs, although this should be quantified in future studies explicitly aimed at assessing cost-effectiveness (not the objective of this study). Moreover, young adults may prefer easily accessible and relatively anonymous and self-directed interventions with limited therapeutic contact [14,17]. Meta-analytic findings have indicated that older people are more likely to respond better to web-based treatments for depression; thus, methods to increase efficacy for young adults need to be found [21].

Transdiagnostic web-based approaches may broaden the scope and impact of interventions. This is particularly relevant given the high comorbidity of symptoms among students [5]. There is a growing consensus that depression and anxiety, although often classified as distinct disorders, share common etiological and perpetuating factors, such as proneness to internalization and high levels of negative affect [22,23]. Targeting anxiety and depression in one treatment program may appeal to a large variety of students. However, at the same time, studies using web-based transdiagnostic interventions among students have reached inconsistent findings, although these studies had limited scope and compared treatment to a waitlist control group only [24,25]. In general, the potential of transdiagnostic interventions for large-scale implementation should not be underestimated, and they could have secondary effects on related problems, such as substance use issues [26].

Evidence is accruing that *guided* (cognitive behavioral therapy [CBT]) interventions should be preferred over *unguided* or self-guided interventions with only limited (ie, technological) support, given their greater efficacy and positive impact on adherence [27,28]. Karyotaki et al [27], for example, found some advantageous effects of guided versus unguided internet-based CBT (iCBT) for depression, but these differences disappeared at the 6-month and 12-month follow-up. This has important practical implications, as offering human support exerts greater pressure on extant resources and could thus limit implementation on a large scale (ie, in a university setting).

Guided and unguided web-based interventions have rarely been compared head-to-head among students with mental health issues [28]. In this study, we use a slightly different terminology for the guided and unguided iCBT conditions, as we think these terms are unfortunate for several reasons (for a discussion, see the study by Koelen et al [28]).

Aims of This Study

In this study, we compared 3 conditions simultaneously in a large sample of both undergraduate and graduate (including doctoral) university students (henceforth, students): a human-guided iCBT transdiagnostic program, a computer-guided iCBT transdiagnostic program, and care as usual (CAU). This study built upon a previous study [10] comparing a human-guided (transdiagnostic) iCBT against CAU in a sample of students (n=100). This previous study [10] yielded no significant differences (P>.05) between conditions on any of the examined outcomes, at any of the time points (including the 6-month and 12-month follow-up), which may in part be attributable to limited power to detect small effects. In addition, this previous study [10] included only students with mild to moderate anxiety and depression symptoms. As other iCBT trials indicate that interventions are more efficacious when offered to individuals with higher levels of anxiety and depression symptoms [25,27], we also included students with elevated levels of anxiety and depression beyond mild symptoms in this study.

The aims of this study were two-fold: (1) to examine the efficacy of a transdiagnostic iCBT program in a large sample of students with mild, moderate, or severe levels of anxiety, depression, or both, when compared with CAU, and (2) to examine whether the addition of human guidance resulted in greater efficacy when compared with computer guidance. It was expected that (1) individuals who followed either human-guided or computer-guided iCBT would improve more than those in the CAU condition on the primary outcomes of depression and anxiety after treatment and (2) individuals who followed human-guided iCBT would improve more on the primary measures compared with computer-guided iCBT. Exploratory analyses were performed up until 6 and 12 months after baseline and for the secondary outcomes. Higher adherence rates were expected for the human-guided compared with the computer-guided condition [28].

Methods

Design

This study was a 3-arm, randomized controlled superiority trial conducted at the University of Amsterdam (UvA) in the Netherlands. A web-based, personalized, and transdiagnostic intervention (iCare Prevent) [29] with human guidance was compared with the same intervention with computer guidance. Both web-based interventions were also compared with CAU. CAU in this context refers to the standard mental health care that is accessible in the university setting, including help provided by the (student) general practitioners, student psychologists, and study advisers, as well as the secondary services in the broader community (psychologists and psychiatrists). Participants were followed up to 12 months after randomization. Measurements were administered at baseline (t1); midtreatment (5 weeks after randomization; t2); after treatment (8 weeks after randomization; t3); 6-month follow-up (t4); and 12-month follow-up (t5).

Ethical Considerations

This study was approved by the medical ethical committee of the *Centrale Commissie Mensgebonden Onderzoek* (METC number 2018_085, NL64929.018.18).

Participants

Participants were young adults who were enrolled as bachelor’s, master’s, or PhD students at the UvA. Before inclusion in the randomized controlled trial, students participated in an e-survey to screen for elevation of symptoms of anxiety or depression (mild, moderate, or severe) [30]. Students were recruited between February 2019 and March 2022 and were included in the randomized controlled trial based on the following inclusion criteria: (1) aged ≥16 years and (2) mild, moderate, or severe symptoms of depression (as defined by a score within the range of 15 to 60 points on the Center for Epidemiological Studies Depression Scale) [31], anxiety (as defined by scoring above the cut-off score of 4 on the Generalized Anxiety Disorder scale–7 items [GAD-7]) [32] on the e-survey, or both.

Participants were excluded when they fulfilled one or more of the following criteria: (1) comorbid bipolar disorder or psychotic disorder according to the Mini International Neuropsychiatric Interview (MINI) [33]; (2) active high suicide risk according to the MINI; (3) currently receiving psychological treatment for depression or anxiety; (4) having a slow or no internet connection; and (5) no provision of providing written informed consent before participation.

Procedures of Screening and Intervention Phases

Students enrolled at the UvA received an invitation via email from the secured, central research-dedicated platform (LOTUS). LOTUS contained the students’ email addresses, from which links for the e-survey platform Qualtrics and other messages were sent, depending on the phase of the study and the forms completed. The invitation email contained general information about the study and a unique link. Participants who clicked on the invitation link were referred to the e-survey platform. Here, they could find an information letter and were asked to provide informed consent and complete the survey. Invitations were sent in separate cohorts to control the participant flow throughout the study’s distinct phases. Following the original invitation, 2 reminder emails were sent (1 and 2 weeks after the first invitation). In addition, study advisers and study counselors of the UvA were informed about the study. They were asked to refer students interested in participating in the study to the research team. The study spanned 4 academic years (2018/2019 to 2021/2022), and participants were asked at the beginning of the screening if they consented to being invited again later that academic year. Participants could opt out of the study at any time, and they could also indicate that they did not want to receive further emails. During the COVID-19 pandemic (April 2020 until the end of the study), the screening e-survey contained several additional questions about their activities and coping style (for details, see the study by Koelen et al [34]).

Participants who scored above the cut-off on either anxiety symptoms, depressive symptoms, or both received an email with a web-based information brochure and informed consent document. Participants could also choose to receive the information brochure and informed consent document through the post. Likewise, informed consent could be signed on the web, returned by post, or handed in at several *collection points* at the university campus. Next, they were telephoned by the administrators of the project to book an appointment to conduct the MINI diagnostic interview (see the *Measures* section). When eligibility was confirmed, participants were immediately randomized to one of the 3 conditions. The participants who were randomized to either intervention were instructed to create an account on the intervention platform (Minddistrict).

Participants were reimbursed for completing the 6-month and 12-month follow-up assessments. In the initial phases of the trial, participants were paid €10 (US $10.8) when they completed both assessments. To increase adherence, students were paid €5 (US $5.4) for each of the assessments separately in the final stages of the trial. A raffle to win a tablet or e-reader was held among every hundred participants that completed both the 6-month and 12-month assessments.

Randomization, Blinding, and Treatment Allocation

Participants who were eligible to participate in the study were randomly assigned to either the human-guided intervention, the computer-guided intervention, or CAU (1:1:1 allocation ratio) group directly following the baseline measurement, which was done automatically through an algorithm built into the LOTUS platform. Randomization was based on computer-generated random numbers. Participants were stratified by gender and anxiety and depressive symptoms to guarantee an even distribution of male and female participants and symptom severity across conditions. Allocation was concealed from psychologists involved in this study; 2 of the psychologists who provided guidance were also administrators with access to the backend of the software. Owing to the nature of the intervention, participants and psychologists could not be blinded to the assigned treatment condition. Participants were informed about the 3 conditions and whether they were assigned to either an intervention condition or CAU. It was not specified to participants to which intervention (human-guided or computer-guided) they were assigned. Psychologists were aware of the participants' intervention condition because they provided participants with personalized feedback in the human-guided condition. The assessments were all conducted on the web and were not accessible for the psychologists (ie, blinded), except for the presession questions that were available through the intervention platform.

Interventions

The web-based transdiagnostic intervention that was used in this study, *iCare Prevent*, was originally developed by Weisel et al [35] for the German-speaking general population and translated and adapted by Bolinski et al [29] for a Dutch undergraduate student population into Dutch and English (see Textbox 1 for modules of iCare Prevent). For this study, we created a second English version of the intervention for PhD students that included a small adaptation of the examples to match their situation. The intervention is based on principles from CBT for anxiety and depression and includes web-based exercises and homework assignments. The intervention comprised 7 regular sessions (45-60 min/session) and 1 booster session (4 weeks after the completion of the last session). From the second session onward, participants were able to follow 8 additional optional modules based on their personal needs, including sleep, perfectionism, alcohol use, rumination, self-worth, acceptance, appreciation and gratitude, and rest and relaxation. In sessions 5 and 6, participants could decide to either engage in content directed at changing negative cognitions or at exposure to fear situations. They could decide to choose 1 additional module per session, and they were free to repeat this module or choose other modules in later sessions. It was advised to do at least 1 and no more than 2 treatment sessions per week. For a full description of the intervention, see the study by Karyotaki et al [10].

Textbox 1. Overview of content of the iCare Prevent training.

Sessions

1. Behavioral activation: reducing incongruence
2. Behavioral activation: overcoming difficulties and pleasant activity scheduling
3. Psychoeducation
4. Cognitive restructuring
5. Problem solving I or exposure I
6. Problem solving II or exposure II
7. Plan for the future
8. Booster session (after 4 week)

Optional modules (sessions 2-7)

1. Rumination and worries
2. Acceptance
3. Relaxation
4. Reducing alcohol
5. Self-worth
6. Perfectionism
7. Appreciation and gratefulness
8. Sleep

Guidance

During the intervention, participants in both web-based conditions received support in the form of brief standardized emails (reminders) in the chat function of the web-based environment in case they were inactive. Participants received up to 3 weekly reminders by email. Moreover, participants in both web-based conditions could use the chat or messaging function to ask technical or user-related questions (eg, problems getting web-based access).

Participants in the computer-guided condition received automatically generated feedback messages upon completion of a module with the main aim of motivating students to carry on. In contrast, in the human-guided condition, the counselors provided detailed therapeutic feedback based on the student's output of the modules. The counselors spent approximately 30 minutes on providing feedback per session. On average, participants received feedback on 3.9 (SD 2.0; range 0-12) days after they completed their session. The counselors were 6 female bachelor's-level psychologists, 5 female and 1 male master's-level psychologist, and 1 male PhD-level health care psychologist with over 10 years of prior clinical experience, who also supervised the web-based guidance and baseline intakes weekly.

CAU Condition

Participants in the CAU condition were informed about or referred to conventional care services, both internal and external to the UvA. It should be noted that information about the available services was also provided to participants in both intervention groups. However, students in the CAU group were strongly advised to seek support. Medical health services used during the trial were monitored in all groups through self-report questions after treatment (t3), at 6-month follow-up (t4), and at 12-month follow-up (t5). Participants in the CAU group were assessed at the same time points (with the same measures) as in the 2 intervention conditions.

Safety Monitoring

Ample attention was paid to warrant safety and decrease adverse effects during the trial. For this purpose, a *suicide protocol* was developed, describing in detail what the collaborating psychologists should do in case of an alert, which is available (only in Dutch) upon request (see Multimedia Appendix 1 for details about the safety measures). To monitor for sharp increases in complaints, the Patient Health Questionnaire (PHQ; PHQ-4) [36] was administered before each treatment session or weekly via email (CAU). Suicide risk was monitored using item 3 ("feeling down, depressed, or hopeless") of the PHQ-4 and item 9 ("thought that you would be better off dead, or of hurting yourself") of the Beck Depression Inventory-II [37]. All items were rated on a 0 to 3 scale. The counselors and main researchers received an automatic alert and contacted the participant by telephone if they (1) reported feeling down, depressed, or hopeless "more than half the days" or "nearly every day" (score>1) and reported having thoughts that they would be better off dead "several days" (score>0) or (2) reported having thoughts that they would be better off dead "more than half of the days" or "nearly every day" (score>1). The counselors then interviewed the participant using 6 standardized questions to rate the suicide risk and took the necessary precautions after consulting a supervisor (JK). When deemed necessary, participants were called once more, and then referred to the appropriate services. Of those allocated to 1 of the web-based intervention groups, 13.3% (71/534) had at least 1 alert and were contacted. Participants in the CAU condition were also contacted, but their data were not stored for pragmatic reasons.

Measures

Participants who fulfilled the inclusion criteria were telephoned by appointment by a trained psychologist. The MINI [33] was administered by telephone to establish *Diagnostic and Statistical Manual of Mental Disorders* (Fourth Edition) classifications with respect to mood and anxiety disorders, bipolar disorder, psychosis, and suicidal ideation. After the interview, students were briefly informed about their complaints and whether they could be included in the study.

The primary outcome measures were the PHQ (depression; PHQ-9) and the GAD-7, which both have good psychometric properties and are often used in the context of web-based interventions. The PHQ-9 [38] is a 9-item self-report questionnaire focused on depressive symptoms experienced over the past 2-week period, such as mood, sleep, and appetite. Items are rated from 0 (not at all) to 3 (nearly every day), with total scores ranging from 0 to 27. The PHQ-9 is suited for samples at risk for depression, with high specificity (0.94) and somewhat low sensitivity (0.77) in an unselected primary care sample, which is comparable to a student sample [39]. PHQ-9 scores of 5, 10, 15, and 20 represented mild, moderate, moderately severe, and severe depression, respectively [39]. The Cronbach α value in this study was .81 at baseline. The 7-item GAD [32] is a self-report questionnaire measuring anxiety symptoms (eg, "Not being able to stop or control worrying" and "Feeling nervous, anxious, or on edge"). Items are rated from 0 (not at all) to 3 (nearly every day), with total scores ranging from 0 to 21. Mild, moderate, and severe levels of anxiety are indicated by cut-off scores of 5, 10, and 15 on the GAD-7, respectively [32]. The GAD-7 questionnaire has good psychometric properties, including a good test-retest reliability (intraclass correlation coefficient=0.83) [32] and a good internal consistency (.79<Cronbach α<.91) [40]. The Cronbach α level in this study was .84 at baseline.

Secondary outcome measures for alcohol and drug use were administered at all time points. The Drug Abuse Screening Test–10 items [41] was used to screen for drug abuse over the past 12 months. The 10 items can be answered with "yes" (score=1) or "no" (score=0), and 1 item ("Are you always able to stop abusing drugs when you want to?") is reverse keyed. In the case where the first item ("Have you used drugs other than those required for medical reasons?") was answered with no (0), the other items were not administered because all remaining items are concerned with problems related to drug use. The internal consistency of this scale in this study was 0.70 at baseline. Alcohol use was measured using the abbreviated Alcohol Use Disorder Identification Test [42]. This 3-item instrument assesses the quantity and frequency of drinking and binge-drinking sessions. Items are ranked from 0 to 4; total scores range from 0 to 12. Higher scores indicate more hazardous drinking. Alcohol Use Disorder Identification Test has been validated in student populations [6]. The internal consistency of this scale in this study was 0.73 at baseline.

We examined the quality of life with the subjective health item (visual analog scale) from the EQ-5D-5L at all time points except for midtreatment [43]. Medical service use was assessed with 2 items from the Treatment Inventory of Costs in Patients with Psychiatric Disorders [44] after treatment and at 6 months and 12 months. The first item includes the frequency of contact with conventional care services (eg, general practitioner, study adviser, psychologist or psychiatrist, medical specialist), and the second item includes the use of medication. We did not use this questionnaire to calculate implicit medical costs but to compare the use of medical services across treatment groups. Client satisfaction was assessed after treatment with the Client Satisfaction Questionnaire (CSQ-8) [45]. The CSQ-8 consists of 8 items (eg, "In an overall, general sense, how satisfied are you with the web-based health support you have received?"), scored on a 1 ("quite dissatisfied") to 4 ("very satisfied") scale, with total scores ranging between 8 and 32. A higher score on the CSQ-8 indicates a higher level of satisfaction related to the intervention. The CSQ-8 is a standardized satisfaction measure reporting very good internal consistency (Cronbach α=.83 to .93) and high validity [46]. The internal consistency of this measure was 0.93 in this study after treatment.

Finally, treatment adherence was measured by tracking the activities in Minddistrict. As outcome measures, we collected the average number of sessions completed and whether participants completed all sessions [47]. For descriptive purposes, we also extracted the duration of the treatment for those completing it and the number of times they scored above the cut-offs for a "suicide trigger."

Statistical Analysis

Descriptive statistics, ANOVAs (for categorical and continuous data), and χ^2 tests (for categorical data only) were used to determine whether patient characteristics (sociodemographic and clinical) or use data (adherence) were similar across experimental conditions. To handle missing values for the main outcomes, multiple imputation was used. The Markov Chain algorithm was used to impute 50 data sets, with a maximum of 100 iterations for each imputation. This approach is particularly useful in case of high attrition (approximately 50%), as it minimizes the loss of statistical power when examining the relationship between variables [48]. On the basis of all 48 variables in our imputation model (including auxiliary variables), the (research) attrition rate was 48.7%. Note that recent studies have shown that the proportion of missing data should not be used as a guide for imputation per se, and when done responsibly, imputation for large amounts of missing data can still reduce bias [49]. Continuous data were imputed using predictive mean matching, while categorical variables were imputed using logistic regression. Various sociodemographic variables (eg, age, gender, and student status), the allocated experimental condition, and some auxiliary variables (eg, social anxiety, social phobia, and loneliness) were used as predictors only. In addition, all main outcomes of this study were used as predictors as well as values to be imputed. In the main analyses, the results were pooled across the 50 imputed data sets using the rules to account for the uncertainty introduced by the imputation process by Rubin [50]. Our a priori power calculations were based on the mixed factors 3×3 interaction difference between the 3 treatment arms and 3 time points (baseline, midtreatment, and after treatment), and taking a dropout rate of 35% into account, as well as a correlation between repeated measures of r=.50. A total sample size of 276 (92 in each group) would be needed to estimate a small within-between interaction effect [11]. With a corrected P value of .025 based on 2 primary outcomes, the total sample size required would be 369, which is amply exceeded by our sample size of 801.

Linear mixed models were used to estimate treatment effects over time. We included a fixed between-subjects effect for intervention group and a fixed within-subjects effect for time (ie, assessments at 3 or 5 sequential time points, respectively), as well as their 2-way interaction. We modeled linear, quadratic, or asymptotic effects of time by, including first or flattened second-degree polynomials created with the poly function in R statistical software (R Foundation for Statistical Computing) [51] and compared their fit. We only interpreted results from the best-fitting model. This was a convex pattern for complaints (depression, anxiety, and drug and alcohol use) and a concave pattern for health concerning the long-term outcomes, and a linear pattern for the short-term primary outcomes. As we compared 3 groups in the main analyses, we coded condition as a factor with 3 levels. We also estimated a random intercept to allow for individual variations in the intercepts for each subject. We imposed an autoregressive structure to the residual covariance matrix to correct for autoregression of repeated assessments within the same subject. It assumes that correlations between any 2 elements are equal to (Pearson) r for adjacent elements, r^2 for 2 elements separated by a third, and so on. Maximum likelihood was used as the overall estimation method. Plots were made using the Grammar of Graphics framework (*ggplot*) in R statistical software.

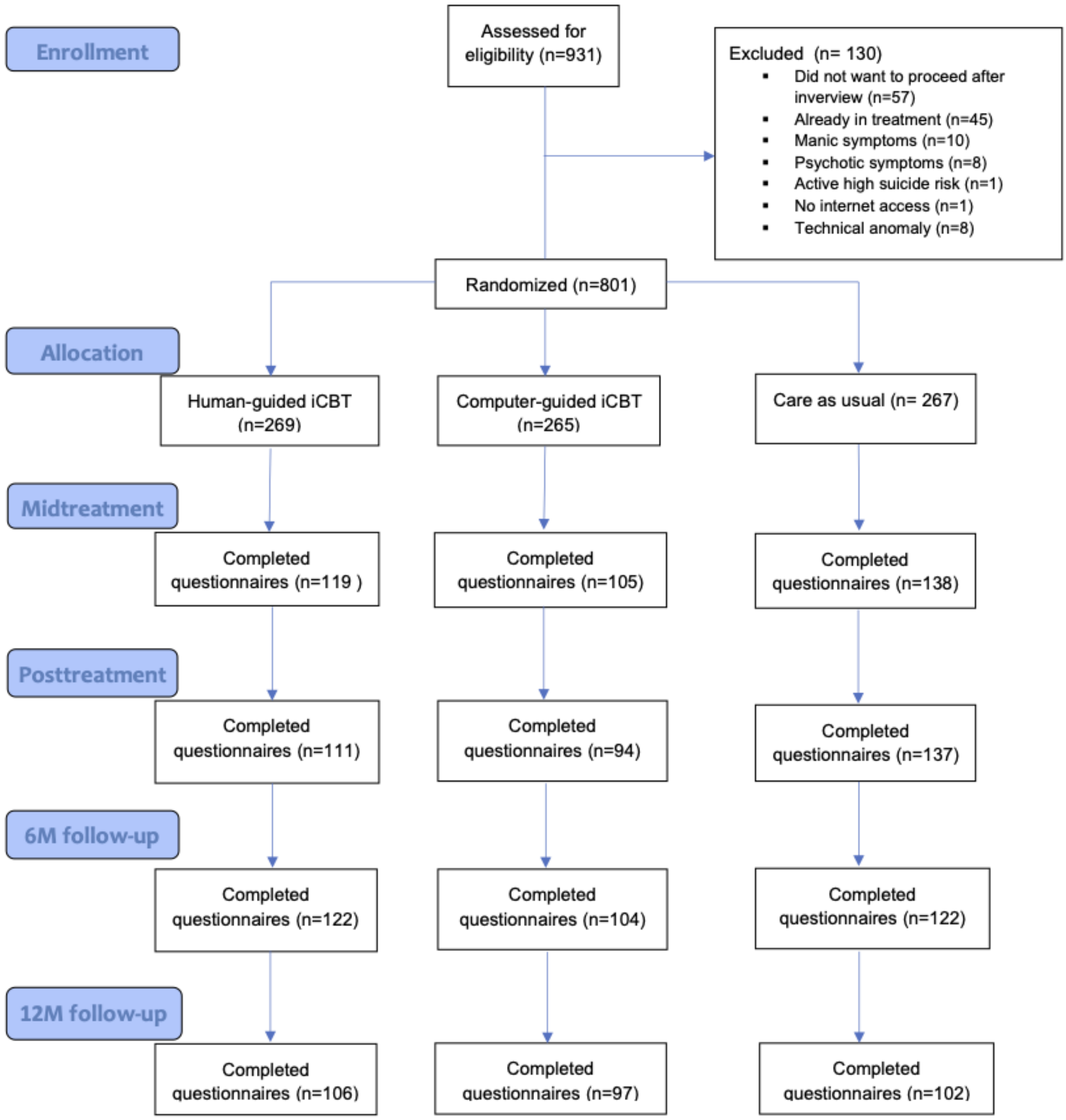
For the primary outcomes, both short-term (t3) and long-term (t4 and t5) analyses were performed based on the 3 conditions, using CAU as the reference group and subsequently computer-guided iCBT as the reference group (for pairwise comparisons). For the secondary outcomes, only long-term outcomes were analyzed (t5). Bonferroni corrections were applied to all exploratory multilevel models to control for type I error related to multiple testing, while avoiding an increased risk for type II error. Thus, based on approximately 30 tests, corrected P values ≤.002 were considered statistically significant. P values between .002 and .025 were considered marginally significant. Effect sizes were calculated for marginally significant interactions, transforming t values to Pearson r [52]. For all main analyses, SPSS (version 28.0.1.0; IBM Inc) [53] was used.

Results

Random Allocation and Characteristics of Participants

A total of 801 participants were eligible and thus randomized across conditions: 269 were allocated to the human-guided group, 265 to the computer-guided group, and 267 to CAU group (Figure 1). The average age of participants was 23.9 (SD 4.6, range 17-55) years; 71.5% (573/801) were female, and 7.5% (6/801) of the participants indicated that their gender was *Other*. The sample contained 387 (48.3%) undergraduate students (mean age 21.6, SD 3.6 y; range 17-55 y), 315 (39.3%) master's students (mean age 25.5, SD 3.9 y; range 18-53 y), and 89 (11.1%) PhD students (mean age 29.4, SD 4.8 y; range 24-51 y).

Figure 1. Flow diagram illustrating the progression of participants through the stages of the study. iCBT: internet-based cognitive behavioral therapy.



The average level of alcohol use (mean 3.72, SD 2.09) was below the average for students and far below the cut-off for harmful drinking [6]. At baseline, the average number of classifications according to the MINI was 1.79 (SD 1.74). The most common diagnosis was a current depressive episode (287/801, 35.8%), followed by generalized anxiety disorder (154/801, 19.2%). Around half of the sample had either no classification (222/801, 27.7%) or 1 potential disorder (219/801, 27.3%); the other half (360/801, 45%) had more than 1 potential disorder. For an overview of diagnostic classifications, see Table 1.

Table 1. Diagnostic and Statistical Manual of Mental Disorders (Fourth Edition) classifications based on the Mini-International Neuropsychiatric Interview administered at baseline (N=801).

	Current, n (%)	Lifetime, n (%)
Classifications		
Depression	287 (35.8)	199 (24.8)
Dysthymic disorder	36 (4.5)	N/A[a]
Manic episode	N/A[b]	11 (1.4)
Hypomanic episode	4 (0.5)	43 (5.4)
Panic disorder	36 (4.5)	92 (11.5)[c]
Agoraphobia	65 (8.1)	N/A
Social phobia	149 (18.6)	N/A
Simple phobia	17 (2.1)[d]	N/A
Obsessive-compulsive disorder	46 (5.7)[e]	N/A
Post-traumatic stress disorder	27 (3.4)[f]	N/A
Generalized anxiety disorder	154 (19.2)	N/A
Mixed anxiety-depressive disorder	85 (10.6)[g]	N/A
Psychotic syndrome	N/A[b]	16 (2)
Comorbidity		
No classification	222 (27.7)	N/A
1 disorder	219 (27.3)	N/A
2 disorders	111 (13.9)	N/A
3 disorders	114 (14.2)	N/A
>3 disorders	135 (16.9)	N/A
Suicide risk		
Low	130 (16.2)	N/A
Moderate	59 (7.4)	N/A
High	19 (2.4)	N/A
No	591 (73.8)	N/A

[a]N/A: not applicable.
[b]This was an exclusion criterion for the study.
[c]Based on a lower number of participants (n=726).
[d]Based on a lower number of participants (n=713).
[e]Based on a lower number of participants (n=770).
[f]Based on a lower number of participants (n=771).
[g]Based on a lower number of participants (n=717).

Adherence and Satisfaction

For the human-guided condition, the largest groups were composed of those doing either 0 sessions (63/268, 23.5%) or 8 sessions (48/268, 17.9%). For the computer-guided condition, the most common number of sessions completed was also 0 (81/264, 30.7%), followed by 1 (53/264, 20.1%). The average number of sessions completed in the human-guided condition was 3.32 (SD 3.02; median 2, IQR 6); in the computer-guided condition, this was 2.54 (SD 2.75; median 1, IQR 4; F_1=9.85; P=.002). The completion rate (defined as 7 or 8 sessions done) for the human-guided condition was also significantly higher than that for the computer-guided condition: 26.9% (72/268) versus 15.5% (41/264) cases completed treatment (n=532; χ^2_1=10.2; P=.001). Finally, participants in both the human-guided and computer-guided conditions were significantly more satisfied after treatment compared with those in the CAU condition (F_2=45.2; P<.001).

Optional Modules and Psychological Care

Only those participants who completed more than 1 session were enabled to choose optional modules (289/534, 54.1%). Among those participants, a total of 28.5% (83/291) completed

1 optional module, 21.6% (63/291) completed 2 modules, and 49.8% (145/291) completed 3 to 6 modules (median 2, IQR 3). There were no significant differences between the iCBT conditions in terms of optional modules made (F_1=0.74; P=.39). Of those enabled to choose an optional module, 13.7% (40/291) of participants chose the optional module for alcohol use-related problems. Asked retrospectively at 6 months, participants in the CAU condition reported to have visited a psychologist or psychiatrist more often (65/265, 24.5%) during the 6 months prior than the participants in the human-guided (38/269, 14.1%) or computer-guided (49/267, 18.3%) conditions (n=795; χ^2_2=7.96; P=.02).

Efficacy Over Time

For the course of symptoms over time across the 3 conditions, see Figure 2 (depression) and Figure 3 (anxiety). For the estimated marginal means of the 2 primary outcome measures across treatment groups at each assessment point, see Tables 2 and 3.

Figure 2. Mean depression over time by condition. PHQ-9: Patient Health Questionnaire–9 Note. Vertical lines represent error bars (with a 95% confidence interval); non-overlapping error bars indicate that the true means are likely to be different from each other.

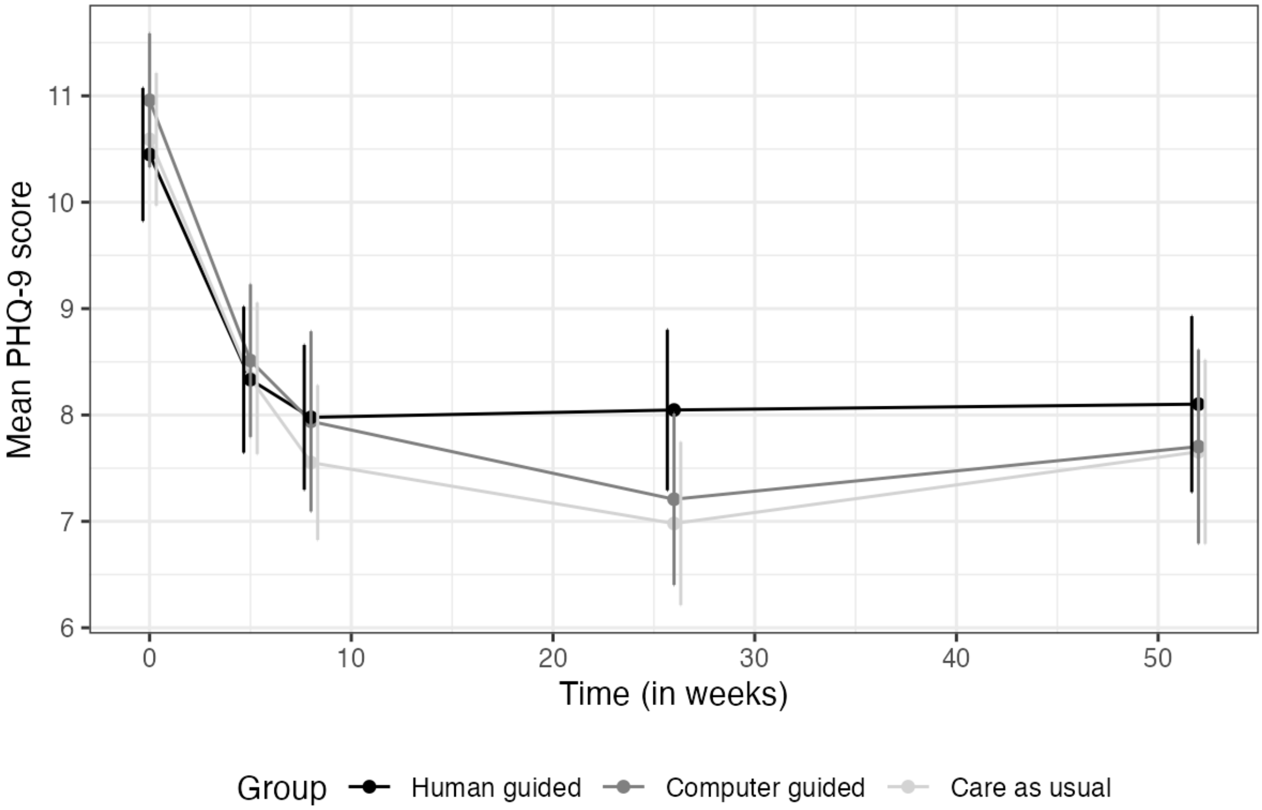


Figure 3. Mean anxiety over time by condition. GAD-7: Generalized Anxiety Disorder scale–7 items. Note. Vertical lines represent error bars (with a 95% confidence interval); non-overlapping error bars indicate that the true means are likely to be different from each other.

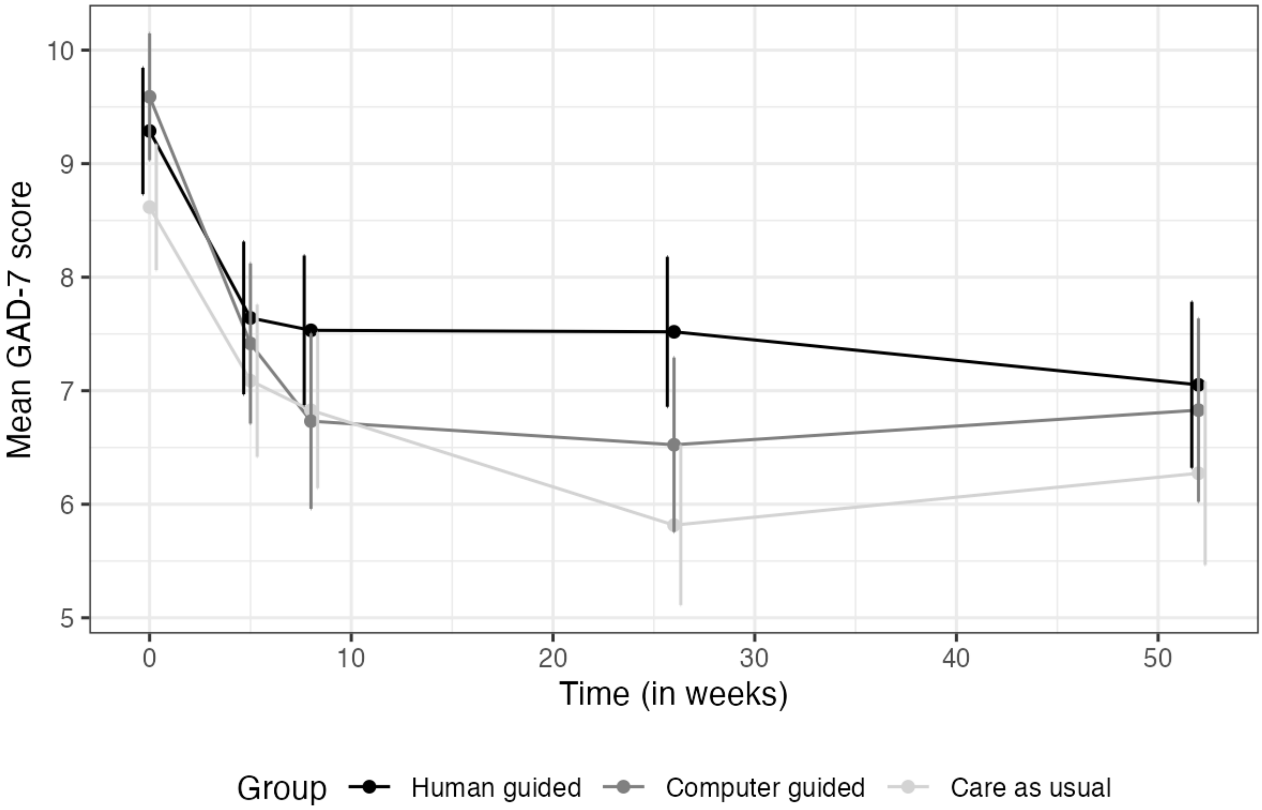


Table 2. Estimated marginal means and SEs for the primary outcome measures at all assessment points for the 3 conditions—depression (Patient Health Questionnaire–9).

Condition	Time				
	Baseline, mean (SE)	Midtreatment, mean (SE)	After treatment, mean (SE)	6 months, mean (SE)	12 months, mean (SE)
Human guided	10.591 (0.318)	8.345 (0.363)	7.553 (0.371)	6.980 (0.391)	7.654 (0.441)
Computer guided	10.958 (0.321)	8.511 (0.366)	7.940 (0.432)	7.208 (0.411)	7.703 (0.465)
Care as usual	10.449 (0.319)	8.332 (0.350)	7.978 (0.347)	8.047 (0.385)	8.102 (0.422)

Table 3. Estimated marginal means and SEs for the primary outcome measures at all assessment points for the 3 conditions—anxiety (Generalized Anxiety Disorder scale–7 items).

Condition	Time				
	Baseline, mean (SE)	Midtreatment, mean (SE)	After treatment, mean (SE)	6 months, mean (SE)	12 months, mean (SE)
Human guided	8.617 (0.283)	7.088 (0.342)	6.828 (0.350)	5.816 (0.359)	6.273 (0.412)
Computer guided	9.589 (0.285)	7.416 (0.360)	6.732 (0.394)	6.524 (0.392)	6.829 (0.412)
Care as usual	9.288 (0.284)	7.641 (0.343)	7.532 (0.336)	7.519 (0.337)	7.052 (0.373)

Short-Term Efficacy (t3)

According to our preregistered protocol, we analyzed the efficacy over time until after treatment (t3) for the 2 primary outcome measures. For depression, no significant time×group interaction was observed when comparing both the human-guided (B=.08, 95% CI –.05 to .21; *P*=.25) and the computer-guided (B=.06, 95% CI –.09 to .21; *P*=.43) conditions to CAU. For the direct comparison between human-guided and computer-guided iCBT, no significant time by treatment interaction was found (B=.02, 95% CI –.14 to .17; *P*=.86).

For anxiety, none of the interventions differed significantly from CAU over time: *human-guided versus CAU* (B=.01, 95% CI −.11 to .13; *P*=.89) and *computer-guided iCBT versus CAU* (B=.13, 95% CI –.003 to .27; *P*=.05). No significant differences were found when comparing human-guided and computer-guided iCBT directly with each other (B=−.12, 95% CI –.26 to .01; *P*=.08).

Long-Term Efficacy for the Primary Outcomes (t4)

Next, we analyzed the primary outcomes until t4 (6 months) follow-up. Over 6 months, no significant time by condition interactions were found for either depression or anxiety, indicating that the slopes for the 3 conditions were similar. Some comparisons were marginally significant (see Tables 4 and 5 for coefficients). Next, we added the covariate of "psychological care" to control for this higher level of care in the CAU condition. For both depression and anxiety, results remained marginally significant for both intervention conditions compared with CAU yet did not reach the corrected *P* value of .002. For depression, human-guided iCBT was marginally more efficacious than CAU (B=3.21, 95% CI 0.54-5.89; *P*=.02; *r*=0.08). The computer-guided condition was also marginally more efficacious compared with CAU (B=3.33, 95% CI 0.61-6.05; *P*=.02; *r*=0.08). For anxiety, similar results were obtained: *human-guided* iCBT *versus CAU* (B=3.08, 95% CI 0.77-5.40; *P*=.01; *r*=0.09) and *computer-guided iCBT versus CAU* (B=3.12, 95% CI 0.62-5.62; *P*=.02; *r*=0.08).

Table 4. Long-term (6-month) results of linear mixed model for pairwise comparisons of conditions (depression).

Fixed effects	Comparison					
	Human guided vs care as usual		Computer guided vs care as usual		Human guided vs computer guided	
	B (SE)	*P* value	B (SE)	*P* value	B (SE)	*P* value
Intercept	*8.22* [a] *(0.29)*	*<.001*	*8.22 (0.29)*	*<.001*	*7.66 (0.30)*	*<.001*
Time	*4.49 (0.97)*	*<.001*	*4.49 (0.97)*	*<.001*	*7.39 (0.98)*	*<.001*
Condition	−0.56 (0.41)	.18	−0.30 (0.44)	.50	−0.26 (0.43)	.55
Time×condition	2.91 (1.36)	.03	3.15 (1.38)	.02	−0.25 (1.39)	.86

[a]Italicized values are statistically significant (*P*<.05).

Table 5. Long-term (6-month) results of linear mixed model for pairwise comparisons of conditions (anxiety).

Fixed effects	Comparison					
	Human guided vs care as usual		Computer guided vs care as usual		Human guided vs computer guided	
	B (SE)	*P* value	B (SE)	*P* value	B (SE)	*P* value
Intercept	*7.72* [a] *(0.25)*	*<.001*	*7.72 (0.25)*	*<.001*	*6.59 (0.27)*	*<.001*
Time	*3.40 (0.83)*	*<.001*	*3.40 (0.83)*	*<.001*	*6.01 (0.88)*	*<.001*
Condition	−1.13 (0.36)	.002	−0.65 (0.41)	.11	−0.48 (0.40)	.23
Time×condition	2.61 (1.15)	.02	2.84 (1.27)	.03	−0.24 (1.29)	.85

[a]Italicized values are statistically significant (*P*<.05).

Long-Term Efficacy for the Primary and Secondary Outcomes (t5)

All material related to the outcomes at t5 can be found in Multimedia Appendix 2 for the primary outcomes and Multimedia Appendix 3 for the secondary outcomes. In summary, none of the interactions between time and treatment conditions were significant. However, time effects for depression, anxiety, and drug use were highly significant, indicating that all groups improved over time (*P*<.001). The time effects for alcohol use and subjective health were not significant (*P*=.08 and .30, respectively). Figures for the secondary measures are shown in Figures 4-6.

Figure 4. Alcohol use over time. AUDIT-C: Alcohol Use Disorder Identification Test alcohol consumption questions. Note. Vertical lines represent error bars (with a 95% confidence interval); non-overlapping error bars indicate that the true means are likely to be different from each other.

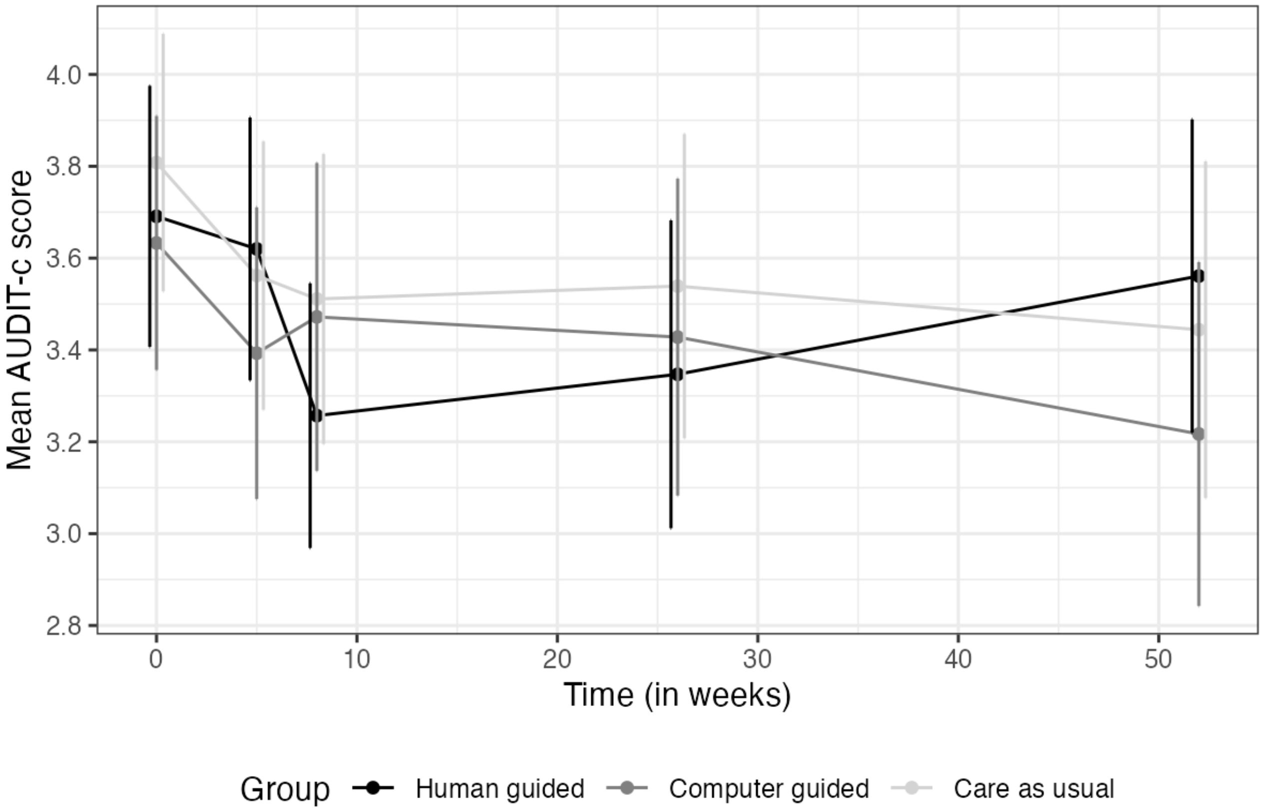


Figure 5. Drug use over time. DAST-10: Drug Abuse Screening Test–10 items. Note. Vertical lines represent error bars (with a 95% confidence interval); non-overlapping error bars indicate that the true means are likely to be different from each other.

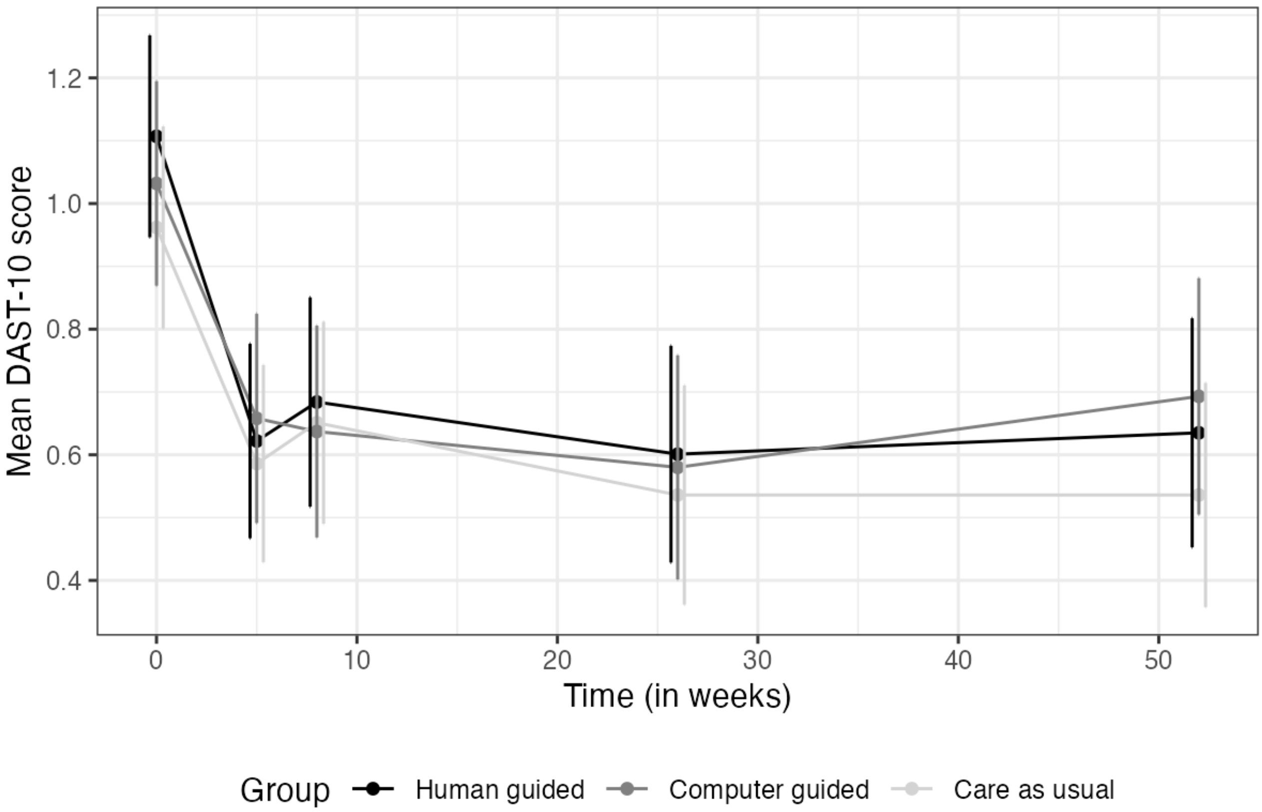
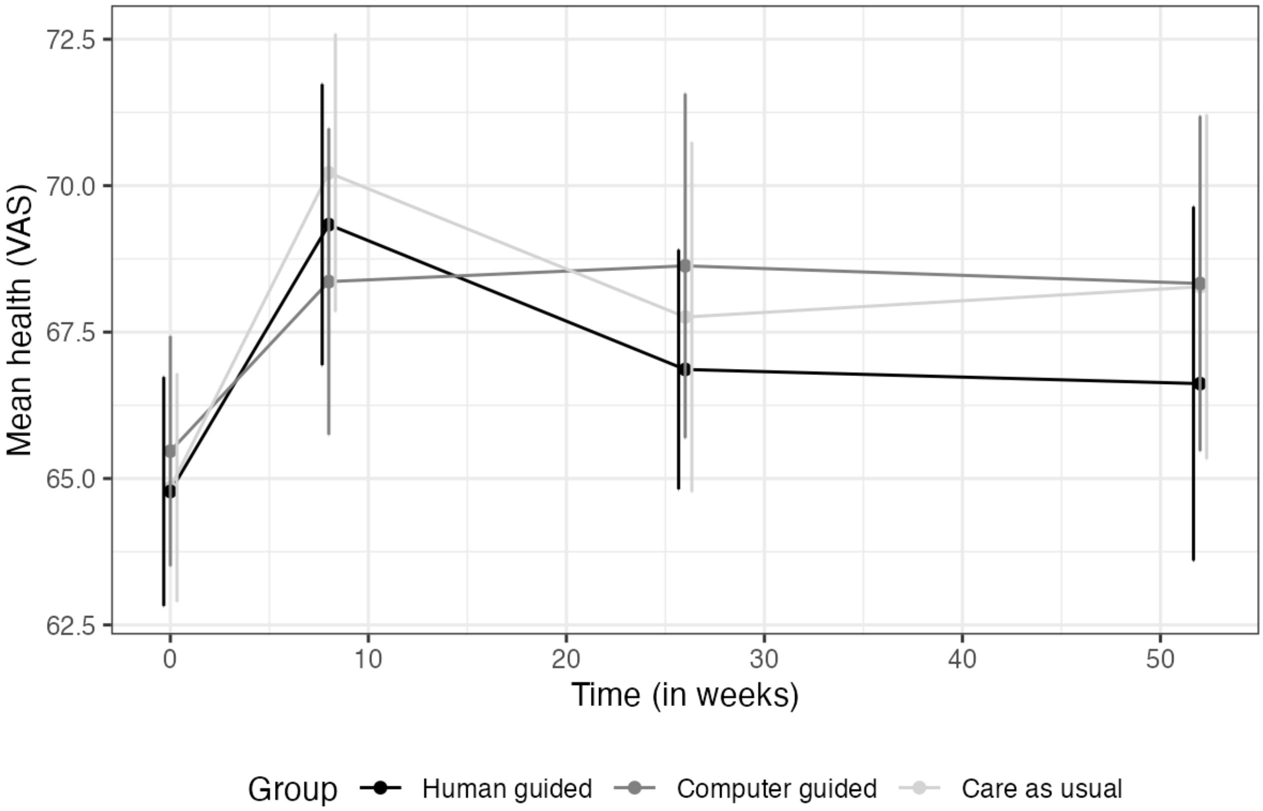


Figure 6. Subjective health over time. VAS: visual analog scale. Note. Vertical lines represent error bars (with a 95% confidence interval); non-overlapping error bars indicate that the true means are likely to be different from each other.



Sensitivity Analyses (t4)

To add further validity to our results, we included 2 additional sensitivity analyses (see Multimedia Appendix 4 for details). First, we reran analyses on the groups that did at least 1 session (655/801, 81.8%), excluding those who did 0 sessions. The intent of this analysis was to assess whether those with at least some adherence to treatment, regardless of completed assessments, had a similar outcome compared to the whole group (this is an imputed data set). For both depression and anxiety, human-guided and computer-guided iCBT were marginally more efficacious than CAU (.02<*P*<.05; 0.07<*r*<0.09).

Second, we reran the analyses for the original data set without imputation. This latter data set contained all subjects, yet most of them had at least 1 missing value which was not corrected for. With this analysis, we examined a potential difference between the imputed and the unimputed data set, which may indicate a bias. For both depression and anxiety, human-guided and computer-guided iCBT were significantly more efficacious than CAU (*P*<.001; 0.11<*r*<0.16).

Discussion

Principal Findings

In this study, we compared 2 types of web-based interventions with CAU, both regarding the short and long-term outcomes in a large sample of university students with elevated levels of depression and anxiety symptoms. Overall, reductions in anxiety and depression were seen in all 3 conditions over time, yet no significant differences between conditions were observed over time.

In the CAU group, approximately one-quarter (65/265, 24.5%) of participants sought care from a mental health care professional during the intervention period. When controlling for this psychological care in the CAU condition, both iCBT conditions were marginally more efficacious compared with CAU across the 6-month period. This implies that iCBT is a viable alternative for CAU in a large educational setting. It seems that most students found the help they needed, or at least felt reassured to know that help was available.

No significant differences in symptoms were found in the direct comparison between the human-guided intervention and computer-guided intervention. However, as expected, the difference between human-guided and computer-guided was present for the adherence measures; students in the human-guided condition completed more sessions and subsequently did more optional modules that focused on specific problem areas. It could be that human guidance encouraged students to persist with the intervention. However, higher adherence to the human-guided intervention was not necessarily associated with larger benefits, as indicated by our analyses of symptoms. This could be due to the small yet significant difference in adherence between the conditions (approximately 1 session), which may be too small to yield an observable and clinically relevant difference. Future studies could test the hypothesis whether human guidance indirectly impacts symptom outcomes through its effect on adherence (ie, mediation analysis).

Comparison to Previous Studies

In line with Karyotaki et al [10], both web-based interventions did not outperform the CAU condition. Our finding replicates and extends this previous finding, as our study included more students and students with higher levels of anxiety and depression. Some studies suggested that therapist-guided iCBT is more efficacious for individuals with elevated levels of anxiety or depression [25,27], but this assumption was not supported by our findings. The fact that the interventions were not more efficacious than CAU could have several reasons. First, it could be considered that our pattern of findings reflects regression to the mean, which is typical in longitudinal studies in which participants are selected on the basis of elevated scores at baseline [54]. Any potential effect of the interventions may have been overshadowed by the effect of this statistical artifact, which would have been present in all conditions. Second, although comparable to other web-based intervention studies, the adherence rates in this study were rather low, resulting in a potential loss of efficacy. Only a quarter (72/268, 26.9%) of the participants completed the human-guided intervention. In the computer-guided condition, only half (130/264, 49.2%) of the participants did ≥2 sessions. The attrition rates are generally high in internet interventions (especially in so-called unguided formats), and this should be improved in future trials [28]. A third explanation could be that students who do not seek help on their own initiative may have a diminished treatment response, due to factors of decreased intrinsic motivation, low awareness of their illness burden, and unrealistic expectations about the time investment required.

The lack of any significant differences between human-guided and computer-guided conditions (*P* values>.025 as per preregistration) is not in line with previous findings from a meta-analysis showing that human guidance is slightly yet significantly more efficacious than *technological guidance* [28], at least in the short run, and begs further explanation. It could be a result of the fact that the intensity of guidance in both conditions was quite similar. Although students in the human-guided condition received weekly personalized feedback from psychologists, students in the computer-guided condition received a considerable amount of support as well, including automated feedback, regular reminders sent by a psychologist, personalized responses to queries in the chat, and phone calls in case of suicide alerts. Of those who were actively engaged on the eHealth platform, around 13.3% (71/534) received such phone calls. This may have led to the impression among students in the computer-guided condition that they did receive personalized support and that they perceived a sense of *therapeutic presence*, that is, someone in the outside world that had their concerns in mind. An alternative reason could be that psychologists in the human-guided condition, although they received supervision, were not specifically trained for this intervention and they were not delivering therapy in a strict sense, but rather some form of support, which may have been experienced as disappointing for some students, who expected more e-therapy, as qualitative interviews after the trial suggested (JA Koelen, unpublished data, April 2024). Third, students in the computer-guided condition received immediate, automatized feedback, whereas human guidance was delivered after relatively long intervals between sessions, which took an average of 3.9 (SD 2.0) days. Further research should aim to clarify to what extent these factors contribute to varieties in effectiveness. A final explanation could be that the differences between *guided* and *unguided* web-based interventions (often iCBT) reported in the literature refer to short-term outcomes, as emerging evidence indicates that the differences may be short-lived [27].

The lack of a significant difference between the human-guided and computer-guided conditions also raises questions concerning the working mechanisms of web-based therapy [55]. As with regular therapy, it is still unclear whether web-based therapy works through mechanisms common to all therapies (eg, the therapeutic alliance, the mobilization of hope), or those specific to certain types of therapies (eg, cognitive restructuring in CBT), or both. Notably, technical components of iCBT (ie, specific mechanisms) were present in both conditions, and the nonexistent difference between conditions would plea for the effectiveness of such components. Previous qualitative analyses of emails from web-based counselors providing iCBT suggested that interventions that were regarded as most helpful (and were associated with better outcomes) were best classified as *common factors* [56,57]. These interventions included validation of completed exercises, anticipation of pending assignments (ie, changing expectations of personal effectiveness), as well as empathy and self-disclosure of web-based counselors. Some of these interventions were also provided with the automated feedback in the computer-guided condition, which could also explain the absence of significant differences between conditions.

Strengths and Limitations

A major strength of this trial, first, is that we compared a human-guided and a computer-guided form of iCBT head-to-head, and to an active control group. Second, our sample was relatively large, enabling us to detect small interaction effects. Third, we included students with elevated levels of symptoms, with nearly three quarters of the students (579/801, 72.3%) presenting with at least one potential mental disorder, which makes this a representative sample of students in terms of burden of disease [3]. Finally, we assessed the long-term outcome of our interventions.

Several limitations need to be considered when interpreting the results of this study. First, nearly all participants were included during the COVID-19 pandemic, which reached its first peak in the Netherlands in May 2020. More precisely, only 1.9% (15/801) of the participants completed the entire study before the onset of the COVID-19 pandemic. Intervention effects may have been distorted due to this crisis, for instance, due to an increase in symptoms during the trial [34]. Second, the number of PhD students included in this study was relatively low, which needs to be considered when extrapolating our findings to US student samples. This small number is mainly attributable to the clear distinction between master’s and PhD students that is made in the Netherlands, where a Master of Science degree is the typical end point of study and only few students are selected to do a PhD trajectory. Third, as expressed by some students, the modules were *text heavy* and lacked interactive functionalities, which may have diminished its attractiveness, so this could be improved. Yet, the intervention had been adapted especially for students, and the text was enriched by elements of persuasive design, which are known to increase adherence, such as creating an overview, use of metaphors, and creating a sense of control and purpose [58]. Fourth, in hindsight, we might have recruited more psychologists for guidance, which would have resulted in swifter feedback, although it is debatable whether swifter feedback results in greater efficacy [59]. Fifth, the relatively large amount of missing data could have introduced a bias in our data. However, we limited that risk of bias by using the Markov Chain algorithm, which is adept at capturing data dependencies, and maintaining data set integrity [60]. A recent meta-analysis [28] of web-based interventions found an average of 48% attrition for human-guided interventions and an even higher percentage (51%) for those using technological guidance only. These figures are comparable to those reported in this study. Nevertheless, a major interest of future studies should be to decrease dropout. Sixth, future research could aim to clarify the optimal length and intensity for web-based treatments, as we followed a classic format with weekly session. Some studies of regular (face-to-face) therapy have shown, for example, that a more concentrated approach (ie, an equal number of sessions with increasing frequency) can bring about more optimal outcomes [61]. Seventh, some studies have shown that rewarding students for completing questionnaires, as was done in this trial, could lead to an underestimation of intervention effects [20]. Eighth, we could not establish the effect of sessions made irrespective of the intervention type because the number of sessions completed was contingent upon the allocated condition. This requires another type of experimental research. Ninth, future studies should pay more attention to individual differences and moderators of outcome. Finally, we assessed the outcome only in terms of symptoms, and not in terms of underlying vulnerabilities (eg, elevated levels of negative affect), which would have suited a transdiagnostic approach.

Conclusions

In conclusion, the results of this study indicated both iCBT interventions were not inferior to CAU (which did include the possibility to visit other mental health care professionals, who had a limited capacity). Moreover, we found that 1 type of iCBT (human-guided) was not more efficacious than the other (computer-guided). We conclude that iCBT offers a feasible and viable alternative for the regular health services in a university setting, which are often hard to access for students. Future studies should also examine cost-effectiveness to determine whether the additional investment in web-based psychologists is worthwhile.

Acknowledgments

The authors would like to thank Ruben Atteveld, Matilda Nottage Chantalle Roetink, Dieuwke Vanwersch, Pieta Vanwersch, Lotte van der Vegt, Sabine de Vet, and Anne Vonk for their contributions as web-based counselors. This study was funded by the University of Amsterdam (UvA). Open access publication fees were paid for by the UvA Student Physicians. We are indebted to Peter ten Klooster at the University of Twente for his invaluable support with the data imputation

JK was affiliated with the University of Amsterdam, Amsterdam, Netherlands at the time of the trial and is currently affiliated with the University of Twente, Enschede, Netherlands.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Safety measures.
[PDF File (Adobe PDF File), 95 KB - mental_v11i1e50503_app1.pdf]

Multimedia Appendix 2
Long-term results (t5) for the primary outcomes.
[PDF File (Adobe PDF File), 225 KB - mental_v11i1e50503_app2.pdf]

Multimedia Appendix 3
Long-term results (t5) for the secondary outcomes.
[PDF File (Adobe PDF File), 305 KB - mental_v11i1e50503_app3.pdf]

Multimedia Appendix 4
Sensitivity analyses.
[PDF File (Adobe PDF File), 209 KB - mental_v11i1e50503_app4.pdf]

Multimedia Appendix 5

CONSORT checklist.
[PDF File (Adobe PDF File), 1192 KB - mental_v11i1e50503_app5.pdf]

References

1. Karyotaki E, Cuijpers P, Albor Y, Alonso J, Auerbach RP, Bantjes J, et al. Sources of stress and their associations with mental disorders among college students: results of the World Health Organization World Mental Health Surveys International College Student Initiative. Front Psychol 2020;11:1759 [FREE Full text] [doi: 10.3389/fpsyg.2020.01759] [Medline: 32849042]
2. Branje S, de Moor EL, Spitzer J, Becht AI. Dynamics of identity development in adolescence: a decade in review. J Res Adolesc 2021 Dec;31(4):908-927 [FREE Full text] [doi: 10.1111/jora.12678] [Medline: 34820948]
3. Auerbach RP, Mortier P, Bruffaerts R, Alonso J, Benjet C, Cuijpers P, et al. WHO World Mental Health Surveys International College Student Project: prevalence and distribution of mental disorders. J Abnorm Psychol 2018 Oct;127(7):623-638 [FREE Full text] [doi: 10.1037/abn0000362] [Medline: 30211576]
4. Auerbach RP, Alonso J, Axinn WG, Cuijpers P, Ebert DD, Green JG, et al. Mental disorders among college students in the World Health Organization World Mental Health Surveys. Psychol Med 2016 Oct;46(14):2955-2970 [FREE Full text] [doi: 10.1017/S0033291716001665] [Medline: 27484622]
5. Auerbach RP, Mortier P, Bruffaerts R, Alonso J, Benjet C, Cuijpers P, et al. Mental disorder comorbidity and suicidal thoughts and behaviors in the World Health Organization World Mental Health Surveys International College Student Initiative. Int J Methods Psychiatr Res 2019 Jun;28(2):e1752 [FREE Full text] [doi: 10.1002/mpr.1752] [Medline: 30450753]
6. Verhoog S, Dopmeijer JM, de Jonge JM, van der Heijde CM, Vonk P, Bovens RH, et al. The use of the Alcohol Use Disorders Identification Test - consumption as an indicator of hazardous alcohol use among university students. Eur Addict Res 2020 Sep 27;26(1):1-9 [FREE Full text] [doi: 10.1159/000503342] [Medline: 31563902]
7. Alonso J, Mortier P, Auerbach RP, Bruffaerts R, Vilagut G, Cuijpers P, WHO WMH-ICS Collaborators. Severe role impairment associated with mental disorders: results of the WHO World Mental Health Surveys International College Student Project. Depress Anxiety 2018 Sep;35(9):802-814 [FREE Full text] [doi: 10.1002/da.22778] [Medline: 29847006]
8. Wilks CR, Auerbach RP, Alonso J, Benjet C, Bruffaerts R, Cuijpers P, et al. The importance of physical and mental health in explaining health-related academic role impairment among college students. J Psychiatr Res 2020 Apr;123:54-61 [FREE Full text] [doi: 10.1016/j.jpsychires.2020.01.009] [Medline: 32036074]
9. Bolinski F, Boumparis N, Kleiboer A, Cuijpers P, Ebert DD, Riper H. The effect of e-mental health interventions on academic performance in university and college students: a meta-analysis of randomized controlled trials. Internet Interv 2020 Apr;20:100321 [FREE Full text] [doi: 10.1016/j.invent.2020.100321] [Medline: 32382515]
10. Karyotaki E, Klein AM, Ciharova M, Bolinski F, Krijnen L, de Koning L, et al. Guided internet-based transdiagnostic individually tailored cognitive behavioral therapy for symptoms of depression and/or anxiety in college students: a randomized controlled trial. Behav Res Ther 2022 Mar;150:104028 [FREE Full text] [doi: 10.1016/j.brat.2021.104028] [Medline: 35066365]
11. Klein A, Wolters NE, Bol EJ, Koelen J, de Koning L, Roetink SS, et al. Online computer or therapist-guided cognitive behavioral therapy in university students with anxiety and/or depression: study protocol of a randomised controlled trial. BMJ Open 2021 Nov 26;11(11):e049554 [FREE Full text] [doi: 10.1136/bmjopen-2021-049554] [Medline: 34836897]
12. Bruffaerts R, Mortier P, Auerbach RP, Alonso J, Hermosillo De la Torre AE, Cuijpers P, et al. Lifetime and 12-month treatment for mental disorders and suicidal thoughts and behaviors among first year college students. Int J Methods Psychiatr Res 2019 Jun 20;28(2):e1764 [FREE Full text] [doi: 10.1002/mpr.1764] [Medline: 30663193]
13. Beatty L, Binnion C. A systematic review of predictors of, and reasons for, adherence to online psychological interventions. Int J Behav Med 2016 Dec;23(6):776-794 [FREE Full text] [doi: 10.1007/s12529-016-9556-9] [Medline: 26957109]
14. Boydell KM, Hodgins M, Pignatiello A, Teshima J, Edwards H, Willis D. Using technology to deliver mental health services to children and youth: a scoping review. J Can Acad Child Adolesc Psychiatry 2014 May;23(2):87-99 [FREE Full text] [Medline: 24872824]
15. Czyz EK, Horwitz AG, Eisenberg D, Kramer A, King CA. Self-reported barriers to professional help seeking among college students at elevated risk for suicide. J Am Coll Health 2013;61(7):398-406 [FREE Full text] [doi: 10.1080/07448481.2013.820731] [Medline: 24010494]
16. Dopmeijer JM, de Jonge JM, Visscher TL, Bovens RH, Wiers RW. Predicting disclosure and help-seeking in university students with psychosocial problems based on stigma and attitudes towards disclosure and help-seeking. Ment Health Addict Res 2020;5(2):1-8. [doi: 10.15761/mhar.1000193]
17. Eisenberg D, Hunt J, Speer N, Zivin K. Mental health service utilization among college students in the United States. J Nerv Ment Dis 2011 May;199(5):301-308. [doi: 10.1097/NMD.0b013e3182175123] [Medline: 21543948]
18. Cuijpers P, Miguel C, Ciharova M, Aalten P, Batelaan N, Salemink E, et al. Prevention and treatment of mental health and psychosocial problems in college students: an umbrella review of meta-analyses. Clin Psychol Sci Pract 2021 Sep;28(3):229-244. [doi: 10.1037/cps0000030]

19. Ebert DD, Zarski AC, Christensen H, Stikkelbroek Y, Cuijpers P, Berking M, et al. Internet and computer-based cognitive behavioral therapy for anxiety and depression in youth: a meta-analysis of randomized controlled outcome trials. PLoS One 2015;10(3):e0119895 [FREE Full text] [doi: 10.1371/journal.pone.0119895] [Medline: 25786025]
20. Harrer M, Adam SH, Baumeister H, Cuijpers P, Karyotaki E, Auerbach RP, et al. Internet interventions for mental health in university students: a systematic review and meta-analysis. Int J Methods Psychiatr Res 2019 Jun;28(2):e1759 [FREE Full text] [doi: 10.1002/mpr.1759] [Medline: 30585363]
21. Karyotaki E, Ebert DD, Donkin L, Riper H, Twisk J, Burger S, et al. Do guided internet-based interventions result in clinically relevant changes for patients with depression? an individual participant data meta-analysis. Clin Psychol Rev 2018 Jul;63:80-92 [FREE Full text] [doi: 10.1016/j.cpr.2018.06.007] [Medline: 29940401]
22. Craske MG. Transdiagnostic treatment for anxiety and depression. Depress Anxiety 2012 Sep;29(9):749-753 [FREE Full text] [doi: 10.1002/da.21992] [Medline: 22949272]
23. Newby JM, McKinnon A, Kuyken W, Gilbody S, Dalgleish T. Systematic review and meta-analysis of transdiagnostic psychological treatments for anxiety and depressive disorders in adulthood. Clin Psychol Rev 2015 Aug;40:91-110 [FREE Full text] [doi: 10.1016/j.cpr.2015.06.002] [Medline: 26094079]
24. Day V, McGrath PJ, Wojtowicz M. Internet-based guided self-help for university students with anxiety, depression and stress: a randomized controlled clinical trial. Behav Res Ther 2013 Jul;51(7):344-351 [FREE Full text] [doi: 10.1016/j.brat.2013.03.003] [Medline: 23639300]
25. Mullin A, Dear BF, Karin E, Wootton BM, Staples LG, Johnston L, et al. The UniWellbeing course: a randomised controlled trial of a transdiagnostic internet-delivered cognitive behavioural therapy (CBT) programme for university students with symptoms of anxiety and depression. Internet Interv 2015 May;2(2):128-136 [FREE Full text] [doi: 10.1016/j.invent.2015.02.002]
26. Păsărelu CR, Andersson G, Bergman Nordgren L, Dobrean A. Internet-delivered transdiagnostic and tailored cognitive behavioral therapy for anxiety and depression: a systematic review and meta-analysis of randomized controlled trials. Cogn Behav Ther 2017 Jan;46(1):1-28 [FREE Full text] [doi: 10.1080/16506073.2016.1231219] [Medline: 27712544]
27. Karyotaki E, Efthimiou O, Miguel C, Bermpohl FM, Furukawa TA, Cuijpers P, Individual Patient Data Meta-Analyses for Depression (IPDMA-DE) Collaboration, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. JAMA Psychiatry 2021 Apr 01;78(4):361-371. [doi: 10.1001/jamapsychiatry.2020.4364] [Medline: 33471111]
28. Koelen JA, Vonk A, Klein A, de Koning L, Vonk P, de Vet S, et al. Man vs. machine: a meta-analysis on the added value of human support in text-based internet treatments ("e-therapy") for mental disorders. Clin Psychol Rev 2022 Aug;96:102179 [FREE Full text] [doi: 10.1016/j.cpr.2022.102179] [Medline: 35763975]
29. Bolinski F, Kleiboer A, Karyotaki E, Bosmans J, Zarski AC, Weisel KK, et al. Effectiveness of a transdiagnostic individually tailored internet-based and mobile-supported intervention for the indicated prevention of depression and anxiety (ICare Prevent) in Dutch college students: study protocol for a randomised controlled trial. Trials 2018 Feb 20;19(1):118 [FREE Full text] [doi: 10.1186/s13063-018-2477-y] [Medline: 29458407]
30. Van der Heijde CM, Vonk P, Meijman FJ. Self-regulation for the promotion of student health. Traffic lights: the development of a tailored web-based instrument providing immediate personalized feedback. Health Psychol Behav Med 2015 Jul 13;3(1):169-189. [doi: 10.1080/21642850.2015.1049950]
31. Beekman AT, Deeg DJ, Van Limbeek J, Braam AW, De Vries MZ, Van Tilburg W. Criterion validity of the center for epidemiologic studies depression scale (CES-D): results from a community-based sample of older subjects in the Netherlands. Psychol Med 1997 Jan 01;27(1):231-235. [doi: 10.1017/s0033291796003510] [Medline: 9122304]
32. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med 2006 May 22;166(10):1092-1097. [doi: 10.1001/archinte.166.10.1092] [Medline: 16717171]
33. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 1998;59 Suppl 20:22-34. [Medline: 9881538]
34. Koelen JA, Mansueto AC, Finnemann A, de Koning L, van der Heijde CM, Vonk P, et al. COVID-19 and mental health among at-risk university students: a prospective study into risk and protective factors. Int J Methods Psychiatr Res 2022 Mar;31(1):e1901 [FREE Full text] [doi: 10.1002/mpr.1901] [Medline: 34932250]
35. Weisel KK, Zarski AC, Berger T, Schaub MP, Krieger T, Moser CT, et al. Transdiagnostic tailored internet- and mobile-based guided treatment for major depressive disorder and comorbid anxiety: study protocol of a randomized controlled trial. Front Psychiatry 2018;9:274 [FREE Full text] [doi: 10.3389/fpsyt.2018.00274] [Medline: 30022954]
36. Kroenke K, Spitzer RL, Williams JB, Löwe B. An ultra-brief screening scale for anxiety and depression: the PHQ-4. Psychosomatics 2009;50(6):613-621. [doi: 10.1176/appi.psy.50.6.613] [Medline: 19996233]
37. Beck AT, Steer RA, Brown GK. Beck Depression Inventory BDI–II. London, UK: Pearson Prentice Hall; 1996.
38. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. J Gen Intern Med 2001 Sep;16(9):606-613 [FREE Full text] [doi: 10.1046/j.1525-1497.2001.016009606.x] [Medline: 11556941]

39. Wittkampf KA, Naeije L, Schene AH, Huyser J, van Weert HC. Diagnostic accuracy of the mood module of the patient health questionnaire: a systematic review. Gen Hosp Psychiatry 2007 Sep;29(5):388-395. [doi: 10.1016/j.genhosppsych.2007.06.004] [Medline: 17888804]
40. Dear BF, Titov N, Sunderland M, McMillan D, Anderson T, Lorian C, et al. Psychometric comparison of the generalized anxiety disorder scale-7 and the Penn State Worry Questionnaire for measuring response during treatment of generalised anxiety disorder. Cogn Behav Ther 2011 Sep;40(3):216-227. [doi: 10.1080/16506073.2011.582138] [Medline: 21770844]
41. Skinner HA. The Drug Abuse Screening Test. Addict Behav 1982 Jan;7(4):363-371. [doi: 10.1016/0306-4603(82)90005-3] [Medline: 7183189]
42. Bush K, Kivlahan DR, McDonell MB, Fihn SD, Bradley KA. The AUDIT alcohol consumption questions (AUDIT-C): an effective brief screening test for problem drinking. Ambulatory Care Quality Improvement Project (ACQUIP). Alcohol use disorders identification test. Arch Intern Med 1998 Sep 14;158(16):1789-1795. [doi: 10.1001/archinte.158.16.1789] [Medline: 9738608]
43. Herdman M, Gudex C, Lloyd A, Janssen MF, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011 Dec;20(10):1727-1736 [FREE Full text] [doi: 10.1007/s11136-011-9903-x] [Medline: 21479777]
44. Bouwmans C, De Jong K, Timman R, Zijlstra-Vlasveld M, Van der Feltz-Cornelis C, Tan Swan S, et al. Feasibility, reliability and validity of a questionnaire on healthcare consumption and productivity loss in patients with a psychiatric disorder (TiC-P). BMC Health Serv Res 2013 Jun 15;13:217 [FREE Full text] [doi: 10.1186/1472-6963-13-217] [Medline: 23768141]
45. de Brey H. A cross-national validation of the Client Satisfaction Questionnaire: the Dutch experience. Eval Program Plann 1983;6(3-4):395-400 [FREE Full text] [doi: 10.1016/0149-7189(83)90018-6] [Medline: 10267266]
46. Attkisson CC, Greenfield TK. Client satisfaction questionnaire-8 and service satisfaction scale-30. In: Maruish ME, editor. The Use of Psychological Testing for Treatment Planning and Outcome Assessment. Mahwah, NJ: Lawrence Erlbaum Associates; 1994:402-420.
47. Donkin L, Christensen H, Naismith SL, Neal B, Hickie IB, Glozier N. A systematic review of the impact of adherence on the effectiveness of e-therapies. J Med Internet Res 2011 Aug 05;13(3):e52 [FREE Full text] [doi: 10.2196/jmir.1772] [Medline: 21821503]
48. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? some practical clarifications of multiple imputation theory. Prev Sci 2007 Sep;8(3):206-213 [FREE Full text] [doi: 10.1007/s11121-007-0070-9] [Medline: 17549635]
49. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. J Clin Epidemiol 2019 Jun;110:63-73 [FREE Full text] [doi: 10.1016/j.jclinepi.2019.02.016] [Medline: 30878639]
50. Rubin DB. Multiple Imputation for Nonresponse in Surveys. Hoboken, NJ US: John Wiley & Sons; 1987.
51. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. URL: https://www.R-project.org/ [accessed 2024-04-02]
52. Rosenthal R. Meta-Analytic Procedures for Social Research. Newbury Park, CA: Sage Publications; 1991.
53. IBM SPSS statistics for macintosh. IBM Corp. URL: https://www.ibm.com/products/spss-statistics [accessed 2024-04-02]
54. Rapee RM, Kennedy S, Ingram M, Edwards S, Sweeney L. Prevention and early intervention of anxiety disorders in inhibited preschool children. J Consult Clin Psychol 2005;73(3):488-497. [doi: 10.1037/0022-006x.73.3.488]
55. Cuijpers P, Reijnders M, Huibers MJ. The role of common factors in psychotherapy outcomes. Annu Rev Clin Psychol 2019 May 07;15:207-231 [FREE Full text] [doi: 10.1146/annurev-clinpsy-050718-095424] [Medline: 30550721]
56. Holländare F, Gustafsson S, Berglind M, Grape F, Carlbring P, Andersson G, et al. Therapist behaviours in internet-based cognitive behaviour therapy (ICBT) for depressive symptoms. Internet Interv 2016 Mar;3:1-7 [FREE Full text] [doi: 10.1016/j.invent.2015.11.002] [Medline: 30135783]
57. Paxling B, Lundgren S, Norman A, Almlöv J, Carlbring P, Cuijpers P, et al. Therapist behaviours in internet-delivered cognitive behaviour therapy: analyses of e-mail correspondence in the treatment of generalized anxiety disorder. Behav Cogn Psychother 2012 May 01;41(3):280-289 [FREE Full text] [doi: 10.1017/s1352465812000240]
58. Kelders SM, Kok RN, Ossebaard HC, Van Gemert-Pijnen JE. Persuasive system design does matter: a systematic review of adherence to web-based interventions. J Med Internet Res 2012 Nov 14;14(6):e152 [FREE Full text] [doi: 10.2196/jmir.2104] [Medline: 23151820]
59. Hadjistavropoulos H, Peynenburg V, Nugent M, Karin E, Titov N, Dear BF. Transdiagnostic Internet-delivered cognitive behaviour therapy with therapist support offered once-weekly or once-weekly supplemented with therapist support within one-business-day: pragmatic randomized controlled trial. Internet Interv 2020 Dec;22:100347 [FREE Full text] [doi: 10.1016/j.invent.2020.100347] [Medline: 32995303]
60. Schunk D. A Markov chain Monte Carlo algorithm for multiple imputation in large surveys. Adv Stat Anal 2008 Jan 29;92(1):101-114. [doi: 10.1007/s10182-008-0053-6]
61. Bruijniks SJ, Lemmens LH, Hollon SD, Peeters FP, Cuijpers P, Arntz A, et al. The effects of once- versus twice-weekly sessions on psychotherapy outcomes in depressed patients. Br J Psychiatry 2020 Apr 07;216(4):222-230. [doi: 10.1192/bjp.2019.265] [Medline: 32029012]

Abbreviations

CAU: care as usual
CBT: cognitive behavioral therapy
CSQ-8: Client Satisfaction Questionnaire
GAD-7: Generalized Anxiety Disorder scale–7 items
iCBT: internet-based cognitive behavioral therapy
MINI: Mini International Neuropsychiatric Interview
PHQ: Patient Health Questionnaire
UvA: University of Amsterdam

Edited by J Torous; submitted 03.07.23; peer-reviewed by A Ghosh, P Watson; comments to author 13.10.23; revised version received 23.01.24; accepted 28.01.24; published 19.06.24.

Please cite as:
Koelen J, Klein A, Wolters N, Bol E, De Koning L, Roetink S, Van Blom J, Boutin B, Schaaf J, Grasman R, Van der Heijde CM, Salemink E, Riper H, Karyotaki E, Cuijpers P, Schneider S, Rapee R, Vonk P, Wiers R
Web-Based, Human-Guided, or Computer-Guided Transdiagnostic Cognitive Behavioral Therapy in University Students With Anxiety and Depression: Randomized Controlled Trial
JMIR Ment Health 2024;11:e50503
URL: https://mental.jmir.org/2024/1/e50503
doi:10.2196/50503
PMID:38896474

Original Paper

Action Opportunities to Pursue Responsible Digital Care for People With Intellectual Disabilities: Qualitative Study

Nienke M Siebelink[1], PhD; Kirstin N van Dam[1,2], MSc; Dirk R M Lukkien[3,4], MSc; Brigitte Boon[1,2,5], PhD; Merlijn Smits[6], PhD; Agnes van der Poel[1], PhD

[1]Academy Het Dorp, Arnhem, Netherlands

[2]Tranzo, Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, Netherlands

[3]Vilans, Utrecht, Netherlands

[4]Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, Netherlands

[5]Siza, Arnhem, Netherlands

[6]PBLQ, The Hague, Netherlands

Corresponding Author:
Nienke M Siebelink, PhD
Academy Het Dorp
Kemperbergerweg 139e
Arnhem, 6816RP
Netherlands
Phone: 31 0651656387
Email: nienke.siebelink@academyhetdorp.nl

Abstract

Background: Responsible digital care refers to any intentional systematic effort designed to increase the likelihood of a digital care technology developed through ethical decision-making, being socially responsible and aligned with the values and well-being of those impacted by it.

Objective: We aimed to present examples of action opportunities for (1) designing "technology"; (2) shaping the "context" of use; and (3) adjusting the behavior of "users" to guide responsible digital care for people with intellectual disabilities.

Methods: Three cases were considered: (1) design of a web application to support the preparation of meals for groups of people with intellectual disabilities, (2) implementation of an app to help people with intellectual disabilities regulate their stress independently, and (3) implementation of a social robot to stimulate interaction and physical activity among people with intellectual disabilities. Overall, 26 stakeholders participated in 3 multistakeholder workshops (case 1: 10/26, 38%; case 2: 10/26, 38%; case 3: 6/26, 23%) based on the "guidance ethics approach." We identified stakeholders' values based on bottom-up exploration of experienced and expected effects of using the technology, and we formulated action opportunities for these values in the specific context of use. Qualitative data were analyzed thematically.

Results: Overall, 232 effects, 33 values, and 156 action opportunities were collected. General and case-specific themes were identified. Important stakeholder values included quality of care, autonomy, efficiency, health, enjoyment, reliability, and privacy. Both positive and negative effects could underlie stakeholders' values and influence the development of action opportunities. Action opportunities comprised the following: (1) technology: development of the technology (eg, user experience and customization), technology input (eg, recipes for meals, intervention options for reducing stress, and activities), and technology output (eg, storage and use of data); (2) context: guidelines, training and support, policy or agreements, and adjusting the physical environment in which the technology is used; and (3) users: integrating the technology into daily care practice, by diminishing (eg, "letting go" to increase the autonomy of people with intellectual disabilities), retaining (eg, face-to-face contact), and adding (eg, evaluation moments) certain behaviors of care professionals.

Conclusions: This is the first study to provide insight into responsible digital care for people with intellectual disabilities by means of bottom-up exploration of action opportunities to take account of stakeholders' values in designing technology, shaping the context of use, and adjusting the behavior of users. Although part of the findings may be generalized, case-specific insights and a complementary top-down approach (eg, predefined ethical frameworks) are essential. The findings represent a part of an ethical discourse that requires follow-up to meet the dynamism of stakeholders' values and further develop and implement action

opportunities to achieve socially desirable, ethically acceptable, and sustainable digital care that improves the lives of people with intellectual disabilities.

(JMIR Ment Health 2024;11:e48147) doi:10.2196/48147

KEYWORDS

ethics; value-based health care; digital technology; intellectual disability; digital care

Introduction

Digital Care

As digital tools have shown great potential to enhance health care and well-being services, digital care plays a central role in the policies and plans of governments and care organizations to continue to provide good care efficiently [1-3]. Digital care refers to technology and data that inform and improve health care provision [4]. Unfortunately, today, digital care is often not aligned with the needs and values of its users and other stakeholders [5,6]. Not aligning digital care with stakeholders' needs and values results in low technology uptake. Although the importance of involving all relevant stakeholders in digital care innovation is widely accepted [7,8], they are insufficiently involved and, often, involved very late in digital care innovation [9]. Consequently, time and effort are wasted [10], and the clinical appropriateness and usability of digital care are compromised [11].

Digital care is found to have a significant impact on people's lives, especially for those with intellectual disabilities who often receive life-long care. For people with intellectual disabilities, technologies are not only applied to promote health but also to enhance independence and quality of life, such as being able to participate in the society, which creates more educational, vocational, and leisure opportunities [12-14]. In long-term care organizations, integrating technology in daily practice for people with intellectual disabilities requires insight into the needs and values of the people with intellectual disabilities themselves, as well as those of their care professionals who guide them through daily life and need to adapt their guiding strategies, the IT support staff of the care organization who provide technical support for the digital solutions, the human resources professionals who need to integrate the technology in their regular training programs within the care organization, and the data specialists who need to make decisions about incorporating the data provided by the digital care technology. Considering the increasing influence of digital care on several domains of the life of people with intellectual disabilities [13], the design and implementation of technologies should be well considered.

Responsible Design and Implementation

Ideally, the design and implementation of digital care for people with intellectual disabilities are "responsible": to include any intentional systematic effort designed to increase the likelihood of a digital care technology developed through ethical decision-making, being socially responsible and aligned with the values and well-being of those influenced by it [15]. Values, defined as "convictions or matters that people feel should be strived for in general and not just for themselves to be able to lead a good life or realise a good society" [16], are commonly considered within ethical discourse. These values function as moral compasses that guide certain actions, for example, in the design and implementation of technology. However, the ethics of digital care is not a common subject of study [17,18]. There is limited empirical evidence describing how to address stakeholders' values within their context, in this case, the context of long-term care for people with intellectual disabilities, even though it is broadly recognized that responsible design and implementation of digital care require insight into and sensitivity toward specific contexts of use [17,19,20]. There is a need for context-specific studies about how certain values matter to the stakeholders of particular technologies and how these values can be accounted for in technology design and implementation.

Guidance Ethics

The "guidance ethics approach" [21] is a relatively new method for reflection about and guidance for the responsible design and implementation of technologies in the context of use, developed by the ECP (Platform for the Information Society), the Netherlands. The approach is applied in various fields, such as municipalities, government, security, police, and health care sector, and regarding various cases, such as a biofeedback app for people with profound intellectual and multiple disabilities and challenging behavior or the use of artificial intelligence for nighttime monitoring in disability care [22]. The guidance ethics method involves a multistakeholder workshop, in which stakeholders' values are identified based on an exploration of the experienced and expected positive and negative effects of using a specific technology. Subsequently, workshop participants mutually formulate action opportunities to account for these values in the specific context of use. In this study, we used this method to identify stakeholders' values and formulate action opportunities for the responsible design and implementation of specific technologies used in long-term care for people with intellectual disabilities.

The guidance ethics approach—more extensively described in the *Methods* section—has several advantages compared with other research methods focused on ethics in design. One of the most well-known methodologies considering ethics through values in design is "Value Sensitive Design" [23]. Although the idea of embedding values in technology originated from this method, it does not provide the tools to empirically study values. However, guidance ethics is practical and hands-on, allowing stakeholders in a workshop to contribute to identifying the values affected by the use of technology. There are tools, such as the "interactive technology assessment" [24], that also provide hands-on tools, but these solely focus on studying the values of 1 user. Guidance ethics enables to involve a diverse group of stakeholders to identify a comprehensive set of values [7] and facilitates stakeholders to better understand the position of others [25]. Although there are methods, such as an evidence-informed, deliberative process approach to a health technology assessment [7,26] that involves multiple stakeholders also, this method, in contrary to guidance ethics, does not translate insights into concrete action opportunities for responsible technology use. To the best of our knowledge, guidance ethics is the only method that enables the study of values involving multiple stakeholders and directly translates these into concrete action opportunities.

Objective

In this study, we applied the guidance ethics approach to three digital care technologies that are currently being developed or implemented within care organizations for people with intellectual disabilities:

1. The design of "Kookapp for groups" (developed by care organizations *Amerpoort* and *Reinaerde* and IT company *Ilionx*, Utrecht): this is a web application to support group workers with the preparation of healthy and tasty meals for groups of people with intellectual disabilities, from choosing recipes and buying ingredients to cooking and serving the meals.
2. The implementation of the "SignaLEREN" app (developed by care organization *Koraal* and IT company *Ivengi*, Maastricht): this app is used by people with intellectual disabilities or autism spectrum disorder to regularly gauge their emotional state; in the case of increased stress, they can choose a personalized stress-reducing activity within the app, such as watching a video clip or listening to certain music.
3. The implementation of SARA (developed by *SARA Robotics*, Eindhoven): this is a social robot that provides activities (eg, exercises, games, and music) during day care to stimulate interaction and physical activity among older people with intellectual disabilities.

With these 3 cases, we aimed to present examples of action opportunities to guide the responsible use of digital care for people with intellectual disabilities.

Methods

Participants

The 4 care organizations of the 3 cases participated in the Innovation Impulse Disability Care, a 3-year program initiated by the Dutch Ministry of Health, Welfare, and Sport in 2019. This program aimed to accelerate digital transformation in long-term care by providing support in implementing technology in the everyday practice of 26 disability care organizations [27]. In each organization, the implementation started by defining a topical care issue from the perspective of and together with people with a disability, followed by the selection of a technology that contributed to the solution of this care issue. In addition, organizations evaluated their IT and organizational readiness to implement the selected technology [28,29]. The care issues included, for example, improving day structure [30] or sleep-wake patterns [31], lowering stress levels, and increasing independent living [32]. Digital care technologies included sensors, domotics, social robotics, and apps. The Innovation Impulse program also entailed researching the factors influencing the implementation (NM Siebelink, unpublished data, 2024).

All 26 care organizations were invited to apply for participation in this study. Guidance ethics workshops were conducted in 4 care organizations, for the 3 cases described previously. Project leaders of the Innovation Impulse program within each care organization invited a purposefully diverse group of workshop participants from their organizations, for example, people with intellectual disabilities, relatives, care professionals, policy advisors, managers, members of the board of directors, IT staff, and technology developers.

In total, 26 individuals participated in this study. Participants' characteristics are presented in Table 1. No personal data such as sex or age were collected. Almost all participants had some knowledge about the specific technology: of the 26 participants, 13 (50%) considered themselves informed, 9 (35%) had practical experience with the technology, and 4 (15%) were unfamiliar with the respective technology before participating in the workshop.

Table 1. Study participants' characteristics.

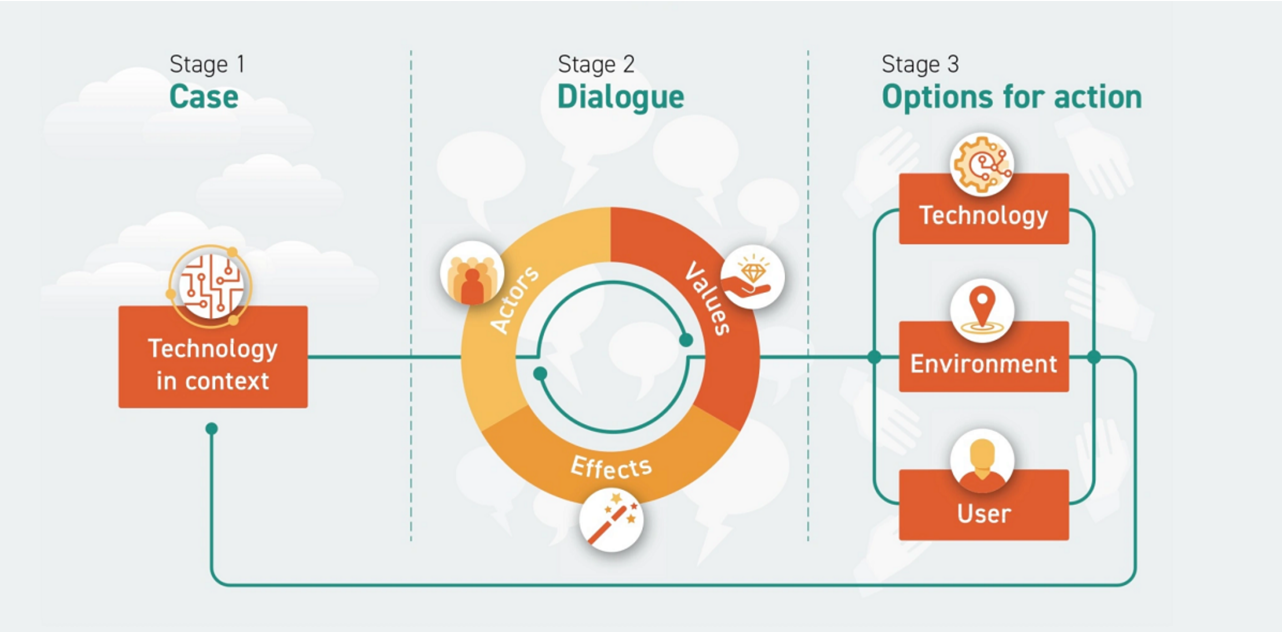
Participant	Case 1: Kookapp for groups (n=10), n (%)	Case 2: SignaLEREN app (n=10), n (%)	Case 3: social robot, SARA (n=6), n (%)
People with intellectual disabilities (or a representative)	2 (20)	0 (0)	0 (0)
Management or policy maker	2 (20)	1 (10)	1 (17)
IT staff or technology developer	1 (10)	2 (20)	2 (33)
Care professional or team leader	4 (40)	3 (30)	2 (33)
Other (eg, project leader or consultant)	1 (10)	4 (40)	1 (17)

Procedure and Materials

This study had a qualitative research design, using guidance ethics workshops to collect data. In total, three 3.5-hour multiple stakeholder workshops were conducted by trained workshop leaders from ECP (2 per workshop). The workshops were attended in person in May, June, and September 2022. Data were collected by means of a questionnaire (described in this section) completed on paper by the participants during the workshop. In addition, the information that the workshop leaders wrote on the flip charts was collected by taking photographs of the flip charts. In total, 2 researchers (KNvD and NMS or AvdP) were present during each workshop to observe and explain the study and the questionnaire; they did not engage in the workshop.

The questionnaire—constructed by researchers (NMS and KNvD) for this study—followed the workshop outline (Figure 1 [21,33]). In stage 1 of the workshop (case), the project leader presented the case, that is, information about the technological solution, its aim, the way it works, for which target group, and in which daily (care) process. Thereafter, data about the participants' characteristics and their familiarity with the technology were collected.

Figure 1. Outline of the guidance ethics approach (adapted from Verbeek and Tijink 2020 [21], which is published under Creative Commons Attribution 4.0 International License [33]).



In stage 2 (dialogue), participants were first asked to call out all actors that are or should be affected by or involved with the use of the technology; the workshop leaders wrote these actors on a flip chart. Second, participants were asked to write down any positive and negative effects of the technology they could think of, not only from their own perspective but also any effect that came to mind. Next, all of them were asked to mention an effect until all effects were called out. Again, the workshop leader wrote these effects on a flip chart, and the effects were discussed, supplemented, and clustered by the workshop leaders and participants. Third, the workshop leaders identified values based on the clustered effects, and these values were adjusted in discussion with the participants. Finally, in stage 2, each participant determined the top 3 values that they deemed most important for their professional role in the particular case. These values were marked on the flip chart and discussed, after which the top 3 values of the total group were determined.

For stage 3 (action opportunities), participants were divided into 3 subgroups with diverse stakeholders in each subgroup. These subgroups were invited to come up with action opportunities to achieve a highly value-driven use of the respective technology in the context of the specific case, using the top 3 values of the group as starting point. A slight deviation from the protocol was that the group of the third case (social robot, SARA)—which was relatively small—was not divided into subgroups in stage 3 and did not explicitly focus on the top 3 values. The subgroups were instructed to come up with action opportunities for the following (respectively): the design of the technology (technology), shaping the environment or context of use (context), and adjusting the behavior of the users (users). Workshop leaders explained the meaning of "action opportunities" by using the example of the technology "car." Driving a car should be safe (value); therefore, cars have seat belts and automatic brakes (action opportunities for the technology to improve safety), traffic lights and other infrastructure guide drivers (action opportunities for the environment to improve safety), and drivers practice driving and learn the rules before receiving a license (action opportunities for the user's behavior to improve safety). Participants were asked to write down the action opportunities that came to mind, which were then discussed and collected on a flip chart in the subgroups. Each subgroup presented their action opportunities, which were discussed plenarily. At the end of the workshop, participants were asked to write down in the questionnaire any new effects, values, or action opportunities that came to mind that were not mentioned during the workshop.

Ethical Considerations

Participants were informed about the study and privacy statement, and they provided consent by completing the questionnaire anonymously. The local Medical Research Ethics Committee Oost-Nederland deemed the research in the Innovation Impulse program not subject to the Medical Research Involving Human Subjects Act ("Wet medisch-wetenschappelijk onderzoek met mensen"; file 2021-8293).

Analyses

Data from the questionnaires about positive and negative effects, values, and action opportunities were entered in Excel (Microsoft Corporation; 2018). The data set was checked for completeness using notes of the workshop observations, pictures of the flip charts from the workshops, and the workshop reports made by ECP's workshop leaders. Analyses were conducted using a bottom-up approach: participants' descriptions were the starting point leading to the derivation of themes. First, 2 researchers (KNvD and NMS) independently derived themes from the "effects data" per case. That is, effects regarding a similar subject were given a descriptive name. For example, "joyous end users" and "end users can experience more enjoyment" were named "enjoyment of end users," which was then considered an effect theme. The 2 researchers compared and discussed their effect themes until consensus was reached. Next, the effect themes of all 3 cases were written on digital Post-it notes and visually arranged, so that related effect themes were near each other. Digital Post-it notes on which values were presented were added for each theme from which the values were abstracted. Furthermore, 2 researchers (KNvD and NMS) also derived themes from the action opportunity data per case. Next, analyses and discussions were conducted regarding which values were represented by the action opportunity themes. For example, the action opportunity theme "Give the person with intellectual disability some self-direction in the use of the technology" is mainly related to the value "autonomy," whereas the action opportunity theme "keep the goal in mind and deploy technology as a means rather than a goal in itself" is related to the value "quality of care." The preliminary results were presented and discussed in an interpretation meeting with workshop leaders and project leaders from the care organizations (participants from all 3 workshops were present). Input and feedback from this meeting were used for further analyses through an iterative process.

Results

Overview

The 3 workshops provided insight into how effects were translated into values and, subsequently, how values were translated into action opportunities for technology, context, and users of a specific technology. The numbers of collected effects, values, and action opportunities for each of the 3 cases are presented in Table 2. An overview of their content is provided in Multimedia Appendix 1; for readability, effects and action opportunities were shortened, and similar ones were combined in the overview.

Data about values and action opportunities were missing from a participant who could only attend the first part of the workshop. Furthermore, the person with intellectual disability from case 1 formulated effects and action opportunities together with a care professional.

Table 2. Number of participants per case and the amount of data collected.

Case	Participants (n=26), n (%)	Effects (n=232), n (%)	Values (n=33), n (%)	Action opportunities (n=156), n (%)
1—Kookapp for groups	10 (38.5)	102 (43.9)	13 (39.4)	74 (47.4)
2—SignaLEREN app	10 (38.5)	66 (28.4)	7 (21.2)	54 (34.6)
3—Social robot, SARA	6 (23.1)	64 (27.6)	13 (39.4)	28 (17.9)

Examples of Action Opportunities

Table 3 presents examples of action opportunities for (1) designing "technology"; (2) shaping the "context" of use; and (3) adjusting the behavior of the "users" to guide the responsible use of digital care for people with intellectual disabilities. Given that describing all results (which can be found in Multimedia Appendix 1) is beyond the scope of this paper, Table 3 highlights 1 example per case based on one of the most prominent values in that case, and we have described the effects and action opportunities linked to that value. Following the examples, we have reflected about general observations within and overarching the 3 cases.

Table 3. Examples of combinations of effects, a value, and action opportunities for the 3 cases.

Cases and categories	Examples
Case 1: Kookapp for groups—a web application to support healthy cooking for groups of people with intellectual disabilities	
Selected value	• Quality of care
Positive effects	• Connectedness through choosing, cooking, and eating together • Continuity of meal quality; not being dependent on care professionals' skills • Awareness of the importance of good nutrition • More time for the people with intellectual disabilities • Equality (differences regarding meals for people with intellectual disabilities between care organizations become small when they use the web application)
Negative effects	• Excessive focus on health compared with enjoying tasty food
Action opportunities—technology	• None mentioned regarding the selected value
Action opportunities—context (care organization)	• Evaluate efficiency, health, and eating pleasure continuously • Provide the care professionals with instructions about how to use the web application along with a manual
Action opportunities—users (care professionals)	• Invest in understanding the web application to use it properly • Have a backup plan for situations in which the web application does not work • Know the dietary preferences and needs of each person with intellectual disability in the group • Know what to do when a person with intellectual disability does not want to participate (or experiences less fun) in cooking with the web application
Case 2 : SignaLEREN app—an app to support people with intellectual disabilities in autonomously dealing with stress or anxiety	
Selected value	• Autonomy

Positive effects	• The following were the positive effects for the people with intellectual disabilities:
	  • More self-direction, independence, and personal autonomy
	  • An extra support option besides support from care professionals
	  • Increased awareness of own stress and its causes

Negative effects	• Counterproductive effects of the app on stress if it does not work • Less autonomy when using the app feels obligatory • The app as a barrier to seeking contact with the care professional

Action opportunities—technology	• Enable people with intellectual disabilities to make choices themselves:
	  • Set the regularity of question pop-ups in the app
	  • Disregard the notifications at unsuitable moments by choosing the response option "I don't want to answer this question (right now)"
	  • Delete data
	  • Schedule an appointment with their care professional via the app

Action opportunities—context	• Give people with intellectual disabilities access to the back end of the app, so that they can adjust specific content in the app themselves

Action opportunities—users	• Give people with intellectual disabilities some self-direction regarding the use of the app:
	  • Discuss what using the app entails and what happens with the data
	  • Give guidance in setting up and using the app at their own pace
	  • Evaluate app use and effects frequently during coaching moments
	• Support people with intellectual disabilities in a different way, eg, redirect them to the app first, "Have you completed the app?"

Case 3: Social robot, SARA—a robot to support the physical and social activities for people with intellectual disabilities	
Selected value	• Privacy
Positive effects	• None mentioned regarding the selected value
Negative effects	• Risk of privacy infringement owing to storage and sharing of personal data • Insufficient insight into what data are collected and stored when the functionalities of the robot are expanded

Cases and categories	Examples
Action opportunities—technology	• Differentiate between accounts for administrators and care professionals using the robot to restrict access to personal data • Establish a maximum storage period for personal data before they are automatically deleted

Action opportunities—context	• Train the care professionals using the robot in accordance with the privacy law
	• Restrict the number of people who have access to personal data
	• Read and reconsider the consent statements annually with people with intellectual disabilities (or their representatives) who use the robot—taking into account any changes in the functionalities of the robot and therefore storage of other data
	• Revise the organization's privacy policy
	• When people with intellectual disabilities use the robot to have contact with relatives, the privacy of both should be protected:
	  • Use headphones during this contact
	  • Create a cozy private "corner" in the location

Action opportunities—users	• None mentioned regarding the selected value

Observations About and Differences and Similarities Among the Cases

Effects

Part of the effects that were collected was case specific. For example, effects regarding healthy food were mentioned only in case 1 (Kookapp for groups), effects regarding insight into the stress of the person with intellectual disability were mentioned only in case 2 (SignaLEREN app), and effects regarding activation or development of cognitive skills of the person with intellectual disability were mentioned only in case 3 (social robot, SARA). Apart from case-specific effects, several general themes that were extracted from the effects across all 3 cases are presented in Table 4.

Table 4. General (ie, not case specific) effect themes, values, and action opportunity themes.

Categories and themes found across the 3 cases (Kookapp for groups; SignaLEREN app; and social robot, SARA)	Actors to whom the themes mainly apply
Positive effects	
Customization of care	People with intellectual disabilities
Increase in self-reliance or self-direction	People with intellectual disabilities
Ease of work or job satisfaction	Care professionals
Efficiency or labor saving or time saving	Care organizations
Negative effects	
Dependency on IT infrastructure	People with intellectual disabilities, care professionals, and care organizations
Risks related to the storage of privacy-sensitive data	People with intellectual disabilities, care professionals, and care organizations
Perception of having "yet another system"	Care professionals
Cost or time investment	Care organizations
Values	
Quality of care	People with intellectual disabilities
Autonomy	People with intellectual disabilities
Privacy	People with intellectual disabilities, care professionals, and care organizations
Job satisfaction	Care professionals
Efficiency or affordability of care	Care organizations and society
Action opportunities	
Keep the content (recipes, interventions, and activities) up to date	Care professionals
Train the care professionals regarding how to use the technology well	Care organizations
Connect the use of the technology to goals in the individual care plan or electronic health record	Care professionals and technology developers
Focus on upscaling (more users of the technology in the care organization)	Care organizations

Values

Table 5 shows the values that were identified as the top 3 values during the 3 workshops. Note that only in case 1 (Kookapp for groups), participants chose 4 values. Values identified in all 3 cases were "quality of care," "autonomy," "privacy," "job satisfaction," and "efficiency or affordability of care" (Table 4). In all 3 cases, certain values apply to a specific actor. For example, "job satisfaction" applied to care professionals, whereas "autonomy" was primarily related to people with intellectual disabilities. Other values applied to several actors (eg, "reliability") or an actor group; for example, "efficiency or affordability of care" was related to the care organization or even society.

Table 5. Top values identified during "guidance ethics" workshops, based on the personal top 3 values of all participants per case.

Case	Top values
1—Kookapp for groups	Quality of care, efficiency, health, and enjoyment
2—SignaLEREN app	Quality of care, autonomy, and reliability
3—Social robot, SARA	Quality of care, autonomy, and privacy

Action Opportunities

Action opportunities regarding the technology comprised the development of the technology itself (eg, optimization of user experience and customization), input into the technology (eg, recipes for meals in case 1—Kookapp for groups; intervention options for reducing tension in case 2—SignaLEREN app; and activities in case 3—social robot, SARA), and output of the technology (eg, storage and use of data). Action opportunities regarding the context covered the need for guidelines, training and support, policy or agreements, and adjustments to the physical environment in which the technology is used. Action opportunities regarding the users mainly focused on how care professionals can integrate the technology into their daily care practice. Some behaviors need to be diminished (eg, care professionals need to "let go" instead of "take over" to give the person with intellectual disability more autonomy), some behaviors must be retained (eg, face-to-face contact moments), and some behaviors need to be added (eg, evaluation moments). Although some general themes across all 3 cases were identified (Table 4), most action opportunities were context specific.

For case 1 (Kookapp for groups), most action opportunities were listed for the value "user convenience." Action opportunities included optimizing the user experience with the web application (eg, using icons and less text) and providing resources (eg, placing magnets on the kitchen wall to hold tablets while cooking). Action opportunities that stood out because they were mentioned by several participants were related to an attractive design and ease of operation of the web application. Although the value "health" was a top value in case 1, relatively few action opportunities were formulated for this value.

For case 2 (SignaLEREN app), none of the values stood out, but 4 values were evenly represented among most action opportunities: "quality of care," "autonomy" (Table 3), "reliability," and "efficiency." "Quality of care," "autonomy," and "reliability" were the top 3 values. For "quality of care," action opportunities included keeping the goal in mind and deploying the app as a means rather than a goal in itself (eg, personal goals of the person with intellectual disability as starting point for the conversation about how to use the app), training the care professionals on the use of the app, and having a person-centered approach (ie, customizing the app). Action opportunities for the value "reliability" included continuous provision of easily accessible support (eg, assigning SignaLEREN coaches and arranging a 24-h helpdesk) and the maintenance of the app organized within the own organization. Finally, action opportunities for the value "efficiency" included integrating the app in the care process (eg, embedding the use of the app in a particular care methodology), scaling up the use and adoption of the app (eg, deploying the app with all care professionals to whom it applies), and extracting and using data from the app (eg, built-in notifications in the app for when the person's stress level is likely to become very high). The more frequently mentioned action opportunities included giving the person with intellectual disability access to the personal settings of the app (eg, frequency of prompts and data access rights), securing face-to-face contact of the person with intellectual disability and care professional, using data from the app to provide insights into stress level trends, connecting the app with the electronic health record, and assigning a SignaLEREN coach.

For case 3 (social robot, SARA), most action opportunities were related to the value "privacy" (Table 3). Action opportunities that stood out because they were mentioned by several participants included linking the use of the robot to individual care goals, connecting the robot with the electronic health record, and expanding the content that the robot can present. Although the value "autonomy" was a top value in case 3, relatively few action opportunities were related to this value, whereas relatively many action opportunities were linked to "effectiveness" (eg, optimizing the content that the robot can present and recurrent evaluation), which was not chosen as a top value.

Discussion

Relevance

Often, the development, implementation, and use of digital care does not entail an intentional and systematic effort to include the ethical considerations of all involved stakeholders [34]. Therefore, new technologies and the processes to integrate them into daily practice are often not aligned with the values and well-being of those influenced by them [15]. Instead, ethics is merely considered a separate area of attention (eg, a separate line of investigation or work package within projects) discussed by a distinct group of experts [35].

This study illustrates the types of insights that are gained when various stakeholders are involved in the reflection about the ethical impact of specific technologies and how this impact can be influenced for the better. This is illustrated using 3 cases of different digital care technologies for people with intellectual disabilities: a web application for cooking for groups (Kookapp for groups), an app for stress regulation (SignaLEREN app), and a robot for interaction and physical activity (social robot, SARA). Our findings may help researchers, innovators, and users of technology to move from a rather abstract thinking about ethics and responsible innovation toward effective practical approaches in which all stakeholders can be involved.

Principal Findings

In a short amount of time (three 3.5-h workshops), relatively much information was gathered in a multistakeholder setting regarding (1) positive and negative effects for various stakeholders of a specific digital care technology for people with intellectual disabilities; (2) values underlying these effects; and (3) action opportunities to take into account important values in the design, implementation, and use of the specific technology. The effects were primarily case specific, as they described the implementation of a technology in a specific context, but several general themes were also recognized. The latter included the effects of the technology on customization of care, dependency on IT infrastructure, self-reliance or self-direction of people with intellectual disabilities, risk of privacy infringement, care professionals' ease of work, workload, and efficiency and investment. When all the effects were abstracted into values, several values were identified in all 3 cases and were found to be related to the general effects. These values were quality of care, autonomy, privacy, job satisfaction, and efficiency or affordability of care. Most action opportunities were related to the top values from the respective cases, as can be expected when the guidance ethics approach (stage 3) is followed. Hence, many action opportunities from case 1 (Kookapp for groups) were related to "enjoyment," "efficiency," and "quality of care." However, relatively few action opportunities involved the top value, "health." Notably, most action opportunities from case 1 were related to "user convenience"; however, this was not identified as a top value. In case 2 (SignaLEREN app), most action opportunities were related to the top values ("quality of care," "autonomy," and "reliability") and to "efficiency." In case 3 (social robot, SARA), the top values, "quality of care" and "privacy," were well represented among the action opportunities, but few were related to the top value, “autonomy,” and relatively many were related to “efficiency” and “effectivity.”

Although action opportunities can only be described in relation to specific sociomaterial contexts [36] (ie, specific technologies in their contexts of use), our study reveals that, at a higher level, there are similarities regarding effects, values, and action opportunities for different cases. Thus, it may be wise for stakeholders of digital care technologies to not only learn how technologies can be responsibly used within their own context but also seek inspiration from similar contexts in which other technologies are used and from different contexts in which the same or comparable technologies are used. However, caution should be exercised when generalizing case-specific effects, values, and action opportunities to a broad scope.

Comparisons With Previous Studies

This is the first study to provide a broad overview of actual action opportunities for responsible digital care for people with intellectual disabilities. In their 2020 reports, the Dutch Centre of Ethics and Health advised the Dutch Ministry of Health, Welfare, and Sport about ethics regarding digital care such as apps and robots [37,38]. The themes discussed in the report regarding apps are cost savings, increase of autonomy, increase of well-being, unrest, information overload, decrease of human contact, overemphasis on health that can lead to medicalization, and increase of differences in health and inequality [37]. Notably, the diametrically opposed side of inequality was raised in our study, namely that differences between care organizations would decrease if they would organize meals using the Kookapp for groups. The report regarding care robots discusses meaningful contact, dignity, autonomy, dependency, privacy, and justice [38]. Apart from information overload and justice, all themes also appeared in ≥1 of the 3 cases in our study. This shows that our results covered most of the essential topics that ethics experts recognized.

Studies of ethics and health often describe themes that have a positive or negative load (eg, increase of autonomy or increase of inequality, respectively) or identify ethical harms [17,18,37,39]. Our data revealed that, in most cases, 2 sides of the same coin were considered in the multistakeholder setting, for instance, technology as a facilitator and a burden for care professionals’ work (all cases); cost or labor savings and high costs or time investment (all cases); positive and negative effects of the focus on a health theme (Kookapp for groups: awareness of the importance of healthy food vs a lot of emphasis on health at the cost of enjoying tasty food; SignaLEREN app: improving the stress signaling plan vs risk of medicalization of normal stress); and increase and risks of autonomy (SignaLEREN app: person with intellectual disability is less dependent on care professional but possibly also less “visible”). The advantage of using values—instead of themes or harms—as a starting point for fostering responsible use of digital care is that values are neutral and hence facilitate the consideration of both sides of the same coin [40].

Strengths and Limitations

As this study illustrates, the guidance ethics approach can be a valuable and low-key method to gain insight into different stakeholders’ experienced and expected positive and negative effects and values affected (or at stake) when using a specific technology and insight into action opportunities for responsible digital care. However, we recognize that the insights gained in this respect may fall short in terms of correctness (ie, being in agreement with facts or with what is generally accepted), concreteness (ie, being specific and detailed), and completeness (ie, the extent to which all relevant effects, values, and action opportunities have been identified) [41].

The correctness of our results about effects may be limited owing to, among others, a general lack of methodologically sound studies of the effects of digital care for people with intellectual disabilities. Hence, there was little to no evidence from scientific studies of the specific care technologies to be presented in stage 1 of the workshops. Therefore, the collected positive and negative effects are mainly based on subjective effects but from stakeholders with lived experience with the specific technology. In addition, an inherent characteristic of qualitative data analysis is that deriving themes from the data (including identifying values based on the effects of the technology) involves interpretation by the analyst. To limit subjectivity, values were discussed during the workshop with all participants, and themes were created independently by 2 researchers and discussed until consensus was reached. Another discussion point linked to correctness is that the values were discussed and presented as relatively stable entities, while they are neither stable nor singular [42,43]. Values may be affected by time and thus constantly defined and redefined (value dynamism), for instance, because users have gained experience with technology [41,43]. Although it may be challenging for researchers, innovators, and other stakeholders to continuously respond to this dynamism, a starting point could be to regularly collect the stakeholders’ perspectives about effects, values, and action opportunities.

The method used in this study has advantages regarding concreteness. For example, although the values are abstract, their definitions are embedded in the concrete effects from which they are derived. Moreover, the method results in relatively concrete output (action opportunities) compared with most ethics research on digital care [44,45], and action opportunities apply to the specific context of use. However, it was not always deducible from the workshop data what a participant specifically meant by an effect or action opportunity or to whom (eg, care professional or person with intellectual disability) specific insights applied. In the cases used in this study, it is not straightforward who is meant by the “user” of the care technology. To improve this, workshop leaders should be alert and ask participants to clarify whether they mean the care professional or the person with intellectual disability.

Regarding completeness, this study did not aim to be exhaustive in collecting effects, values, and action opportunities. However, we aimed to include a diverse sample of participants. Despite the accessibility of the workshops for people with intellectual disabilities, a participant with intellectual disability was included only in case 1 (Kookapp for groups). In case 3 (social robot, SARA), the person with intellectual disability withdrew on the morning of the workshop (with a valid reason), and in case 2 (SignaLEREN app), no person with intellectual disability was invited. In addition, other relevant stakeholders were absent, for example, relatives of the people with intellectual disabilities, the board of directors, or representatives of health insurers. Although all relevant stakeholders that participants could think of were identified during the workshop and participants were asked to keep them all in mind, some perspectives may be missing in the results. Including people with disabilities requires special attention, as this is not common in the co-design or cocreation of digital care technologies [46]; however, this is upcoming [13,47,48]. To improve stakeholder inclusion in general, it may be useful to consult "design principles" for stakeholder engagement [49].

Furthermore, the method of deriving values from effects is suitable for identifying "proximal" values on the micro level that are specific to the technology and context of use [21], but mesolevel or macrolevel effects and more "distal" values may be missed [50] (such as "social justice," which is an important theme in studies of ethics and digital care [45,51,52]). Whenever missed in bottom-up ethical dialogues (such as in this study), proximal and distal values (or principles) from predefined ethical frameworks could be brought in as "top-down" guidance. At the same time, the bottom-up approach is a strength of the guidance ethics approach, revealing important topics such as enjoyment of the person with intellectual disability or job satisfaction, which may be missed when an ethical theory is applied to a case instead [17,53]. In this sense, we argue that the top-down and bottom-up approaches are complementary. Hence, we suggest moving back and forth between the perspectives of stakeholders affected by technology when implementing digital care and ethical frameworks or perspectives of experts about digital care ethics [19,54].

Finally, the action opportunities identified in this study require follow-up in practice. Responsible use of technology requires being continuously responsive and adaptive to new insights that are gained regarding effects, values, and action opportunities, from early design to local implementation and use [55,56]. In addition, it is conceivable that trade-offs between action opportunities need to be made owing to value conflicts (eg, autonomy vs duty of care [57]) and costs. Future studies may shed light on how action opportunities, once formulated, are further operationalized and applied by technology designers, user organizations, and individual end users of the technology and on what factors withhold stakeholders from doing so. Previous studies indicated that ethical concerns of stakeholders might considerably slow the pace of digital care innovation, implying that responsible innovation could be a core catalyst for the progress of digital care overall [18]. Through explicit attention to and communication about responsible digital care, not only are ethical concerns taken into account but also support and acceptance among the involved stakeholders are generated. This increases the chances for the successful implementation of socially desirable, ethically acceptable, and sustainable digital care that improves the lives of people with disabilities.

Acknowledgments

This study was conducted as a part of "Innovation Impulse Disability Care" (in Dutch: Innovatie-impuls Gehandicaptenzorg), a 3-year program funded by the Dutch Ministry of Health, Welfare, and Sport. The authors thank the workshop leaders of the Platform for the Information Society and all study participants for their participation.

Authors' Contributions

NMS, DRML, MS, BB, and AvdP were involved in conceptualization. NMS, BB, and AvdP contributed to the methodology. NMS and KNvD were responsible for the software, conducted the validation and formal analysis, were responsible for the resources, and performed data curation and visualization. NMS, KNvD, and AvdP conducted the investigation. NMS, KNvD, DRML, MS, BB, and AvdP wrote the original draft and were involved in reviewing and editing. AvdP and BB were involved in supervision and funding acquisition. NMS and AvdP were involved in project administration.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of the effects, values, and action opportunities collected from the workshops for the 3 cases: Kookapp for groups; SignaLEREN app; and social robot, SARA.
[DOCX File , 43 KB - mental_v11i1e48147_app1.docx]

References

1. Integraal zorg akkoord: samen werken aan gezonde zorg. Rijksoverheid. 2022 Sep. URL: https://www.rijksoverheid.nl/documenten/rapporten/2022/09/16/integraal-zorgakkoord-samen-werken-aan-gezonde-zorg [accessed 2024-02-21]
2. Study on health data, digital health and artificial intelligence in healthcare. Publications Office of the European Union. 2022. URL: https://op.europa.eu/en/publication-detail/-/publication/179e7382-b564-11ec-b6f4-01aa75ed71a1/language-en [accessed 2024-02-21]
3. Global strategy on digital health 2020-2025. World Health Organization. 2021 Aug 18. URL: https://www.who.int/publications/i/item/9789240020924 [accessed 2024-02-21]

4. The Lancet Digital Health. A digital (r)evolution: introducing The Lancet Digital Health. Lancet Digit Health 2019 May;1(1):e1 [FREE Full text] [doi: 10.1016/S2589-7500(19)30010-X] [Medline: 33323234]
5. Sarri G, Freitag A, Szegvari B, Mountian I, Brixner D, Bertelsen N, et al. The role of patient experience in the value assessment of complex technologies - do HTA bodies need to reconsider how value is assessed? Health Policy 2021 May;125(5):593-601 [FREE Full text] [doi: 10.1016/j.healthpol.2021.03.006] [Medline: 33814201]
6. Smits M, Ludden GD, Verbeek PP, van Goor H. How digital therapeutics are urging the need for a paradigm shift: from evidence-based health care to evidence-based well-being. Interact J Med Res 2022 Oct 20;11(2):e39323 [FREE Full text] [doi: 10.2196/39323] [Medline: 36264624]
7. Baltussen R, Jansen MP, Bijlmakers L, Grutters J, Kluytmans A, Reuzel RP, et al. Value assessment frameworks for HTA agencies: the organization of evidence-informed deliberative processes. Value Health 2017 Feb;20(2):256-260 [FREE Full text] [doi: 10.1016/j.jval.2016.11.019] [Medline: 28237205]
8. Van Velsen L, Wentzel J, Van Gemert-Pijnen JE. Designing eHealth that matters via a multidisciplinary requirements development approach. JMIR Res Protoc 2013 Jun 24;2(1):e21 [FREE Full text] [doi: 10.2196/resprot.2547] [Medline: 23796508]
9. Baines R, Bradwell H, Edwards K, Stevens S, Prime S, Tredinnick-Rowe J, et al. Meaningful patient and public involvement in digital health innovation, implementation and evaluation: a systematic review. Health Expect 2022 Aug;25(4):1232-1245 [FREE Full text] [doi: 10.1111/hex.13506] [Medline: 35526274]
10. Panyard DJ, Ramly E, Dean SM, Bartels CM. Bridging clinical researcher perceptions and health IT realities: a case study of stakeholder creep. Int J Med Inform 2018 Feb;110:19-24 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.11.014] [Medline: 29331251]
11. Pagliari C. Design and evaluation in eHealth: challenges and implications for an interdisciplinary field. J Med Internet Res 2007 May 27;9(2):e15 [FREE Full text] [doi: 10.2196/jmir.9.2.e15] [Medline: 17537718]
12. Sheehan R, Hassiotis A. Digital mental health and intellectual disabilities: state of the evidence and future directions. Evid Based Ment Health 2017 Nov;20(4):107-111 [FREE Full text] [doi: 10.1136/eb-2017-102759] [Medline: 28947677]
13. Oudshoorn CE, Frielink N, Nijs SL, Embregts PJ. eHealth in the support of people with mild intellectual disability in daily life: a systematic review. J Appl Res Intellect Disabil 2020 Nov 23;33(6):1166-1187 [FREE Full text] [doi: 10.1111/jar.12758] [Medline: 32578361]
14. Owuor J, Larkan F. Assistive technology for an inclusive society for people with intellectual disability. Stud Health Technol Inform 2017;242:805-812. [Medline: 28873889]
15. Ahmadpour N, Ludden G, Peters D, Vold K. Editorial: responsible digital health. Front Digit Health 2021;3:841477 [FREE Full text] [doi: 10.3389/fdgth.2021.841477] [Medline: 35128520]
16. van de Poel I, Royakkers · L. Ethics, Technology, and Engineering: An Introduction. Hoboken, NJ: John Wiley & Sons; Feb 25, 2011.
17. Shaw JA, Donia J. The sociotechnical ethics of digital health: a critique and extension of approaches from bioethics. Front Digit Health 2021 Sep 23;3:725088 [FREE Full text] [doi: 10.3389/fdgth.2021.725088] [Medline: 34713196]
18. Landers C, Vayena E, Amann J, Blasimme A. Stuck in translation: stakeholder perspectives on impediments to responsible digital health. Front Digit Health 2023 Feb 6;5:1069410 [FREE Full text] [doi: 10.3389/fdgth.2023.1069410] [Medline: 36815171]
19. Lukkien DR, Nap HH, Buimer HP, Peine A, Boon WP, Ket JC, et al. Toward responsible artificial intelligence in long-term care: a scoping review on practical approaches. Gerontologist 2023 Jan 24;63(1):155-168 [FREE Full text] [doi: 10.1093/geront/gnab180] [Medline: 34871399]
20. Kiran AH, Oudshoorn N, Verbeek PP. Beyond checklists: toward an ethical-constructive technology assessment. J Responsible Innov 2015 Jan 13;2(1):5-19. [doi: 10.1080/23299460.2014.992769]
21. Verbeek PP, Tijink D. Guidance ethics approach: an ethical dialogue about technology with perspective on actions. ECP | Platform voor de InformatieSamenleving. 2020. URL: http://tinyurl.com/m3n8ydyu [accessed 2024-02-21]
22. Cases. Aanpak begeleidigsethiek. URL: https://begeleidingsethiek.nl/cases/ [accessed 2024-02-21]
23. Friedman B. Value-sensitive design. Interactions 1996 Dec;3(6):16-23. [doi: 10.1145/242485.242493]
24. Winkler T, Spiekermann S. Twenty years of value sensitive design: a review of methodological practices in VSD projects. Ethics Inf Technol 2018 Aug 21;23(1):17-21. [doi: 10.1007/s10676-018-9476-2]
25. Fraaije A, Flipse SM. Synthesizing an implementation framework for responsible research and innovation. J Responsible Innov 2019 Nov 21;7(1):113-137. [doi: 10.1080/23299460.2019.1676685]
26. van der Wilt GJ, Reuzel R, Grin J. Design for values in healthcare technology. In: van den Hoven J, Vermaas P, van de Poel I, editors. Handbook of Ethics, Values, and Technological Design. Dordrecht, The Netherlands: Springer; 2015.
27. Innovatie-impuls. Ministerie van Volksgezondheid, Welzijn en Sport. URL: https://www.volwaardig-leven.nl/projecten/innovatie-impuls [accessed 2023-01-10]
28. van der Poel A, Vos N, Buimer H, ter Stal M, van der Weegen S, Gielissen M, et al. Duurzame implementatie van technologie in de gehandicaptenzorg: over deelnemers, vraagstukken en ervaringen uit de kwartiermakersfase van de innovatie-impuls. NTZ. 2021. URL: https://www.ntzonline.nl/media/3/ntzt202102_p66-75_vdpoel_ea_duurzame.pdf [accessed 2024-02-21]

29. Eindrapportage innovatie-impuls gehandicaptenzorg 2019-2022. De Rijksoverheid. URL: https://www.rijksoverheid.nl/documenten/rapporten/2023/01/27/eindrapportage-innovatie-impuls-gehandicaptenzorg-2019-2022 [accessed 2024-02-21]
30. van Dam K, Patel S, de Groot V, ter Stal M, Boon B, van der Poel A. Inzet van technologie die ondersteunt bij dagstructuur bij mensen met niet-aangeboren hersenletsel of een licht verstandelijke beperking: review van literatuur. LVB Onderzoek Praktijk Voorjaar 2023;21(1):13-23 [FREE Full text] [doi: 10.1007/978-90-368-0328-1_3]
31. Buimer HP, Siebelink NM, Gaasterland A, van Dam K, Smits A, Frederiks K, et al. Sleep-wake monitoring of people with intellectual disability: examining the agreement of EMFIT QS and actigraphy. J Appl Res Intellect Disabil 2023 Nov;36(6):1276-1287. [doi: 10.1111/jar.13146] [Medline: 37489295]
32. van Dam K, Gielissen M, Bles R, van der Poel A, Boon B. The impact of assistive living technology on perceived independence of people with a physical disability in executing daily activities: a systematic literature review. Disabil Rehabil Assist Technol 2023 Jan 10:1-10 Forthcoming. [doi: 10.1080/17483107.2022.2162614] [Medline: 36628433]
33. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: https://creativecommons.org/licenses/by/4.0/ [accessed 2024-02-26]
34. Gibson RC, Bouamrane MM, Dunlop M. Design requirements for a digital aid to support adults with mild learning disabilities during clinical consultations: qualitative study with experts. JMIR Rehabil Assist Technol 2019 Mar 04;6(1):e10449 [FREE Full text] [doi: 10.2196/10449] [Medline: 30829575]
35. Politi V, Grinbaum A. The distribution of ethical labor in the scientific community. J Responsible Innov 2020 Feb 16;7(3):263-279. [doi: 10.1080/23299460.2020.1724357]
36. Orlikowski WJ, Scott SV. Sociomateriality: challenging the separation of technology, work and organization. Acad Manag Ann 2008 Jan;2(1):433-474. [doi: 10.5465/19416520802211644]
37. Gezondheidsapps en wearables: de ethiek van e-health deel I. Centrum voor Ethiek en Gezondheid. URL: https://www.gezondheidsraad.nl/documenten/adviezen/2020/01/30/gezondheidsapps-en-wearables [accessed 2024-02-21]
38. Robotisering in de langdurige zorg – de ethiek van e-health deel 2. Centrum voor Ethiek en Gezondheid. URL: https://www.gezondheidsraad.nl/documenten/adviezen/2020/06/30/robotisering-in-de-langdurige-zorg [accessed 2024-02-21]
39. Snijders D, Horsman S, Kool L, van Est R. Verantwoord virtueel – bescherm consumenten in virtual reality. Rathenau Instituut. 2019. URL: https://www.rathenau.nl/sites/default/files/inline-files/Verantwoord%20virtueel.pdf [accessed 2024-02-21]
40. Hursthouse R, Pettigrove G. Virtue ethics. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy. Stanford, CA: Stanford University; 2016.
41. Smits M, Ludden GD, Verbeek PP, van Goor H. Responsible design and assessment of a SARS-CoV virtual reality rehabilitation programme: guidance ethics in context. J Responsible Innov 2022 May 24;9(3):344-370. [doi: 10.1080/23299460.2022.2076986]
42. Boenink M, Kudina O. Values in responsible research and innovation: from entities to practices. J Responsible Innov 2020 Aug 23;7(3):450-470. [doi: 10.1080/23299460.2020.1806451]
43. van Hees S, Greubel C, Moors E, Peine A. Valuation in health and ageing innovation practices. Ageing Soc 2021 Oct 27;43(9):2022-2040. [doi: 10.1017/s0144686x21001483]
44. Skär L, Söderberg S. The importance of ethical aspects when implementing eHealth services in healthcare: a discussion paper. J Adv Nurs 2018 May 29;74(5):1043-1050. [doi: 10.1111/jan.13493] [Medline: 29105151]
45. Keenan AJ, Tsourtos G, Tieman J. The value of applying ethical principles in telehealth practices: systematic review. J Med Internet Res 2021 Mar 30;23(3):e25698 [FREE Full text] [doi: 10.2196/25698] [Medline: 33783366]
46. Henni SH, Maurud S, Fuglerud KS, Moen A. The experiences, needs and barriers of people with impairments related to usability and accessibility of digital health solutions, levels of involvement in the design process and strategies for participatory and universal design: a scoping review. BMC Public Health 2022 Jan 06;22(1):35 [FREE Full text] [doi: 10.1186/s12889-021-12393-1] [Medline: 34991540]
47. Burke EA, Dennehy H, Bakker AR, Bowman S, Murphy E, Maes-Festen D, et al. The methodological approach to the co-creation of online health education with and for individuals with intellectual disability. Glob J Intellect Dev Disabil 2021 Mar 23;7(5):555725. [doi: 10.19080/GJIDD.2021.07.555725]
48. González HS, Córdova VV, Cid KE, Azagra MJ, Álvarez-Aguado I. Including intellectual disability in participatory design processes: methodological adaptations and supports. In: Proceedings of the 16th Participatory Design Conference 2020 - Participation(s) Otherwise - Volume 1. 2020 Presented at: PDC '20; June 15-20, 2020; Manizales, Colombia. [doi: 10.1145/3385010.3385023]
49. Boaz A, Hanney S, Borst R, O'Shea A, Kok M. How to engage stakeholders in research: design principles to support improvement. Health Res Policy Syst 2018 Jul 11;16(1):60 [FREE Full text] [doi: 10.1186/s12961-018-0337-6] [Medline: 29996848]
50. van Velsen L, Ludden G, Grünloh C. The limitations of user-and human-centered design in an eHealth context and how to move beyond them. J Med Internet Res 2022 Oct 05;24(10):e37341 [FREE Full text] [doi: 10.2196/37341] [Medline: 36197718]

51. Hung L, Mann J, Perry J, Berndt A, Wong J. Technological risks and ethical implications of using robots in long-term care. J Rehabil Assist Technol Eng 2022;9:20556683221106917 [FREE Full text] [doi: 10.1177/20556683221106917] [Medline: 35733613]
52. Yew GC. Trust in and ethical design of carebots: the case for ethics of care. Int J Soc Robot 2021;13(4):629-645 [FREE Full text] [doi: 10.1007/s12369-020-00653-w] [Medline: 32837630]
53. Perry J, Beyer S, Holm S. Assistive technology, telecare and people with intellectual disabilities: ethical considerations. J Med Ethics 2009 Feb 01;35(2):81-86. [doi: 10.1136/jme.2008.024588] [Medline: 19181877]
54. Cenci A, Cawthorne D. Refining value sensitive design: a (capability-based) procedural ethics approach to technological design for well-being. Sci Eng Ethics 2020 Oct;26(5):2629-2662 [FREE Full text] [doi: 10.1007/s11948-020-00223-3] [Medline: 32424724]
55. Owen R, Stilgoe J, Macnaghten P, Gorman M, Fisher E, Guston D. A framework for responsible innovation. In: Owen R, Bessant J, Heintz M, editors. Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society. Hoboken, NJ: John Wiley & Sons; 2013.
56. Turja T, Taipale S, Niemelä M, Oinas T. Positive turn in elder-care workers' views toward telecare robots. Int J Soc Robot 2022;14(4):931-944 [FREE Full text] [doi: 10.1007/s12369-021-00841-2] [Medline: 34873425]
57. Niemeijer AR, Frederiks BJ, Riphagen II, Legemaate J, Eefsting JA, Hertogh CM. Ethical and practical concerns of surveillance technologies in residential care for people with dementia or intellectual disabilities: an overview of the literature. Int Psychogeriatr 2010 Nov;22(7):1129-1142. [doi: 10.1017/S1041610210000037] [Medline: 20199699]

Abbreviations

ECP: Platform for the Information Society

Edited by J Torous; submitted 13.04.23; peer-reviewed by B Wernaart, J Hamid; comments to author 14.06.23; revised version received 26.10.23; accepted 12.01.24; published 28.02.24.

Please cite as:
Siebelink NM, van Dam KN, Lukkien DRM, Boon B, Smits M, van der Poel A
Action Opportunities to Pursue Responsible Digital Care for People With Intellectual Disabilities: Qualitative Study
JMIR Ment Health 2024;11:e48147
URL: https://mental.jmir.org/2024/1/e48147
doi:10.2196/48147
PMID:38416547

Corrigenda and Addenda

Correction: Digital Phenotypes for Early Detection of Internet Gaming Disorder in Adolescent Students: Explorative Data-Driven Study

Kwangsu Cho[1], PhD; Minah Kim[2,3], MD, PhD; Youngeun Cho[4], PhD; Ji-Won Hur[5], PhD; Do Hyung Kim[6], BSc; Seonghyeon Park[1], MSc; Sunghyun Park[2], MD; Moonyoung Jang[2,3], MS, MD; Chang-Gun Lee[6], PhD; Jun Soo Kwon[2,3], MD, PhD

[1]3R Innovation Research Center, Seoul, Republic of Korea
[2]Department of Neuropsychiatry, Seoul National University Hospital, Seoul, Republic of Korea
[3]Department of Psychiatry, Seoul National University College of Medicine, Seoul, Republic of Korea
[4]Department of Artificial Intelligence, Hanyang University, Ansan, Republic of Korea
[5]School of Psychology, Korea University, Seoul, Republic of Korea
[6]Department of Computer Science and Engineering, Seoul National University, Seoul, Republic of Korea

Corresponding Author:
Jun Soo Kwon, MD, PhD
Department of Neuropsychiatry
Seoul National University Hospital
101 Dahak-no
Jongno-gu
Seoul, 03080
Republic of Korea
Phone: 82 2 2072 2972
Email: kwonjs@snu.ac.kr

Related Article:

Correction of: https://mental.jmir.org/2024/1/e50259

(JMIR Ment Health 2024;11:e60568) doi:10.2196/60568

In “[Digital Phenotypes for Early Detection of Internet Gaming Disorder in Adolescent Students: Explorative Data-Driven Study]” ([JMIR Ment Health 2024;11:e50259 doi: 10.2196/50259]) the authors noted the following errors.

1. In the Results section of the Abstract, the sign in the following sentence:

Horizontal length of strokes (β= *–0.21)*

was corrected to:

Horizontal length of strokes (β*=0.21)*

2. In Figure 1, the sign of the coefficient of the Horizontal length of strokes:

–0.21

was corrected to:

0.21

3. In Table 2, the sign of the β[a] of the Horizontal length of strokes,

–0.21

was corrected to:

0.21

4. In Table 2, the *t* test value’s sign of the Horizontal length of strokes:

–2.59

was corrected to:

2.59

5. In Table 1, the column titles:

Potential IGD, Non-IGD

were corrected to:

Non-IGD, Potential IGD

The correction will appear in the online version of the paper on the JMIR Publications website on June 6, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 16.05.24; this is a non–peer-reviewed article;accepted 23.05.24; published 06.06.24.

Please cite as:
Cho K, Kim M, Cho Y, Hur JW, Kim DH, Park S, Park S, Jang M, Lee CG, Kwon JS
Correction: Digital Phenotypes for Early Detection of Internet Gaming Disorder in Adolescent Students: Explorative Data-Driven Study
JMIR Ment Health 2024;11:e60568
URL: https://mental.jmir.org/2024/1/e60568
doi:10.2196/60568
PMID:38843516

Corrigenda and Addenda

Correction: Digital Psychotherapies for Adults Experiencing Depressive Symptoms: Systematic Review and Meta-Analysis

Joanna Omylinska-Thurston[1], PsyD; Supritha Aithal[2], PhD; Shaun Liverpool[2], PhD; Rebecca Clark[1], BSc; Zoe Moula[3], PhD; January Wood[1], MSc; Laura Viliardos[1], PhD; Edgar Rodríguez-Dorans[4], PhD; Fleur Farish-Edwards[2], MSc; Ailsa Parsons[1], MSc; Mia Eisenstadt[5], PhD; Marcus Bull[6], MA; Linda Dubrow-Marshall[1], PhD; Scott Thurston[1], PhD; Vicky Karkou[2], PhD

[1]School of Health and Society, University of Salford, Manchester, United Kingdom
[2]Faculty of Health, Social Care and Medicine, Edge Hill University, Ormskirk, United Kingdom
[3]Faculty of Nursing, Midwifery and Palliative Care, King's College London, London, United Kingdom
[4]School of Health and Social Science, University of Edinburgh, Edinburgh, United Kingdom
[5]Evidence Based Practice Unit, University College London, London, United Kingdom
[6]Faculty of Life Sciences and Education, University of South Wales, Newport, United Kingdom

Corresponding Author:
Joanna Omylinska-Thurston, PsyD
School of Health and Society
University of Salford
Frederick Road Campus
Broad Street
Manchester, M6 6PU
United Kingdom
Phone: 44 0161 295 0000
Email: j.omylinska-thurston1@salford.ac.uk

Related Article:

Correction of: https://mental.jmir.org/2024/1/e55500

(JMIR Ment Health 2024;11:e67439) doi:10.2196/67439

In "Digital Psychotherapies for Adults Experiencing Depressive Symptoms: Systematic Review and Meta-Analysis" (JMIR Ment Health. 2024 Sep 30:11:e55500. doi: 10.2196/55500) the authors noted one error.

The surname of author JO-T has been revised from:

> *Omylinska Thurston*

To:

> *Omylinska-Thurston*

Furthermore, the author's two mentions in the body of the paper have been revised from:

> *JOT*

To:

> *JO-T*

The correction will appear in the online version of the paper on the JMIR Publications website on October 21, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 12.10.24; this is a non–peer-reviewed article;accepted 17.10.24; published 21.10.24.

Please cite as:
Omylinska-Thurston J, Aithal S, Liverpool S, Clark R, Moula Z, Wood J, Viliardos L, Rodríguez-Dorans E, Farish-Edwards F, Parsons A, Eisenstadt M, Bull M, Dubrow-Marshall L, Thurston S, Karkou V
Correction: Digital Psychotherapies for Adults Experiencing Depressive Symptoms: Systematic Review and Meta-Analysis
JMIR Ment Health 2024;11:e67439
URL: https://mental.jmir.org/2024/1/e67439
doi:10.2196/67439
PMID:

Impacts of Telehealth Adoption on the Quality of Care for Individuals With Serious Mental Illness: Retrospective Observational Analysis of Veterans Affairs Administrative Data

Camilla Cummings[1,*], PhD; Pushpa Raja[2,*], MSHPM, MD; Sonya Gabrielian[1,3,*], MPH, MD; Neal Doran[4,*], PhD

1

2

3

4

*all authors contributed equally

Corresponding Author:
Sonya Gabrielian, MPH, MD

Abstract

Background: Telehealth implementation can be challenging for persons with serious mental illness (SMI), which may impact their quality of care and health outcomes. The literature on telehealth's impacts on SMI care outcomes is mixed, necessitating further investigation.

Objective: We examined the impacts of facility-level telehealth adoption on quality of care metrics over time among patients with SMI.

Methods: We analyzed Veterans Affairs (VA) administrative data across 138 facilities from January 2021 to December 2022. We performed longitudinal mixed-effects regressions to identify the relationships between the proportion of facility-level telehealth visits and SMI specialty care quality metrics: engagement with primary care; access and continuity of care across a range of mental health services including psychotherapy or psychosocial rehabilitation, SMI-specific intensive outpatient programs, and intensive case management; and continuity of mental health care after a high-risk event (eg, suicide attempt).

Results: Facilities with a higher proportion of telehealth visits had reduced access and continuity of physical and mental health care for patients with SMI (P<.05). Higher telehealth adoption was associated with reduced primary care engagement (z=−4.04; P<.001), reduced access to and continuity in SMI-specific intensive case management (z=−4.49; P<.001; z=−3.15; P<.002), reductions in the continuity of care within psychotherapy and psychosocial rehabilitation (z=−3.74; P<.001), and continuity of care after a high-risk event (z=−2.46; P<.01). Telehealth uptake initially increased access to intensive outpatient but did not improve its continuity over time (z=−4.47; P<.001). Except for continuity within SMI-specific intensive case management (z=2.62; P<.009), continuity did not improve over time as telehealth became routinized.

Conclusions: Although telehealth helped preserve health care access during the pandemic, telehealth may have tradeoffs with regard to quality of care for some individuals with SMI. These data suggest that engagement strategies used by SMI-specific intensive case management may have preserved quality and could benefit other settings. Strategies that enhance telehealth implementation—selected through a health equity lens—may improve quality of care among patients with SMI.

(JMIR Ment Health 2024;11:e56886) doi:10.2196/56886

KEYWORDS

telemedicine; quality of care; serious mental illness; telehealth; adoption; mental illness; patients; patient; veterans; veteran; psychotherapy; psychosocial; mental healthcare; suicide; rehabilitation; mental health care

Introduction

Telehealth accelerated during the COVID-19 pandemic and is likely to persist. Although telehealth is accepted by most psychiatric patients, individuals with primary or co-occurring serious mental illness (SMI; ie, psychotic spectrum disorders and bipolar disorders) show lower engagement [1,2] and are less likely to use video-based services [3,4] than other diagnostic groups. A systematic review found that individuals with (vs without) SMI use telehealth at lower rates, despite much higher in-person mental health care (MH) use among individuals with SMI [2]. A study of patients with SMI who use telehealth during the pandemic found that, although many patients with SMI maintained engagement, individuals with schizophrenia were the least likely to engage in telehealth [3]. Despite lower telehealth engagement among patients with SMI (defined variably in the literature, but often equated to the aforementioned diagnostic categories [5]), few have evaluated the impact of telehealth expansion on this population's quality of care.

Literature on telehealth's impacts on care outcomes for individuals with SMI is mixed. Prepandemic, county-level research linked telehealth uptake to modest increases in the proportion of patients with SMI receiving minimum levels of MH. Though telehealth resulted in a greater likelihood of posthospitalization follow-up, no changes in medication adherence and slight increases in acute care use were seen [6]. During the first 6 months of the pandemic, rapid telehealth growth was associated with lower MH use among individuals with SMI [7]. A recent study of specialty mental health practices with higher telehealth use found that Medicare patients with SMI had more visits per year, but no differences in medication refills, postpsychiatric hospitalization follow-up, or all-cause mortality [8].

The Department of Veterans Affairs (VA), which serves more patients with SMI than any other US health care system, is an opportune setting to examine telehealth's impact on quality of care. VA maintains a comprehensive performance measurement system for outcome monitoring and quality improvement at national and facility levels [9]. This report analyzes national VA administrative data to characterize the impacts of facility-level telehealth adoption on SMI performance metrics, and interaction effects on quality metrics over time as telehealth was routinized.

Methods

Overview

National VA administrative data (from Corporate Data Warehouse) were used to extract all outpatient visits (across specialties) with SMI-specific *International Statistical Classification of Diseases, Tenth Revision* (*ICD-10*) codes as primary or secondary diagnostic codes. In total, 6 of 8 SMI-specific VA performance measures were examined; 2 measures had an expanded definition of SMI, encompassing visits with *ICD-10* codes for severe depression and posttraumatic stress disorder. Visits were stratified by modality (video telehealth vs face to face) and quarter from January 2021 to December 2022.

SMI-related performance metrics were collated across 138 VA facilities. Eight measures assessing access to care and care continuity for individuals with SMI were examined: (1) engagement with primary care (PC; ie, assigned to PC provider and ≥1 PC visit within the last year); (2) access to psychotherapy or psychosocial services (ie, proportion of patients with SMI who visited ≥1 psychotherapy or psychosocial rehabilitation within the last year); (3) continuity (ie, 5 psychotherapy or psychosocial rehabilitation visits within 10 weeks); access to care within 2 SMI specialty programs: (4) intensive outpatient programs (IOPs; ie, ≥3 visits in the past year); (5) intensive case management (ICM; VA's brand of Assertive Community Treatment; ie, ≥5 visits in the past year); continuity in these SMI specialty programs, with (6) ≥3 IOP visits over 90 days and (7) ≥12 ICM visits over 90 days; and (8) continuity of MH after a high-risk event like psychiatric hospitalization or suicide attempt (ie, ≥1 visit per 6 months over the past year).

Stata (version 15.0; StataCorp LLC) was used to perform longitudinal mixed-effects regressions evaluating whether facility unstandardized scores changed in relation to total volume of SMI telehealth visits. We controlled for facility size approximated by facility-level volume of SMI visits and time (as telehealth was routinized after the initial pandemic emergency). Separate models were fit for each measure. Models initially included linear and quadratic time to assess whether scores on quality measures varied across the 2-year period. Initial models included all 2-way interactions between linear and quadratic time and total SMI visits and telehealth SMI visits to evaluate whether associations between SMI visit volume and quality measures changed over time. Nonsignificant interactions were not retained.

Ethical Considerations

Procedures were approved by the VA San Diego Institutional Review Board (V201253). As an operations/quality improvement project, these data were collected in routine clinical care. Informed consent was not required, and no compensation was provided.

Results

No interactions were observed between time and SMI visit volume, and interactions were omitted. Table 1 displays our findings. In total, 7 of 8 measures (all but access to psychotherapy) were significantly associated ($P<.05$) with total SMI volume, telehealth SMI volume, or both.

Four measures were positively associated with facility total SMI volume but negatively associated with facility telehealth volume, including continuity of psychotherapy or psychosocial care ($z=-3.74$), PC engagement ($z=-4.04$), and ICM program access and continuity ($z=-4.49$; $z=-3.15$). Access to IOPs was negatively associated with total SMI visit volume but positively with telehealth volume. Total SMI visit volume was negatively associated with continuity of MH after high-risk events ($z=-2.46$) and positively associated with continuity of IOP ($z=2.24$); neither measure was associated with telehealth visit volume.

In total, 6 of 8 measures changed over time. Access to psychotherapy, PC engagement, and continuity of MH after high-risk events were positively associated with linear time ($z=4.39$ or 2.54 or 8.18) but negatively with quadratic time ($z=-3.78$ or −2.34 or −6.37); these 3 measures demonstrated an inverted U shape over time, with initial improvement followed by a decline back toward baseline. Continuity of psychotherapy ($z=-2.99$) and IOP ($z=-4.47$) declined significantly and linearly over time, whereas continuity of ICM ($z=2.62$) exhibited a linear increase over time.

Table . Summary of models testing associations between SMI[a] visits by modality and unstandardized scores on SMI-related quality measures across 138 Veterans Affairs facilities nationwide. *z* scores greater than 0 indicate a positive association between the quality measure and the time- or SMI visit volume–based predictor; *z* scores less than 0 reflect a negative association.

Quality measure	Definition	Linear time		Quadratic time		Total SMI visits[b]		Telehealth SMI visits	
		z score	*P* value	*z* score	*P* value	*z* score	*P* value	*z* score	*P* value
Access to psychotherapy	≥1 visit within past year	4.39	<.001	−3.78	<.001	−0.10	.92	0.46	.64
Continuity of psychotherapy	≥5 visits within 10 weeks past year	−2.99	.003	—[c]	—	6.15	<.001	−3.74	<.001
Access to primary care	Assigned primary care team + ≥1 primary care visit past 12 months	2.54	.01	−2.34	.02	2.12	.03	−4.04	<.001
Continuity of mental health care after high risk event	≥1 visit per 6 months in past year	8.18	<.001	−6.37	<.001	−2.46	.01	0.46	.65
Access to ICM[d]	≥5 visits within past year	−0.30	.76	—	—	2.22	.03	−4.49	<.001
Continuity of ICM	≥12 visits within 90 days	2.62	.009	—	—	3.88	<.001	−3.15	.002
Access to IOP[e]	≥3 visits past year	−0.46	.65	—	—	−4.62	<.001	2.66	.008
Continuity of IOP[f]	≥3 visits within 90 days	−4.47	<.001	—	—	2.24	.02	−1.58	.11

[a]SMI: serious mental illness.

[b]SMI encompassing schizophrenia and other psychotic disorders and bipolar disorders.

[c]Not applicable.

[d]ICM: intensive case management.

[e]IOP: intensive outpatient program.

[f]Measures for access to and continuity of IOP include a broader definition of SMI including severe depressive disorders and posttraumatic stress disorder.

Discussion

Principal Findings

In VA facilities with higher telehealth adoption, patients with SMI had diminished engagement with PC, lower access and care continuity in ICM programs, and reduced psychotherapy or psychosocial continuity. Although telehealth uptake initially increased access to IOP care for patients with SMI, IOP continuity was not improved. Larger sites, measured by total SMI volume, performed better in PC engagement, access and continuity for ICM, and continuity of psychotherapy, adjusted for telehealth use.

Higher telehealth adoption at the facility-level may have reduced access to physical and MH for patients with SMI. Although telehealth helped preserve health care access during COVID-19, these findings suggest potential tradeoffs in quality metrics for individuals with SMI, particularly in PC engagement and continuity of care in psychotherapy or psychosocial rehabilitation settings and ICM programs. Though others report that practice-level telehealth use was associated with more visits per year for individuals with SMI [8], a subset of patients with SMI may require more robust supports (eg, from peers or case managers) to use telehealth without reducing access. Training and technical support may be necessary to overcome barriers such as hardware and software needs, broadband access, technological illiteracy, and condition-related aversions to technology mediated communication (eg, delusions) [10,11].

Access to primary and MH as well as continuity of MH after a high-risk event initially improved but returned to baseline over time. Continuity of psychotherapy and IOP declined over time. These findings suggest quality metrics did not improve as service providers and patients habituated to telehealth. Perhaps flexibility with modality and platform decreased as telehealth was routinized, or patients open to telehealth as a temporary measure were less open to it long term. The only metric to improve over time was continuity of SMI-specific ICM, suggesting more intensive strategies (eg, frequent in-person case management visits) may have facilitated successful engagement. Such strategies could be scaled to support telehealth implementation for SMI populations in other settings.

In addition to condition-specific barriers to care, individuals with SMI face social and economic inequities, including lower socioeconomic status, education, employment, and higher rates of homelessness, affecting their ability to engage in telehealth [12]. Access and continuity of care are crucial for this group because of their impact on health outcomes and mortality [13]. Our data suggest that reliance on telehealth may exacerbate disparities in health care access and quality, raising concerns about health equity for this vulnerable population.

These analyses are limited by a focus on VA, which differs from community systems in services and patient population. VA's ICM and IOP programs have programmatic elements and performance metrics that may differ from other settings. These analyses used operations data intended for quality improvement; facility-level characteristics (eg, patient complexity) that could explain associations were not available. Future research could explore the impact of characteristics related to facility (eg, urbanicity or rurality), patients (eg, average age or facility), organization (eg, staffing), and provider attitudes regarding telehealth. Such investigations might assess the extent to which quality of care is impacted by telehealth enabling access to individuals previously disengaged from services altogether.

Conclusions

These findings raise concerns about facility-level impacts of telehealth adoption on quality of care and service engagement for individuals with SMI. Though telehealth is effective for persons with SMI [14], it may contribute to quality gaps for some individuals with SMI, negatively impacting facility-level quality metrics. Strategic telehealth supports are likely needed to reduce inequities for individuals with SMI, such as telehealth literacy screening [15], telehealth skills training [16], or in-person support to enable telehealth access. Given existing health disparities among individuals with SMI [10,12], centering health equity in the identification of implementation supports is warranted to ensure equitable and effective service provision for this vulnerable population.

Conflicts of Interest

None declared.

References

1. Matthews EB, Lushin V, Rzewinski J. Patterns & predictors of telehealth utilization among individuals who use substances: implications for the future of virtual behavioral health services. Community Ment Health J 2024 Jan;60(1):148-153. [doi: 10.1007/s10597-023-01166-2] [Medline: 37410213]
2. Neumann A, König HH, Bokermann J, Hajek A. Determinants of patient use and satisfaction with synchronous telemental health services during the COVID-19 pandemic: systematic review. JMIR Ment Health 2023 Aug 18;10:e46148. [doi: 10.2196/46148] [Medline: 37594785]
3. Bareis N, Tepper MC, Wang R, et al. Engagement of individuals with serious mental illness in outpatient mental health services and telehealth use during the COVID-19 pandemic. Psychiatry Res 2023 Nov;329:115497. [doi: 10.1016/j.psychres.2023.115497] [Medline: 37778232]
4. Raja PV, Gabrielian S, Doran N. Access to care for veterans with serious mental illness during the COVID-19 pandemic. Psychiatr Serv 2021 Nov 1;72(11):1324-1327. [doi: 10.1176/appi.ps.202000898] [Medline: 34030456]
5. VHA Handbook 1163.06. Department of Veterans Affairs. 2016. URL: https://www.va.gov/vhapublications/ViewPublication.asp?pub_ID=3164#:~:text=National%20Psychosis%20Registry.&text=The%20NPR%20is%20an%20ongoing,from%201988%20to%20the%20present [accessed 2023-10-25]
6. Wang B, Huskamp HA, Rose S, et al. Association between telemedicine use in nonmetropolitan counties and quality of care received by medicare beneficiaries with serious mental illness. JAMA Netw Open 2022 Jun 1;5(6):e2218730. [doi: 10.1001/jamanetworkopen.2022.18730] [Medline: 35759264]
7. Busch AB, Huskamp HA, Raja P, Rose S, Mehrotra A. Disruptions in care for medicare beneficiaries with severe mental illness during the COVID-19 pandemic. JAMA Netw Open 2022 Jan 4;5(1):e2145677. [doi: 10.1001/jamanetworkopen.2021.45677] [Medline: 35089352]
8. Wilcock AD, Huskamp HA, Busch AB, et al. Use of telemedicine and quality of care among medicare enrollees with serious mental illness. JAMA Health Forum 2023 Oct 6;4(10):e233648. [doi: 10.1001/jamahealthforum.2023.3648] [Medline: 37889483]
9. Lemke S, Boden MT, Kearney LK, et al. Measurement-based management of mental health quality and access in VHA: SAIL mental health domain. Psychol Serv 2017 Feb;14(1):1-12. [doi: 10.1037/ser0000097] [Medline: 28134552]

10. Kopelovich SL, Monroe-DeVita M, Buck BE, et al. Community mental health care delivery during the COVID-19 pandemic: practical strategies for improving care for people with serious mental illness. Community Ment Health J 2021 Apr;57(3):405-415. [doi: 10.1007/s10597-020-00662-z] [Medline: 32562033]
11. Clare CA. Telehealth and the digital divide as a social determinant of health during the COVID-19 pandemic. Netw Model Anal Health Inform Bioinform 2021 Apr;10(1):26. [doi: 10.1007/s13721-021-00300-y] [Medline: 33842187]
12. Shinn AK, Viron M. Perspectives on the COVID-19 pandemic and individuals with serious mental illness. J Clin Psychiatry 2020 Apr 28;81(3):20com13412. [doi: 10.4088/JCP.20com13412] [Medline: 32369691]
13. Jacobs R, Aylott L, Dare C, et al. The association between primary care quality and healthcare utilisation, costs and outcomes for people with serious mental illness: retrospective observational study. Health Serv Deliv Res 2020 Jun;8(25). [doi: 10.3310/hsdr08250]
14. Chakrabarti S. Usefulness of telepsychiatry: a critical evaluation of videoconferencing-based approaches. World J Psychiatry 2015 Sep 22;5(3):286-304. [doi: 10.5498/wjp.v5.i3.286] [Medline: 26425443]
15. Gillie M, Ali D, Vadlamuri D, Carstarphen KJ. Telehealth literacy as a social determinant of health: a novel screening tool to support vulnerable patient equity. J Alzheimers Dis Rep 2022 Feb;6(1):67-72. [doi: 10.3233/ADR-210024] [Medline: 35360275]
16. Camacho E, Torous J. Impact of digital literacy training on outcomes for people with serious mental illness in community and inpatient settings. Psychiatr Serv 2023 May 1;74(5):534-538. [doi: 10.1176/appi.ps.20220205] [Medline: 36164771]

Abbreviations

ICD-10**:** *International Statistical Classification of Diseases, Tenth Revision*
ICM: intensive case management
IOP: intensive outpatient program
MH: mental health care
PC: primary care
VA: Veterans Affairs

Edited by J Torous; submitted 29.01.24; peer-reviewed by A Rotondi, MC Tepper; revised version received 26.04.24; accepted 27.04.24; published 09.07.24.

Please cite as:
Cummings C, Raja P, Gabrielian S, Doran N
Impacts of Telehealth Adoption on the Quality of Care for Individuals With Serious Mental Illness: Retrospective Observational Analysis of Veterans Affairs Administrative Data
JMIR Ment Health 2024;11:e56886
URL: https://mental.jmir.org/2024/1/e56886
doi:10.2196/56886

Original Paper

Effectiveness of One Videoconference-Based Exposure and Response Prevention Session at Home in Adjunction to Inpatient Treatment in Persons With Obsessive-Compulsive Disorder: Nonrandomized Study

Ulrich Voderholzer[1,2,3], Prof Dr med; Adrian Meule[1,2,4], PD Dr; Stefan Koch[1], PhD; Simone Pfeuffer[1], Dr med; Anna-Lena Netter[5], PhD; Dirk Lehr[6], Prof Dr; Eva Maria Zisler[1,2], MSc

[1]Schoen Clinic Roseneck, Prien am Chiemsee, Germany

[2]Department of Psychiatry and Psychotherapy, University Hospital, LMU Munich, Munich, Germany

[3]Medical Center, Department of Psychiatry and Psychotherapy, Faculty of Medicine, University of Freiburg, Freiburg, Germany

[4]Institute of Medical Psychology, Faculty of Medicine, LMU Munich, Munich, Germany

[5]MindDoc Health GmbH, Munich, Germany

[6]Leuphana Universität Lüneburg, Lüneburg, Germany

Corresponding Author:
Eva Maria Zisler, MSc
Department of Psychiatry and Psychotherapy
University Hospital, LMU Munich
Nußbaumstraße 7
Munich, 80336
Germany
Phone: 49 15205886528
Email: Eva.Zisler@med.uni-muenchen.de

Abstract

Background: Therapist-guided exposure and response prevention (ERP) for the treatment of obsessive-compulsive disorder (OCD) is frequently conducted within clinical settings but rarely at places where patients are usually confronted with OCD symptom-provoking situations in daily life (eg, at home).

Objective: This study aimed to investigate patients' views on 1 ERP session at home via videoconference and its impact on treatment outcome.

Methods: A total of 64 inpatients with OCD received 1 session of therapist-guided videoconference-based ERP at home in adjunction to a multimodal inpatient treatment between 2015 and 2020.

Results: Compared with 64 age- and sex-matched controls who received a multimodal inpatient treatment without 1 session of videoconference-based ERP at home, patients who received 1 session of videoconference-based ERP in adjunction to a multimodal inpatient treatment showed stronger reductions in OCD symptom severity from admission to discharge. Before the videoconference-based ERP session, patients reported high rationale credibility and treatment expectancy. After the videoconference-based ERP session, patients reported medium-to-high positive mood as well as depth and smoothness of the session, and they perceived the working alliance as high.

Conclusions: Results highlight the importance of administering therapist-guided ERP sessions in patients' natural environment to enhance treatment response in OCD. Videoconference-based ERP as add-on to treatment as usual is, therefore, a promising approach to facilitate the application of ERP in patients' natural environment and foster the generalization of ERP conducted in clinical settings.

(JMIR Ment Health 2024;11:e52790) doi:10.2196/52790

KEYWORDS

obsessive-compulsive disorder; videoconference-based treatment; therapy; exposure; response prevention; OCD; prevention; inpatient; video; videoconference; therapist; therapists; mood; positive mood; environment; clinical setting

Introduction

Background

Obsessive-compulsive disorder (OCD) is a mental disorder characterized by intrusive and disturbing thoughts as well as repetitive patterns of behavior [1,2]. These are often multifaceted, that is, they include different obsessions and compulsions related to unwanted intrusive thoughts, fears of diseases, and contamination, among others [2,3]. OCD is a common disorder with a lifetime prevalence ranging from 1% to 3% and often has debilitating consequences on the daily functioning, well-being, and quality of life of affected persons as well as family members [4,5]. It usually emerges in late adolescence or early adulthood and has a chronic course if effective treatment is lacking [6,7]. Yet, OCD is often underrecognized and missed in primary care settings [8]. Thus, the duration of untreated illness in adults often exceeds 10 years, which creates a large treatment gap [9].

Exposure and Response Prevention in the Treatment of OCD

Cognitive-behavioral therapy (CBT) with exposure and response prevention (ERP) is the first-line, evidence-based psychotherapeutic treatment for OCD and is recommended as the psychotherapeutic method of choice [10-12]. ERP is a crucial element in CBT for OCD and requires patients to "engage in repeated, prolonged exposure to obsessions while refraining from compulsions" ([13], p. 85) [14]. Recent evidence stemming from inhibitory learning theory suggests that patients learn new associations during ERP (eg, "dirt—no fatal disease"), which in turn inhibit existing maladaptive associations (eg, "dirt—fatal disease" [15]). This acquisition of associations is enabled by expectancy violation that is a mismatch between patients' expectancy and outcome [15]. Although ERP is particularly useful in reducing OCD symptomatology, many patients find it difficult to endure upcoming unpleasant cognitions, feelings, and bodily sensations [16].

Besides the challenging nature of ERP itself, patients with OCD receiving CBT with ERP have to face a variety of difficulties [16]. First, patients are required to understand the underlying principles and measures of a treatment that is referred to as rationale credibility [17]. Second, patients need to expect that the treatment they are undergoing is effective [17-19]. Finally, it is beneficial if a positive working alliance is established between the patient and psychotherapist as it appears to predict treatment outcome [20]. Despite patients having to face various challenges when undergoing ERP, this psychotherapeutic intervention is highly effective for many people with OCD [21].

Home-Based ERP

The most commonly applied form of ERP is therapist-guided ERP in clinical settings (eg, at inpatient wards and in offices of psychotherapists), although the intervention can possibly be provided in several ways and facilities [22]. However, as persons with OCD often face the occurrence of obsessions and compulsions at home and feared situations or triggers cannot be replicated in a hospital or office, it can be hypothesized that home-based ERP may be beneficial in the treatment of OCD [23]. Although the theoretical framework of administering ERP at patients' homes may sound reasonable, evidence on this treatment variant is mixed. Although some studies found that ERP at home was slightly advantageous in terms of symptom reduction [24,25], others reported that home-based ERP was no more effective than standard office-based ERP [23].

There are a variety of reasons why home-based ERP is not administered on a regular basis by the majority of psychotherapists. Specifically, many clinicians lack time or familiarity with this intervention [22,26,27]. Additionally, specialized hospitals are not always located close to the patient's home, making therapist-guided ERP in the patient's living environment difficult to conduct. Even with outpatient therapy close to the patient's home, there is the challenge of therapists having to travel to the patient's place, which is difficult to implement due to limited time resources of therapists. Therefore, patients often receive outpatient treatment that only includes a limited number of therapist-assisted exposure sessions, if any [22,28].

To achieve a better care situation for patients with OCD, there are 2 cost-effective ways of implementing therapist-guided home-based ERP. The first option is telephone-supported ERP, which was shown to be effective in 2 studies [29,30]. With advancing technologies, the second option is videoconference-based ERP, which can also be considered an adequate tool that comes with significant reductions in obsessive-compulsive symptoms, especially in persons with moderate OCD symptoms [27,31-33]. Videoconference-based psychotherapy has several advantages over in-person psychotherapy. First, by using videoconference-based psychotherapy, treatment with ERP can easily be delivered to patients who are homebound or living in rural areas [31,34]. Second, the administration of home-based in vivo exposures allows the generalization of treatment effects to other contexts [31,34]. Third, the therapist is in charge of accompanying and supporting the patient during ERP [35]. Fourth, therapist-assisted ERP has been shown to be more effective than non–therapist-assisted ERP [36], and using videoconference at home might allow for even more therapist-assisted ERP.

This Study

As research on videoconference-based ERP is still limited, we examined treatment effects in patients who received inpatient treatment with an additional videoconference-based ERP at home compared with an age- and sex-matched group of patients who received inpatient treatment without an additional videoconference-based ERP at home. Second, we assessed patients' views on the current intervention (ie, treatment expectancy and rationale credibility) before undergoing the videoconference-based ERP session. Third, we examined patients' evaluations of the videoconference-based ERP session (ie, depth, smoothness, positivity, and arousal) and working alliance with the therapists after having received videoconference-based ERP. We expected stronger reductions in OCD symptom severity in patients who received inpatient treatment with an additional videoconference-based ERP session at home compared with inpatients who received multimodal inpatient treatment without an additional videoconference-based

ERP session at home from admission to discharge. Furthermore, we expected high ratings on rationale credibility and treatment expectancy before as well as high ratings on satisfaction with the therapeutic sessions and quality of the therapeutic relationship from the patients' perspective after the videoconference-based ERP session at home.

Methods

Sample Characteristics

This study was a nonrandomized, 2-group design study, in which a subset of patients who voluntarily participated in the study (videoconference exposure group) were compared with another subset of patients who did not participate in the study (control group). Although this design has disadvantages compared with a randomized controlled trial (RCT; see the *Discussion* section), it can be conducted more conveniently (eg, is less expensive and requires less resources) and may even have higher external validity as randomization may influence participation and outcomes when patients have a treatment preference [37]. Inpatients with OCD treated at the Schoen Clinic Roseneck (Prien am Chiemsee, Germany) between 2015 and 2020 were investigated. In Germany, inpatient treatment is indicated if at least 1 of the following factors applies: absence of or nonresponse to guideline-based disorder-specific outpatient treatment, danger to life, severe neglect, compulsive and avoidant behavior that is either severe or habitual resulting in an inability to maintain a normal daily routine and adherence to outpatient treatment, severe suffering and impairment of psychosocial functioning, psychological or somatic comorbidities aggravating outpatient treatment, and a particularly disease-promoting environment [10,38]. The treatment provided at the Schoen Clinic Roseneck adheres to the German S3 guidelines for the treatment of OCD [10]. Thus, the therapeutic concept is multimodal and consists of symptom-specific, individual CBT and ERP sessions, and other treatment elements, depending on indication (eg, psychopharmacological medication; see Table 1).

Table 1. Sample characteristics (N=128).

Characteristic	Videoconference exposure group (n=64)	Control group (n=64)	Test statistics				
			Chi-square (*df*)	*U*	*V*	*P* value	r_{rb} (*d*)
Subtype of obsessive-compulsive disorder (ICD-10[a] code), n (%)			1.62 (N/A[b])	N/A	0.11	.44	N/A
Obsessions-only subtype (F42.0)	0 (0)	1 (2)					
Compulsions-only subtype (F42.1)	8 (13)	11 (17)					
Mixed subtype (F42.2)	56 (88)	52 (81)					
Sex (female), n (%)	42 (66)	46 (72)	0.58 (N/A)	N/A	0.07	.45	N/A
Age (years), mean (SD)	26.95 (12.26)	29.28 (13.78)	N/A	1888.00	N/A	.45	–0.08 (–0.18)
Length of stay (days), mean (SD)	93.33 (30.54)	85.88 (38.40)	N/A	2302.50	N/A	.23	0.12 (0.22)
Any comorbid mental disorder, n (%)	44 (69)	43 (67)	0.04 (N/A)	N/A	0.02	.85	N/A
Affective disorders	32 (50)	34 (53)	0.13 (N/A)	N/A	0.03	.72	N/A
Anxiety disorders	11 (17)	16 (25)	1.17 (N/A)	N/A	0.10	.28	N/A
Eating disorders	6 (9)	4 (6)	0.43 (N/A)	N/A	0.06	.51	N/A
Antidepressant medication[c], n (%)	28 (49)	23 (51)	0.04 (N/A)	N/A	0.02	.84	N/A
Total score of Obsessive-Compulsive Inventory—Revised at admission, mean (SD)	31.56 (12.62)	31.32 (14.89)	N/A	2095.00	N/A	.83	0.02 (0.02)
Total score of Yale-Brown Obsessive-Compulsive Scale at admission, mean (SD)	23.63 (5.33)	22.57 (6.51)	N/A	2202.50	N/A	.46	0.08 (0.18)

[a]ICD-10: International Classification of Diseases, Tenth Revision.

[b]N/A: not applicable.

[c]Information missing for 7 patients in the videoconference exposure group and 19 patients in the control group.

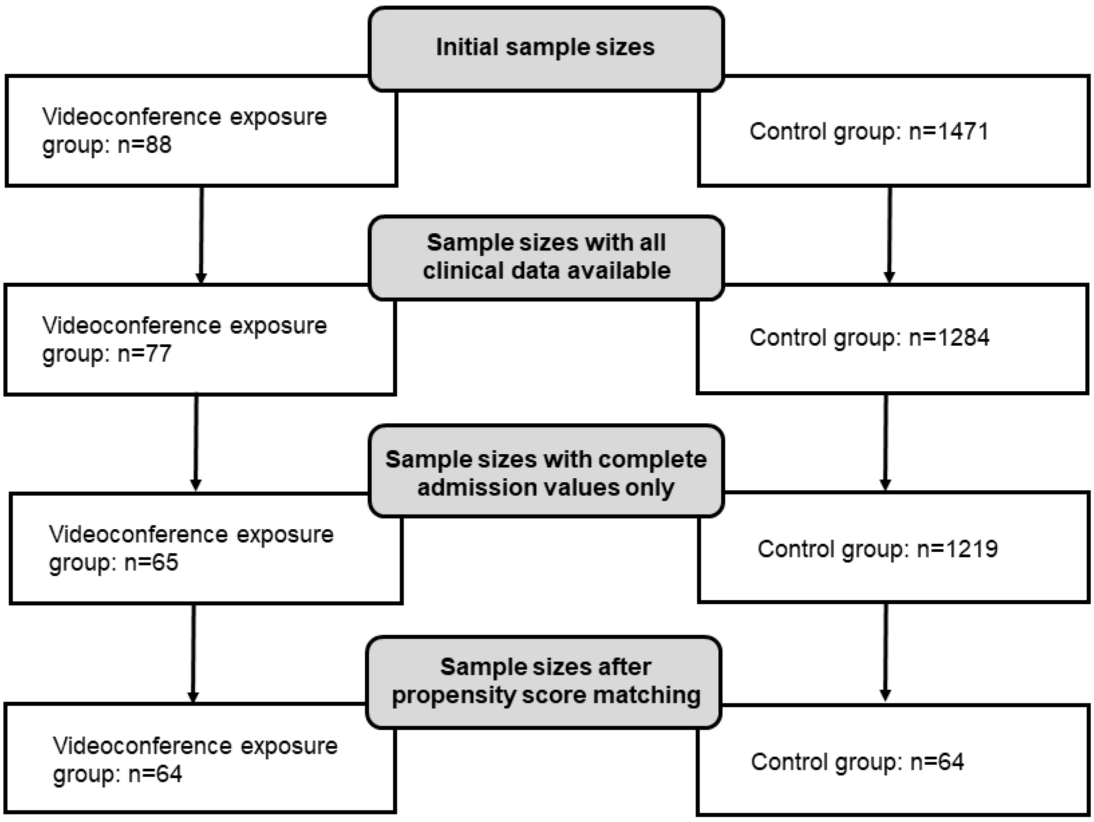
A total of 88 inpatients participated in this study, that is, received 1 videoconference-based ERP session at home in addition to inpatient treatment. As inpatient treatment at the Schoen Clinic Roseneck consists of 3 phases (psychoeducation and motivation, ERP, and transfer to the patients' homes), participating patients were in the third phase of inpatient treatment. Psychotherapists at the hospital who had undergone technical training on videoconference-based ERP were authorized to offer the intervention to their patients. Patients were free to choose whether or not to receive the additional videoconference-based

ERP session at home. On average, persons who received videoconference-based ERP at home had moderate OCD symptom severity according to the self-report version of the Yale-Brown Obsessive-Compulsive Scale (Y-BOCS; mean sum score 23.63, SD 5.33; Table 1; see recommendations by Cervin et al [39]).

Inpatients with OCD who were treated at the hospital within the same time period but who did not receive a videoconference-based ERP session at home were selected as the control group. Yet, these patients also received therapist-guided ERP in the hospital. Similar to the persons having received the videoconference-based ERP session, persons in the control group had, on average, moderate symptom severity according to the Y-BOCS (mean sum score 22.57, SD 6.51; Table 1; see recommendations by Cervin et al [39]). At the Schoen Clinic Roseneck, data from diagnostic assessments (eg, age, sex, diagnoses, medication, length of stay, and questionnaire scores) are automatically transferred to a database from which they can be exported without any identifying information by authorized employees. Thus, accessing individual patient charts is not necessary.

Between 2015 and 2020, a total of 1471 patients with OCD were treated in the hospital who did not receive videoconference-based ERP at home, that is, did not take part in the study. Because of missing data, 1219 patients were available for matching with 65 of the 88 patients in the videoconference exposure group (Figure 1). Groups were matched based on propensity score matching without replacement using the FUZZY extension for SPSS (version 27.0; IBM Corp) [40]. Data were matched in regard to the variables age, sex, any comorbidity, length of stay, Obsessive-Compulsive Inventory–Revised (OCI-R) scores at admission, and Y-BOCS scores at admission. Using a match tolerance with which all 65 persons in the videoconference exposure group were retained did not result in well-matched groups (ie, groups still differed in age and length of stay). Thus, a match tolerance of 0.019 was chosen, which led to the exclusion of 1 person from the videoconference exposure group, resulting in a final sample size of 128 (ie, 64 persons per group; Table 1).

Figure 1. Participant flowchart.



Measures

OCI-R Questionnaire

The OCI-R [41,42] was used to examine obsessive-compulsive symptoms. The OCI-R is an 18-item self-report questionnaire with 6 subscales: washing, checking, ordering, obsessing, hoarding, and neutralizing. Responses are recorded on a 5-point scale ranging from 0 (not at all) to 4 (extremely) and refer to the extent of distress during the past month due to OCD symptoms. In a previous study, internal reliability coefficients for the 6 subscales ranged between α=.76 and .95. In this study, the internal reliability coefficient for the total scale was ω=0.82 at admission and ω=0.86 at discharge.

Y-BOCS Questionnaire

The self-report version [43] of the Y-BOCS [44,45] was used to examine OCD severity. The Y-BOCS is a 10-item self-report questionnaire comprising 2 subscales: obsessions and compulsions. Responses are recorded on a 5-point scale ranging from 0 (no symptoms) to 4 (extreme symptoms). Internal reliability coefficients ranged between α=.78 and .88 in 2 validation studies [46,47] and between ω=0.83 and 0.91 in this study. Convergent validity has been supported by high correlations with other measures for obsessive-compulsive symptomatology, and divergent validity has been supported by moderate correlations with measures for related but distinct constructs such as worry [48-50].

Credibility Expectancy Questionnaire

The Credibility Expectancy Questionnaire (CEQ) [17] was used to assess the rationale credibility and treatment expectancy of the patient. The CEQ is a 6-item self-report questionnaire with 2 subscales: rationale credibility and treatment expectancy. Responses are recorded on a 9-point scale ranging from 1 (not at all) to 9 (very much). Internal reliability coefficients for the subscales ranged between ω=0.71 and 0.88.

Session Evaluation Questionnaire

The Session Evaluation Questionnaire (SEQ) [51] was used to examine the patients' satisfaction with the therapeutic sessions. The SEQ is a 21-item self-report questionnaire with 4 subscales: depth, smoothness, positivity, and arousal. Responses are recorded on a 7-point scale ranging from 1 (unpleasant) to 7 (pleasant). Internal reliability coefficients for the subscales ranged between ω=0.61 and 0.87. A closer inspection revealed that 2 items (1=slow, 7=fast; 1=moved, 7=composed) contributed to a low internal reliability of the arousal subscale. After removing those items, the remaining items of the arousal subscale had an internal reliability of ω=0.76. Thus, internal reliability coefficients for the subscales then ranged between ω=0.76 and 0.87.

Working Alliance Inventory—Short Revised

The Working Alliance Inventory—Short Revised (WAI-SR) [52] was used to examine the quality of the therapeutic relationship from the patient's perspective. The WAI-SR is a 12-item self-report questionnaire with 3 subscales: task, goal, and bond. Responses are recorded on a 7-point scale ranging from 1 (never) to 7 (always). Internal reliability coefficients for the subscales and the total scale ranged between ω=0.84 and 0.88.

Procedure

The videoconference app "VidyoMobile" by Vidyo, Inc was used to enable visual and auditory communication between the patient and therapist [25]. Patients were taught by a research staff member on how to use the smartphone, the tripod, and the videoconference app. Moreover, therapists prepared ERP sessions with their patients in close detail in a preceding session in the hospital. Before the ERP session, patients completed the CEQ. Patients received 1 videoconference-based ERP session each at home either on Friday afternoon or Monday morning. Each session had a duration of 2 hours on average. All videoconference-based ERP sessions were conducted by therapists specialized in CBT and ERP, and only the patient and the therapist were attending the session. The primary goal of the videoconference-based ERP session was to practice difficult situations associated with obsessions and compulsions in the patient's home. The therapist's role was to encourage the patient to face upcoming unpleasant feelings, emotions, and bodily sensations and to accompany them emotionally [16]. The exact execution of actions during ERP (ie, turning off the stove without checking, washing hands only once, etc) was not controlled by the therapist so as to give the patient a sense of personal responsibility in their own home.

After the ERP session, patients completed the SEQ and WAI-SR. In addition, after the videoconference-based ERP session, patients were asked to continue practicing the exposure exercise on their own. These exercises were not accompanied by the therapist, but debriefing followed in subsequent therapy sessions. Questionnaires assessing symptom severity (ie, OCI-R and Y-BOCS) were completed by the patients at admission and discharge.

Data Analyses

Group differences on categorical variables (OCD subtype, sex, comorbid mental disorders, and antidepressant medication) were tested with χ^2 tests and on continuous variables (age, length of stay, and questionnaire scores at admission) with Mann-Whitney *U* tests. Due to missing data at discharge (OCI-R: n=28, Y-BOCS: n=26), we examined changes of OCI-R and Y-BOCS total scores from admission to discharge as a function of a group with robust linear mixed models, which include cases with missing data in the maximum likelihood estimation. For this, we used R [53] and RStudio [54] and, specifically, the R package *robustlmm* [55]. The 2 models (1 for OCI-R scores and 1 for Y-BOCS scores) included fixed effects of time (admission vs discharge), group (videoconference exposure group vs control group), and their interaction term as well as a random intercept (ie, person-level random variability in scores at admission). As the package *robustlmm* does not produce parameter-specific *P* values, we used the workaround by Geniole et al [56]. Specifically, nonrobust models were fitted with the *lme4* package [57], *P* values were obtained with the package *lmertest* [58], and Satterthwaite-approximated degrees of freedom generated by the *lme4* models were combined with the output of the *robustlmm* model [56,59].

Ethical Considerations

The study was approved by the ethics committee of the Psychological Department of the Philipps University of Marburg, Germany. According to the guidelines by the institutional review board of the LMU Munich, retrospective analyses on already available anonymized data are exempt from requiring ethics approval. All participants in the videoconference exposure group signed informed consent before taking part in the study.

Results

As can be seen in Table 1, both groups did not significantly differ in age, sex, having any comorbid mental disorder, OCD subtype, antidepressant medication, OCI-R total scores at admission, and Y-BOCS total scores at admission. Robust linear mixed models revealed statistically significant interactions for group time for OCI-R (*b*=6.27; *P*=.01) and Y-BOCS (*b*=4.58; *P*<.001) scores, indicating that OCD symptom changes from admission to discharge differed as a function of group. As can be seen in Figures 2 and 3, the videoconference exposure group had larger OCD symptom reductions from admission to discharge than the control group. Descriptive statistics for obsessive-compulsive symptoms (total scores for OCI-R and Y-BOCS) at admission and discharge in the videoconference exposure and control groups are displayed in Table 2. On a scale ranging from 1 to 9, patients had mean (SD) values of 8.03 (0.74) on the subscale *rationale credibility* and 7.24 (1.13) on the subscale *treatment expectancy* on the CEQ. On a scale ranging from 1 to 7, patients had mean (SD) values of 5.87 (0.97) on the subscale *depth*, 3.60 (1.29) on the subscale *smoothness*, 4.61 (1.43) on the subscale *positivity*, and 4.11 (1.30) on the (reduced) subscale *arousal* on the SEQ. On a scale ranging from 1 to 7, patients had mean (SD) values of 6.25 (0.65) on the subscale *therapeutic tasks*, 6.52 (0.60) on the subscale *therapeutic goals*, 6.34 (0.75) on the subscale *therapeutic bond*, and 6.37 (0.57) on the total scale of the WAI-SR.

Figure 2. Mean sum scores of the Obsessive-Compulsive Inventory—Revised at admission and discharge as a function of group. The error bars indicate the SE of the mean. Effect sizes (rank biserial correlation coefficients r_{rb} and Cohen d) refer to the changes within each group from admission to discharge.

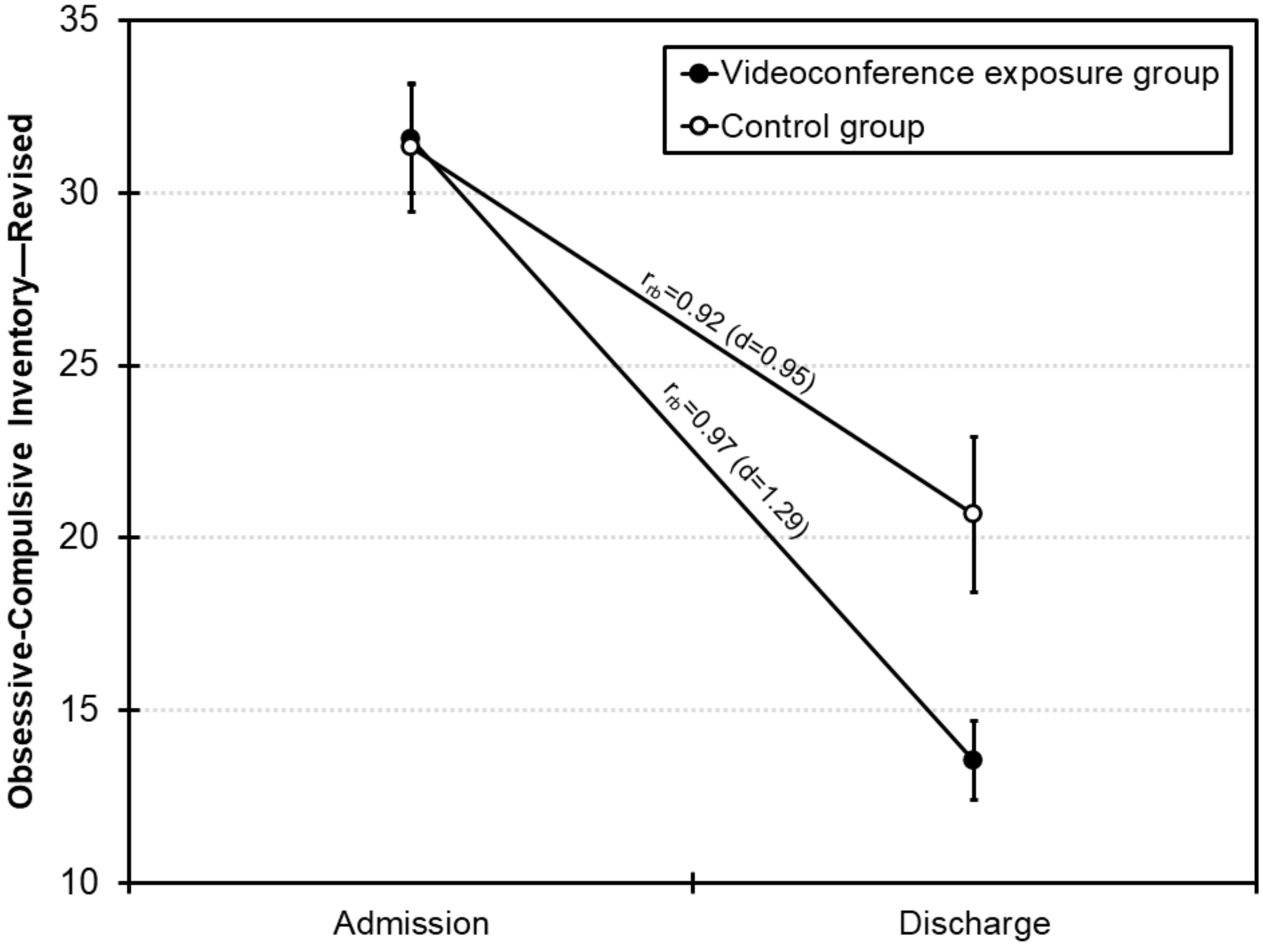


Figure 3. Mean sum scores of the Yale-Brown Obsessive-Compulsive Scale at admission and discharge as a function of group. The error bars indicate the SE of the mean. Effect sizes (rank biserial correlation coefficients r_{rb} and Cohen d) refer to the changes within each group from admission to discharge.

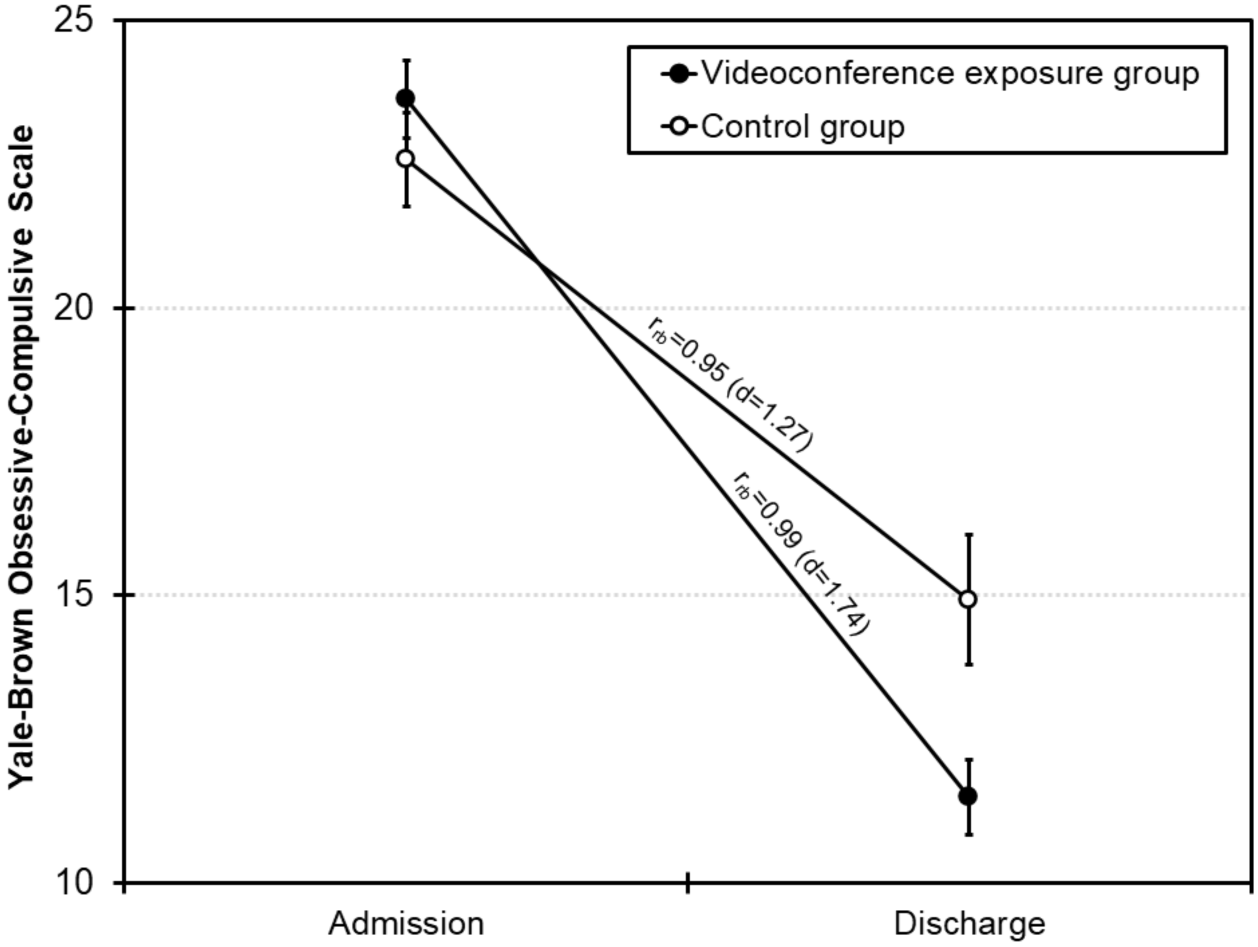


Table 2. Descriptive statistics for obsessive-compulsive symptoms at admission and discharge in the videoconference exposure and control groups.

Time point and statistic	Videoconference exposure group			Control group		
	n (%)	Mean (SD)	Range	n (%)	Mean (SD)	Range
Admission						
Obsessive-Compulsive Inventory—Revised	64 (50)	31.56 (12.62)	6-56	64 (50)	31.32 (14.89)	3-59
Yale-Brown Obsessive-Compulsive Scale	64 (50)	23.63 (5.33)	9-36	64 (50)	22.57 (6.51)	5-35
Discharge						
Obsessive-Compulsive Inventory—Revised	57 (44.5)	13.54 (8.59)	1-35	43 (33.6)	20.67 (14.81)	3-56
Yale-Brown Obsessive-Compulsive Scale	56 (43.8)	11.48 (4.88)	1-25	46 (35.9)	14.92 (7.67)	1-31

Discussion

Summary of Results

This study showed that the group that had an additional therapist-guided, videoconference-based ERP session at home showed greater improvements during inpatient treatment for OCD, that is, displayed larger decreases in OCD symptomatology compared with treatment as usual. Obsessive-compulsive symptoms from admission to discharge decreased for patients who received a videoconference-based ERP session at home as well as for patients who received treatment as usual without a videoconference-based ERP session with medium to large effect sizes. Yet, obsessive-compulsive symptoms decreased even stronger for patients who have received inpatient treatment and a videoconference-based ERP session as an add-on. Furthermore, patients had high treatment expectancy and perceived the rationale as credible before receiving videoconference-based ERP. After undergoing videoconference-based ERP, patients perceived depth (ie, potency and value), smoothness of the session (ie, comfort and relaxation), and mood after the session (ie, positivity and arousal) as medium to high. Patients who received videoconference-based ERP rated working alliance (ie, agreement on therapeutic tasks and goals as well as therapeutic bond) with their therapist as high.

Possible Mechanisms of Videoconference-Based ERP-Enhanced Symptom Reductions

Our results revealed that patients who received videoconference-based ERP at home in adjunction to a multimodal inpatient treatment had higher symptom reductions from admission to discharge with higher effect sizes than the control group. This might be significantly attributable to patients being able to generalize and extend their progresses achieved in the hospital to their own home; that is, with the help of the personal support of their therapist, they are more successful in giving up avoidance behavior at home as well [60]. Yet, it must be considered that there was no randomization in this study, which is why factors other than the additional videoconference-based ERP session might have also contributed to the reduction in OCD symptomatology from admission to discharge.

Alternative explanations for higher OCD symptom reductions in the videoconference exposure group might be that mostly patients who were highly motivated decided to participate in the additional videoconference-based ERP session or that the psychotherapists who treated patients receiving videoconference-based ERP were more motivated compared with other psychotherapists who treated the other patients with traditional ERP in the hospital only. Additionally, it might be possible that psychotherapists themselves expected that the additional ERP session at home would be beneficial for the patients and, thus, were highly engaged in the therapeutic sessions in the hospital as well, which particularly helped patients in reducing their OCD symptoms.

Despite methodological restrictions in nonrandomized study designs such as this study, there are also several disadvantages in RCTs that must be taken into account. First, participants are no passive recipients of interventions and do have treatment preferences. Patients with specific treatment preferences might, thus, refuse to take part in RCTs to avoid being randomized to the nonpreferred treatment, which reduces external validity [37]. Second, patients included in RCTs are strongly preselected, which was not the case in this study. Thus, the characteristics of patients included in this study correspond more to the real care situation. Third, internal validity of RCTs could be reduced as randomization to the nonpreferred treatment might influence patient adherence to the treatment protocol [37]. Accordingly, as this study was a nonrandomized study, patients were able to express and act on their treatment preferences as they could choose to receive the additional videoconference-based ERP session at home. This might have substantially increased patient adherence, which could, in turn, have been a factor contributing to reductions in OCD symptomatology. Furthermore, the 2 groups in this study were matched based on propensity score matching, which aims to account for absent randomization as it imitates some of the characteristics of an RCT [61]. Propensity score matching helps to strengthen causal arguments in observational studies by reducing selection bias [62].

Besides significant reductions in OCD symptomatology from admission to discharge in patients in the videoconference exposure group, the current results indicate that patients mainly had positive views on the videoconference-based ERP session, which became apparent in positive subjective ratings of the sessions. The positive effects of the videoconference-based ERP session on OCD symptomatology might be due to several change factors (ie, treatment expectancy and working alliance) that appear to be targeted in the videoconference setting to a sufficient degree. Several studies have provided evidence that treatment expectancy and understanding of the underlying treatment rationale are powerful predictors of psychotherapy outcome in general [18]. Additionally, as patients rated working alliance in the videoconference setting as high, this might also substantially contribute to the effects shown in this study. Previous studies have already shown that the videoconference setting enables the patient and psychotherapist to establish a strong and stable working alliance that is comparable to that in traditional face-to-face treatment [63,64]. Several studies even highlight that a positive working alliance is predictive of substantial decreases in symptomatology [31]. Although this study cannot show causal associations between working alliance and symptom reductions, a positive working alliance might substantially be linked to improvements of the patients' condition in the face-to-face and videoconference setting.

Limitations

As in every study, interpretation of the current results is limited to the persons and methods investigated. First, the examination of obsessive-compulsive symptoms was based on self-report, and—although the instruments used (OCI-R and Y-BOCS) are characterized by high validity and reliability—future studies may include therapist-rated measurements (eg, Y-BOCS interview version, Clinical Global Impression-Improvement Scale, and Global Assessment of Functioning [65,66]) as the inclusion of multiple views on the patients' OCD symptomatology allows for an even more comprehensive evaluation. Second, due to limited material and human resources in the hospital, only a subset of patients treated at the hospital received an additional videoconference-based ERP session at home. Therefore, future studies might make the treatment available to a larger sample and replicate the effect. Third, future studies might examine the effects of multiple videoconference-based ERP sessions as the current add-on intervention included only 1 ERP session. Fourth, there was no randomization in this study. Hence, there might also be a number of factors other than the additional videoconference-based ERP session at home that might have contributed to significant OCD symptom reductions (eg, motivation to engage in ERP might have differed between the 2 groups and different therapists administered ERP sessions). Thus, conducting RCTs is recommended for future studies.

Conclusions

Altogether, this study showed that the group that received a 1-time home visit of videoconference-based ERP in adjunction to a multimodal inpatient treatment had greater improvements, that is, larger decreases in OCD symptomatology, during inpatient treatment of OCD compared with treatment as usual. In addition, patients' ratings showed that the videoconference setting as well as working alliance with therapists was largely perceived as pleasant. Overall, it is recommended to provide patients with OCD with therapist-guided ERP at home. If it is not possible to accompany the intervention in person due to time constraints or other issues, videoconference-based therapy is a promising alternative to facilitate the application of ERP in patients' natural environment and foster the generalization of treatment effects achieved in clinical settings.

Data Availability

The data sets generated during and/or analyzed during this study are available in the Open Science Framework repository (https://osf.io/pybhw/).

Conflicts of Interest

None declared.

References

1. Jelinek L, Moritz S, Heeren D, Naber D. Everyday memory functioning in obsessive-compulsive disorder. J Int Neuropsychol Soc 2006;12(5):746-749 [FREE Full text] [doi: 10.1017/S1355617706060899] [Medline: 16961955]
2. Abramowitz J, Reuman L. Obsessive compulsive disorder. In: Shackelford TK, Zeigler-Hill V, editors. Encyclopedia of Personality and Individual Differences. Cham: Springer International Publishing; 2009:3304-3306.
3. Thiel A, Thiel J, Oddo S, Langnickel R, Brand M, Markowitsch HJ, et al. Obsessive-compulsive disorder patients with washing symptoms show a specific brain network when confronted with aggressive, sexual, and disgusting stimuli. Neuropsychoanalysis 2014;16(2):83-96. [doi: 10.1080/15294145.2014.976649]
4. Moritz S, Rufer M, Fricke S, Karow A, Morfeld M, Jelinek L, et al. Quality of life in obsessive-compulsive disorder before and after treatment. Compr Psychiatry 2005;46(6):453-459 [FREE Full text] [doi: 10.1016/j.comppsych.2005.04.002] [Medline: 16275213]
5. Macy AS, Theo JN, Kaufmann SCV, Ghazzaoui RB, Pawlowski PA, Fakhry HI, et al. Quality of life in obsessive compulsive disorder. CNS Spectr 2013;18(1):21-33 [FREE Full text] [doi: 10.1017/S1092852912000697] [Medline: 23279901]
6. Bloch MH, Craiglow BG, Landeros-Weisenberger A, Dombrowski PA, Panza KE, Peterson BS, et al. Predictors of early adult outcomes in pediatric-onset obsessive-compulsive disorder. Pediatrics 2009;124(4):1085-1093 [FREE Full text] [doi: 10.1542/peds.2009-0015] [Medline: 19786445]
7. Skoog G, Skoog I. A 40-year follow-up of patients with obsessive-compulsive disorder. Arch Gen Psychiatry 1999;56(2):121-127 [FREE Full text] [doi: 10.1001/archpsyc.56.2.121] [Medline: 10025435]
8. Hirschtritt ME, Bloch MH, Mathews CA. Obsessive-compulsive disorder: advances in diagnosis and treatment. JAMA 2017;317(13):1358-1367 [FREE Full text] [doi: 10.1001/jama.2017.2200] [Medline: 28384832]
9. Albert U, Barbaro F, Bramante S, Rosso G, De Ronchi D, Maina G. Duration of untreated illness and response to SRI treatment in obsessive-compulsive disorder. Eur Psychiatry 2019;58:19-26 [FREE Full text] [doi: 10.1016/j.eurpsy.2019.01.017] [Medline: 30763828]
10. Voderholzer U, Favreau M, Rubart A, Staniloiu A, Wahl-Kordon A, Zurowski B, et al. Treatment of obsessive-compulsive disorders: recommendations of the revised S3 guidelines on obsessive-compulsive disorders [Article in German]. Nervenarzt 2022;93(7):678-687 [FREE Full text] [doi: 10.1007/s00115-022-01336-9] [Medline: 35763051]
11. National Collaborating Centre for Mental Health (Great Britain). Obsessive-Compulsive Disorder: Core Interventions in the Treatment of Obsessive-Compulsive Disorder and Body Dysmorphic Disorder. Leicester, London: British Psychological Society; 2006.
12. American Psychiatric Association, Koran LM, Hanna GL, Hollander E, Nestadt G, Simpson HB. Practice Guideline for the Treatment of Patients with Obsessive-compulsive Disorder. Arlington, VA: American Psychiatric Association; 2007.
13. Foa EB, Kozak MJ. Mastery of Obsessive-Compulsive Disorder: A Cognitive-Behavioral Approach Therapist Guide. USA: Oxford University Press; 2004.
14. Leeuwerik T, Caradonna G, Cavanagh K, Forrester E, Jones AM, Lea L, et al. A thematic analysis of barriers and facilitators to participant engagement in group exposure and response prevention therapy for obsessive-compulsive disorder. Psychol Psychother 2023;96(1):129-147 [FREE Full text] [doi: 10.1111/papt.12430] [Medline: 36302721]
15. Elsner B, Jacobi T, Kischkel E, Schulze D, Reuter B. Mechanisms of exposure and response prevention in obsessive-compulsive disorder: effects of habituation and expectancy violation on short-term outcome in cognitive behavioral therapy. BMC Psychiatry 2022;22(1):66 [FREE Full text] [doi: 10.1186/s12888-022-03701-z] [Medline: 35086513]

16. Reid AM, Garner LE, van Kirk N, Gironda C, Krompinger JW, Brennan BP, et al. How willing are you? willingness as a predictor of change during treatment of adults with obsessive-compulsive disorder. Depress Anxiety 2017;34(11):1057-1064 [FREE Full text] [doi: 10.1002/da.22672] [Medline: 28715850]
17. Devilly GJ, Borkovec TD. Psychometric properties of the credibility/expectancy questionnaire. J Behav Ther Exp Psychiatry 2000;31(2):73-86 [FREE Full text] [doi: 10.1016/s0005-7916(00)00012-4] [Medline: 11132119]
18. Lewin AB, Peris TS, Bergman RL, McCracken JT, Piacentini J. The role of treatment expectancy in youth receiving exposure-based CBT for obsessive compulsive disorder. Behav Res Ther 2011;49(9):536-543 [FREE Full text] [doi: 10.1016/j.brat.2011.06.001] [Medline: 21723534]
19. Kirsch V, Keller F, Tutus D, Goldbeck L. Treatment expectancy, working alliance, and outcome of trauma-focused cognitive behavioral therapy with children and adolescents. Child Adolesc Psychiatry Ment Health 2018;12:16 [FREE Full text] [doi: 10.1186/s13034-018-0223-6] [Medline: 29515647]
20. Hagen K, Solem S, Opstad HB, Vogel PA, Kennair LEO, Kvale G, et al. Therapist variability in the task/goal dimension of the early working alliance predicts outcome in exposure and response prevention treatment for obsessive-compulsive disorder. Clin Neuropsychiatry 2016;13(6):94-99 [FREE Full text]
21. Hezel DM, Simpson HB. Exposure and response prevention for obsessive-compulsive disorder: a review and new directions. Indian J Psychiatry 2019;61(Suppl 1):S85-S92 [FREE Full text] [doi: 10.4103/psychiatry.IndianJPsychiatry_516_18] [Medline: 30745681]
22. Herbst N, Voderholzer U, Stelzer N, Knaevelsrud C, Hertenstein E, Schlegl S, et al. The potential of telemental health applications for obsessive-compulsive disorder. Clin Psychol Rev 2012;32(6):454-466 [FREE Full text] [doi: 10.1016/j.cpr.2012.04.005] [Medline: 22705583]
23. Rowa K, Antony MM, Summerfeldt LJ, Purdon C, Young L, Swinson RP. Office-based vs home-based behavioral treatment for obsessive-compulsive disorder: a preliminary study. Behav Res Ther 2007;45(8):1883-1892 [FREE Full text] [doi: 10.1016/j.brat.2007.02.009] [Medline: 17442265]
24. Rosqvist J, Egan D, Manzo P, Baer L, Jenike MA, Willis BS. Home-based behavior therapy for obsessive-compulsive disorder: a case series with data. J Anxiety Disord 2001;15(5):395-400 [FREE Full text] [doi: 10.1016/s0887-6185(01)00071-8] [Medline: 11583072]
25. Netter AL, Hessler JB, Diedrich A, Lehr D, Berking M, Pfeuffer S, et al. Usability and naturalness of videoconference-based exposure and response prevention for obsessive-compulsive disorder at the patients' homes: a pilot study. J Cogn Psychother 2018;32(4):303-314 [FREE Full text] [doi: 10.1891/0889-8391.32.4.303] [Medline: 32746409]
26. Voderholzer U, Böhm K, Külz AK. Zur Versorgungsrealität der Zwangsstörungen in Deutschland - Versorgungssituation bei Zwangsstörungen. Notfall Hausarztmedizin 2009;35(02):80-84 [FREE Full text] [doi: 10.1055/s-0029-1213746]
27. Feusner JD, Farrell NR, Kreyling J, McGrath PB, Rhode A, Faneuff T, et al. Online video teletherapy treatment of obsessive-compulsive disorder using exposure and response prevention: clinical outcomes from a retrospective longitudinal observational study. J Med Internet Res 2022;24(5):e36431 [FREE Full text] [doi: 10.2196/36431] [Medline: 35587365]
28. Moritz S, Külz A, Voderholzer U, Hillebrand T, McKay D, Jelinek L. "Phobie à deux" and other reasons why clinicians do not apply exposure with response prevention in patients with obsessive-compulsive disorder. Cogn Behav Ther 2019;48(2):162-176 [FREE Full text] [doi: 10.1080/16506073.2018.1494750] [Medline: 30064295]
29. Lovell K, Fullalove L, Garvey R, Brooker C. Telephone treatment of obsessive-compulsive disorder. Behav Cogn Psychother 2000;28(1):87-91 [FREE Full text] [doi: 10.1017/s1352465800000084]
30. Taylor S, Thordarson DS, Spring T, Yeh AH, Corcoran KM, Eugster K, et al. Telephone-administered cognitive behavior therapy for obsessive-compulsive disorder. Cogn Behav Ther 2003;32(1):13-25 [FREE Full text] [doi: 10.1080/16506070310003639] [Medline: 16291531]
31. Vogel PA, Solem S, Hagen K, Moen EM, Launes G, Håland Å, et al. A pilot randomized controlled trial of videoconference-assisted treatment for obsessive-compulsive disorder. Behav Res Ther 2014;63:162-168 [FREE Full text] [doi: 10.1016/j.brat.2014.10.007] [Medline: 25461792]
32. Lisi DM, Hawley LL, McCabe RE, Rowa K, Cameron DH, Richter MA, et al. Online versus in-person delivery of cognitive behaviour therapy for obsessive compulsive disorder: an examination of effectiveness. Clin Psychol Psychother 2023 [FREE Full text] [doi: 10.1002/cpp.2908] [Medline: 37699581]
33. Milosevic I, Cameron DH, Milanovic M, McCabe RE, Rowa K. Face-to-face versus video teleconference group cognitive behavioural therapy for anxiety and related disorders: a preliminary comparison. Can J Psychiatry 2022;67(5):391-402 [FREE Full text] [doi: 10.1177/07067437211027319] [Medline: 34159838]
34. Schiavone E, Freeman J, O'Connor E. Delivering exposure therapy via telehealth: benefits and challenges. Brown Univ Child Adolesc Behav Lett 2021;37(3):1-6 [FREE Full text] [doi: 10.1002/cbl.30527]
35. Hezel DM, Rapp AM, Glasgow S, Cridland G, Simpson HB. Year of zoom in a year of doom: lessons learned delivering ERP remotely during the COVID-19 pandemic. Cogn Behav Pract 2023;30(2):263-272 [FREE Full text] [doi: 10.1016/j.cbpra.2021.12.005] [Medline: 35228790]
36. Tolin DF, Hannan S, Maltby N, Diefenbach GJ, Worhunsky P, Brady RE. A randomized controlled trial of self-directed versus therapist-directed cognitive-behavioral therapy for obsessive-compulsive disorder patients with prior medication trials. Behav Ther 2007;38(2):179-191 [FREE Full text] [doi: 10.1016/j.beth.2006.07.001] [Medline: 17499084]

37. Wasmann KA, Wijsman P, van Dieren S, Bemelman W, Buskens C. Partially randomised patient preference trials as an alternative design to randomised controlled trials: systematic review and meta-analyses. BMJ Open 2019;9(10):e031151 [FREE Full text] [doi: 10.1136/bmjopen-2019-031151] [Medline: 31619428]
38. Wahl-Kordon A, Muche-Borowski C, Hohagen F, Lotz-Rambaldi W, editors. S3-Leitlinie Zwangsstörungen—Langversion. Germany: Springer Berlin Heidelberg; 2022.
39. Cervin M, OCD Severity Benchmark Consortium, Mataix-Cols D. Empirical severity benchmarks for obsessive-compulsive disorder across the lifespan. World Psychiatry 2022;21(2):315-316 [FREE Full text] [doi: 10.1002/wps.20984] [Medline: 35524607]
40. IBM SPSS Statistics for Windows (version 270) [computer software]. IBM Corp. 2020. URL: https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-27 [accessed 2024-01-23]
41. Foa EB, Huppert JD, Leiberg S, Langner R, Kichic R, Hajcak G, et al. The Obsessive-Compulsive Inventory: development and validation of a short version. Psychol Assess 2002;14(4):485-496. [Medline: 12501574]
42. Gönner S, Leonhart R, Ecker W. The German version of the Obsessive-Compulsive Inventory-Revised: a brief self-report measure for the multidimensional assessment of obsessive-compulsive symptoms [Article in German]. Psychother Psychosom Med Psychol 2007;57(9-10):395-404 [FREE Full text] [doi: 10.1055/s-2007-970894] [Medline: 17590836]
43. Schaible R, Armbrust M, Nutzinger DO. Yale-Brown Obsessive Compulsive Scale: sind selbst- und fremdrating äquivalent? Verhaltenstherapie 2002;11(4):298-303 [FREE Full text] [doi: 10.1159/000056673]
44. Baer L, Brown-Beasley MW, Sorce J, Henriques AI. Computer-assisted telephone administration of a structured interview for obsessive-compulsive disorder. Am J Psychiatry 1993;150(11):1737-1738 [FREE Full text] [doi: 10.1176/ajp.150.11.1737] [Medline: 8214187]
45. Baer L. Getting Control: Overcoming Your Obsessions, Compulsions and OCD. Boston: Little, Brown Book Group; 1991.
46. Federici A, Summerfeldt LJ, Harrington JL, McCabe RE, Purdon CL, Rowa K, et al. Consistency between self-report and clinician-administered versions of the yale-brown obsessive-compulsive scale. J Anxiety Disord 2010;24(7):729-733 [FREE Full text] [doi: 10.1016/j.janxdis.2010.05.005] [Medline: 20561767]
47. Rosenfeld R, Dar R, Anderson D, Kobak KA, Greist JH. A computer-administered version of the Yale-Brown Obsessive-Compulsive Scale. Psychological Assessment 1992;4(3):329-332. [doi: 10.1037/1040-3590.4.3.329]
48. Steketee G, Frost R, Bogart K. The Yale-Brown Obsessive Compulsive Scale: interview versus self-report. Behav Res Ther 1996;34(8):675-684 [FREE Full text] [doi: 10.1016/0005-7967(96)00036-8] [Medline: 8870295]
49. Ólafsson RP, Snorrason Í, Smári J. Yale-Brown Obsessive Compulsive Scale: psychometric properties of the self-report version in a student sample. J Psychopathol Behav Assess 2009;32(2):226-235 [FREE Full text] [doi: 10.1007/s10862-009-9146-0]
50. Rapp AM, Bergman RL, Piacentini J, McGuire JF. Evidence-based assessment of obsessive-compulsive disorder. J Cent Nerv Syst Dis 2016;8:13-29 [FREE Full text] [doi: 10.4137/JCNSD.S38359] [Medline: 27594793]
51. Stiles WB, Snow JS. Counseling session impact as viewed by novice counselors and their clients. J Couns Psychol 1984;31(1):3-12. [doi: 10.1037//0022-0167.31.1.3]
52. Hatcher RL, Gillaspy JA. Development and validation of a revised short version of the Working Alliance Inventory. Psychother Res 2006;16(1):12-25 [FREE Full text] [doi: 10.1080/10503300500352500]
53. R: a language and environment for statistical computing. RCore Team. Vienna, Austria: R Foundation for Statistical Computing; 2022. URL: https://www.r-project.org [accessed 2024-01-23]
54. RStudio Team. RStudio: Integrated Development for R. Boston, MA: PBC; 2022. URL: https://cir.nii.ac.jp/crid/1370004240607707919 [accessed 2024-01-23]
55. Koller M. robustlmm: an R package for robust estimation of linear mixed-effects models. J Stat Soft 2016;75(6):1-24 [FREE Full text] [doi: 10.18637/jss.v075.i06]
56. Geniole SN, Proietti V, Bird BM, Ortiz TL, Bonin PL, Goldfarb B, et al. Testosterone reduces the threat premium in competitive resource division. Proc Biol Sci 2019;286(1903):20190720 [FREE Full text] [doi: 10.1098/rspb.2019.0720] [Medline: 31138068]
57. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. ArXiv. Preprint posted online on June 23, 2014 2014 [FREE Full text] [doi: 10.18637/jss.v067.i01]
58. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: tests in linear mixed effects models. J Stat Soft 2017;82(13):1-26 [FREE Full text] [doi: 10.18637/jss.v082.i13]
59. Obtaining p-values in a robustlmm mixed model via satterthwaite-approximated DFs of the equivalent lme4 model? StackExchange. 2019. URL: https://stats.stackexchange.com/questions/430490/obtaining-p-values-in-a-robustlmm-mixed-model-via-satterthwaite-approximated-dfs [accessed 2024-01-23]
60. Steketee G. Treatment of Obsessive-Compulsive Disorder. New York, NY: Guilford; 1993.
61. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat Med 2008;27(12):2037-2049 [FREE Full text] [doi: 10.1002/sim.3150] [Medline: 18038446]
62. Randolph JJ, Falbe K, Manuel AK, Balloun JL. A step-by-step guide to propensity score matching in R. PARE 2014;19(1):18 [FREE Full text] [doi: 10.4135/9781071802854.n5]

63. Herbst N, Franzen G, Voderholzer U, Thiel N, Knaevelsrud C, Hertenstein E, et al. Working alliance in internet-based cognitive-behavioral therapy for obsessive-compulsive disorder. Psychother Psychosom 2016;85(2):117-118 [FREE Full text] [doi: 10.1159/000441282] [Medline: 26807976]
64. Seuling PD, Fendel JC, Spille L, Göritz AS, Schmidt S. Therapeutic alliance in videoconferencing psychotherapy compared to psychotherapy in person: a systematic review and meta-analysis. J Telemed Telecare 2023 Mar 28:1357633X231161774 [FREE Full text] [doi: 10.1177/1357633X231161774] [Medline: 36974478]
65. Guy W. ECDEU Assessment Manual for Psychopharmacology. Rockville, MD: U.S. Dept. of Health, Education, and Welfare, Public Health Service, Alcohol, Drug Abuse, and Mental Health Administration, National Institute of Mental Health, Psychopharmacology Research Branch, Division of Extramural Research Programs; 1976.
66. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders: DSM-IV. Washington DC: American Psychiatric Association; 1994.

Abbreviations

CBT: cognitive-behavioral therapy
CEQ: Credibility Expectancy Questionnaire
ERP: exposure and response prevention
OCD: obsessive-compulsive disorder
OCI-R: Obsessive-Compulsive Inventory—Revised
RCT: randomized controlled trial
SEQ: Session Evaluation Questionnaire
WAI-SR: Working Alliance Inventory—Short Revised
Y-BOCS: Yale-Brown Obsessive-Compulsive Scale

Edited by J Torous; submitted 15.09.23; peer-reviewed by L Jelinek, K Rowa; comments to author 11.10.23; revised version received 24.11.23; accepted 19.12.23; published 13.03.24.

Please cite as:
Voderholzer U, Meule A, Koch S, Pfeuffer S, Netter AL, Lehr D, Zisler EM
Effectiveness of One Videoconference-Based Exposure and Response Prevention Session at Home in Adjunction to Inpatient Treatment in Persons With Obsessive-Compulsive Disorder: Nonrandomized Study
JMIR Ment Health 2024;11:e52790
URL: https://mental.jmir.org/2024/1/e52790
doi:10.2196/52790
PMID:38477970

Original Paper

Design and Implementation of a Brief Digital Mindfulness and Compassion Training App for Health Care Professionals: Cluster Randomized Controlled Trial

Satish Jaiswal[1], PhD; Suzanna R Purpura[1], MS; James K Manchanda[1], BS; Jason Nan[1], BS; Nihal Azeez[1], BS; Dhakshin Ramanathan[2], PhD; Jyoti Mishra[1], PhD

[1]Department of Psychiatry, University of California San Diego, San Diego, CA, United States

[2]Department of Mental Health, Veterans Affairs San Diego Medical Center, San Diego, CA, United States

Corresponding Author:
Satish Jaiswal, PhD
Department of Psychiatry
University of California San Diego
9452 Medical Center Dr
La Jolla, CA
San Diego, CA, 92037
United States
Phone: 1 858 232 2855
Email: s2jaiswal@health.ucsd.edu

Abstract

Background: Several studies show that intense work schedules make health care professionals particularly vulnerable to emotional exhaustion and burnout.

Objective: In this scenario, promoting self-compassion and mindfulness may be beneficial for well-being. Notably, scalable, digital app–based methods may have the potential to enhance self-compassion and mindfulness in health care professionals.

Methods: In this study, we designed and implemented a scalable, digital app–based, brief mindfulness and compassion training program called "WellMind" for health care professionals. A total of 22 adult participants completed up to 60 sessions of WellMind training, 5-10 minutes in duration each, over 3 months. Participants completed behavioral assessments measuring self-compassion and mindfulness at baseline (preintervention), 3 months (postintervention), and 6 months (follow-up). In order to control for practice effects on the repeat assessments and calculate effect sizes, we also studied a no-contact control group of 21 health care professionals who only completed the repeated assessments but were not provided any training. Additionally, we evaluated pre- and postintervention neural activity in core brain networks using electroencephalography source imaging as an objective neurophysiological training outcome.

Results: Findings showed a post- versus preintervention increase in self-compassion (Cohen d=0.57; P=.007) and state-mindfulness (d=0.52; P=.02) only in the WellMind training group, with improvements in self-compassion sustained at follow-up (d=0.8; P=.01). Additionally, WellMind training durations correlated with the magnitude of improvement in self-compassion across human participants (ρ=0.52; P=.01). Training-related neurophysiological results revealed plasticity specific to the default mode network (DMN) that is implicated in mind-wandering and rumination, with DMN network suppression selectively observed at the postintervention time point in the WellMind group (d=–0.87; P=.03). We also found that improvement in self-compassion was directly related to the extent of DMN suppression (ρ=–0.368; P=.04).

Conclusions: Overall, promising behavioral and neurophysiological findings from this first study demonstrate the benefits of brief digital mindfulness and compassion training for health care professionals and compel the scale-up of the digital intervention.

Trial Registration: Trial Registration: International Standard Randomized Controlled Trial Number Registry ISRCTN94766568, https://www.isrctn.com/ISRCTN94766568

(JMIR Ment Health 2024;11:e49467) doi:10.2196/49467

KEYWORDS

compassion; digital app; digital health; digital intervention; digital mental health; digital mindfulness; EEG; health workers; healthcare professionals; mindfulness; neuroplasticity; physicians; training

Introduction

Health care professionals receive intensive hands-on education and training so they may serve as resilient healers. Yet, with high workloads and multitasking demands, physicians can become vulnerable to workplace stress while making critical life-altering decisions for patients [1,2]. Evidence shows high rates of physician burnout ("workplace stress that has not been successfully managed," *International Classification of Diseases*, World Health Organization), currently estimated at 44%, which is much higher than burnout in other professions [3,4]. Such chronic stress among physicians, in turn, may reduce the quality of physician-patient interactions and can lead to unprofessional behavior and attitudes [5,6]. Consistent with this, a study showed that burnout was associated with self-reported unprofessional conduct and less altruistic professional values among medical students at 7 US schools [7]. The COVID-19 pandemic also exacerbated distress in health care workers and has been shown to be associated with significant emotional pain, drop out from training, and reductions in work hours [8,9], which eventually affect the entire health care system. Hence, there is an emerging need for interventions that can help alleviate physician stress and prevent burnout [10].

Previous research has shown that once burnout develops, both individual-level and organizational strategies can result in clinically meaningful reductions in burnout [11], but organizational-level interventions (reducing work hours, increasing staffing, etc) have a larger overall effect size. [12] Notably, research has suggested that individual-level characteristics, particularly those related to mindfulness and self-compassion, may help to prevent burnout from developing [13-15]. Importantly, these are skills that can be developed. Research shows that when mindfulness and interoceptive-awareness exercises are provided as part of continuing medical education to primary care physicians, they help to reduce physician work-related stress and enhance well-being [11,16]. In a recent qualitative study among radio-oncologists, higher-trait mindfulness was found to be a protective factor from burnout and positively associated with life fulfillment [17]. Similarly, studies suggest that compassion training can promote well-being in medical students and improve the quality of clinical care [18]. Research has also shown that quality of care, health care costs, and the well-being of the clinician workforce are interlinked domains [19]. Hence, feasible, scalable, and effective mindfulness strategies delivered early may have far-reaching socioeconomic benefits for the health care system.

Compared to classroom approaches, digital training approaches can be highly scalable. Participants can flexibly engage with these trainings as per their convenience, so it is not a burden on busy work schedules [20]. Further, without the need for a teacher guide for digital training, the motivation for the practice is intrinsic and not reliant on teacher expertise. There are very few empirical studies among health care professionals that examine the efficacy of an app-based mindfulness program and its objective neural implications. A recent study in physician assistants showed a decrease in sleep dysfunction, enhanced connectivity between the medial prefrontal cortex and the superior temporal gyrus, as well as between regions critical for working memory after 8 weeks of intervention [21].

We have previously tested digital meditation approaches in adolescents and healthy young adults and shown positive neuro-cognitive outcomes [22,23]. Moreover, these digital trainings have integrated gamification, feedback, and rewards within closed-loop design systems to drive high user adherence [24].

In this study, we implemented a brief digital closed-loop training for health care professionals that involved an attention-to-breath practice with integrated compassion prompts. Respecting the time constraints of health professionals, each session provided 5-10 minutes of practice, and trainees had access to up to 60 digital sessions. Relative to a no-contact or business-as-usual control group, we evaluated the primary outcomes of change in mindfulness and self-compassion and monitored sessions of training engagement. While we additionally measured burnout, we did not expect this outcome to change with individual-focused training because it has been shown that organization-directed workplace interventions are more effective at addressing burnout [1,25] and burnout was also observed to be low in our sample. Thus, our main goal in this study was to examine whether mindfulness and compassion can be enhanced by brief digitally delivered practice sessions for health care professionals. Positive results from such a study may then serve as a rationale for future implementation integrated within an organizational framework to prevent burnout from developing.

Finally, this study also uniquely assayed objective neurophysiological plasticity associated with the training alongside subjective behavioral changes. For this, we measured electroencephalography (EEG)-based brain signals on an interoceptive attention-to-breathing task, assessed before and after training. The rationale for selecting such a task for neurophysiological measurements is that interoceptive attention to breathing is a core feature of several meditation practices [26,27]. On this task, we were particularly interested in neurophysiological activity within the default mode network (DMN), which has been shown to be modulated by meditation [28-31]. The DMN is a functional network that has not only been consistently associated with autobiographical memory and self-referencing but also on-task behavioral variability, mind-wandering, and rumination [32-37]. We hypothesized that mindfulness and compassion training would suppress DMN activity. Our ultimate analyses focused on whether objective modulation of the DMN relates to subjective behavioral changes in self-compassion and mindfulness.

Methods

Participants

A total of 43 human participants recruited in the study (mean age 28.77, SD 4.13; range 23-43 y; 20 male participants). All human participants were fluent in English. Participants were recruited from the University of California San Diego (UCSD) School of Medicine from Spring 2021 to Fall 2022 academic quarters through email advertisements and campus flyers.

Participants provided demographic data with regards to age, gender, and ethnicity. All participants were healthy adults, that is, they did not have any current medical diagnosis nor were taking any current psychotropic medications. Healthy status and affiliation to the UCSD School of Medicine were the only eligibility criteria.

Participants completed the Maslach Burnout Inventory (MBI) at the time of screening; MBI scores did not reflect high burnout in our sample as all scores were less than the midscore of the MBI score range (see the *Results* section).

Study Design

The study design was interventional and cluster randomized. Of the total 43 study participants, 22 were enrolled in the digital WellMind intervention group, and 21 were part of the no-contact control group. Participants were cluster randomized based on the academic quarter of enrollment to the WellMind or control group. Specifically, all participants recruited during Spring 2021, Spring 2022, and Fall 2022 academic quarters were assigned to the WellMind group, and the no-contact control group participants were recruited during Fall 2021. This was done because individuals within each academic quarter (but not across quarters) were working or studying together and, hence, knew each other professionally and could reveal components of the study intervention to each other. The WellMind group participants received the digital app intervention and had periodic email contact from our research team during the intervention, at about once every 2 weeks, to ensure compliance and help troubleshoot any issues faced by the participants. On the other hand, the no-contact control group had no interaction with the study research team or any digital training resource provided to them between their pre- and postintervention time points.

Sample Size and Power

The sample size within each group was powered to detect medium effect size for pre- or postintervention differences (Cohen d>0.6) at β power of 0.8 and α level of 0.05. Between-group differences met criteria for investigating only large effect size outcomes (Cohen d>0.8) at β power of 0.8 and α level of 0.05. Effect sizes were calculated a priori using the G*Power (Axel Buchner) software [38].

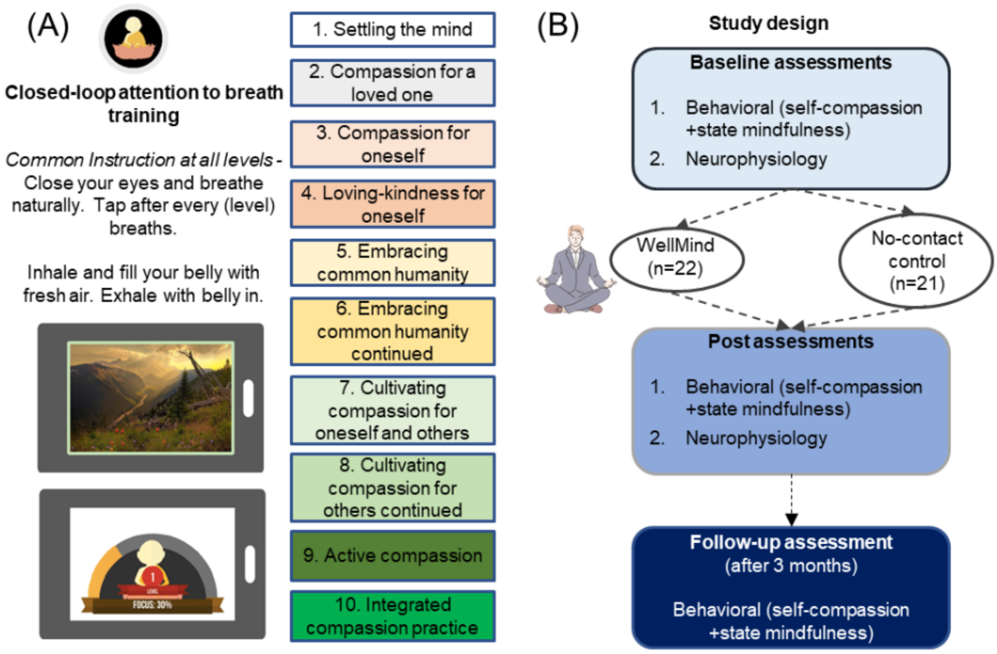
Intervention

The WellMind digital intervention was deployed on the BrainE platform, implemented in Unity Game Engine (Unity Technologies), and available on both iOS and Android phone devices [39]. This digital program is Health Insurance Portability and Accountability Act (HIPAA) compliant and secured by password protection, and each user interacts through an alphanumeric study ID that is not linked to any personal health information. Participants accessed the app in their own free time and engaged in breath-focused mindfulness training, with each session lasting 5-10 minutes for up to 60 sessions. The training was delivered in a game-like format and was performance adaptive. Specifically, individuals were requested to close their eyes, pay attention to their breathing, and tap the mobile screen after a specific number of breaths. The app monitored the consistency of tap responses. If the user was distracted based on the low consistency of breath monitoring taps, a gentle chime reminded the user to let go of the distraction and revert their attention back to mindful breathing. Initially, at level 1, participants tapped the screen after each breath. If they were able to do this consistently for 3 repeats of level 1 of 1 minute duration each, they graduated to level 2 and tracked 2 breaths at a time for 2 minutes, and so on. Thus, in the performance-adaptive task, the level reflected the number of minutes spent at that level and the number of breaths the participant was requested to repeatedly monitor. The maximum achievable level was level 10, that is, monitoring 10 breaths at a time for up to 10 minutes. When the user graduated to the max level, they stayed at this level until the end of all assigned sessions, that is, 60 sessions. Also within the game-like format, when the participant opened their eyes at the end of a level, a peaceful nature scene would slowly unfold as a form of training reward.

Overall, this digital meditative practice is considered closed loop because of its performance-adaptive feature [22,23]. Consistent attention to breathing is emphasized over other types of breathing techniques, such as deep breathing. The moment-to-moment performance tracking further allows quantification of the attentive focus during each session, which is not possible with traditional nondigital meditation.

Finally, the training also introduced standard compassion cultivation instructions as audio and text prompts before the start of each session's breath practice. Prompts were updated every 6 sessions, with a total of 10 prompts gradually increasing in complexity over 60 sessions. These prompts were designed per guidance from the Compassion Cultivation Training program [40] and included (1) settling the mind, (2) compassion for a loved one, (3) compassion for oneself, (4) loving kindness for oneself, (5) embracing common humanity, (6) embracing common humanity continued, (7) cultivating compassion for oneself and others, (8) cultivating compassion for others continued, (9) active compassion, and (10) integrated compassion cultivation practice. The WellMind training app and study design are summarized in Figure 1. Participants received in-app notifications once a day, reminding them to complete their training.

Figure 1. Brief digital mindfulness and compassion training. (A) The WellMind app delivered closed-loop, that is, performance-adaptive, attention to breath training. Common instruction across 10 levels of training is shown. At each level, the user tapped the mobile screen after (certain level) number of breaths while keeping their eyes closed. Feedback included auditory chimes to guide consistent performance and signal the end of training. A distinct, calming nature scene is unveiled at the end of each block in the session, along with focus feedback based on consistency of performance. Levels of breath monitoring were tied to 10 levels of cultivating compassion instructions. (B) The study design incorporated behavioral and neurophysiological assessments at pre- and posttraining and continued behavioral assessments at follow-up conducted in the WellMind training group relative to a no-contact control group.



Behavioral Assessments

At baseline (T1), postintervention completion (T2; or a 3-month no-contact period for the control group), and at follow-up (T3; 6 months following baseline), participants completed validated behavioral self-report scales of self-compassion: a 12-item self-compassion scale [41], and mindfulness: a 14-item Mindful Attention Awareness Scale [42]. These measures served as the primary outcomes. MBI measures were obtained as exploratory outcomes at T1 and T2. The Cronbach α measure of reliability was calculated for each of these behavioral measures at baseline.

Neurocognitive Assessments

In addition, participants completed an objective neurophysiological assessment of interoceptive attention to breathing at T1 and T2. For these assessments, all participants made individual study visits at the Neural Engineering and Translational Labs at the UCSD. Assessments were deployed on the BrainE platform with simultaneous EEG [43], delivered on a laptop (running on the Windows 10 operating system) at a comfortable viewing distance. The Lab Streaming Layer protocol was used to time stamp all user response events in this assessment [44].

In the interoceptive attention to breathing task, participants were instructed to close their eyes, breathe naturally, and respond every 2 breaths by tapping on the spacebar [45,46]. The computer screen appeared gray for the 5-minute duration of the task, implemented in two 2.5-minute blocks. A beep signaled the end of the task, at which time participants opened their eyes.

The median response time (RT) on the interoceptive task was monitored for all human participants so that we could identify and contrast neurophysiological activity on high consistency, that is, attentive breath monitoring trials (trials with RT≤1 median absolute deviation of median RT) versus low consistency, that is, distracted trials (trials with RT>1 median absolute deviation of median RT) in each human participant.

EEG data were collected using a 24-channel cap with saline soaked electrodes following the 10-20 system and a wireless SMARTING amplifier (mBrainTrain). The signals were digitized with a sampling rate of 500 Hz and 24-bit resolution and stored as .xdf files.

Behavioral Data Analyses

For behavioral subjective scales, scores on self-compassion and mindfulness were calculated at T1, T2, and T3 and for MBI at T1 and T2. T2 versus T1 and T3 versus T1 scores were compared within each group using 2-tailed paired *t* tests or its nonparametric equivalent Wilcoxon signed rank test depending on the distribution of the behavioral scores; the normality of distributions was checked using the Levene test. For mindfulness, we compared state mindfulness across the 3 time points, which is a component of the dispositional trait mindfulness scale [47], as we expected state mindfulness but not trait mindfulness to be malleable with training.

Cohen *d* effect sizes were calculated for both within and between-group differences. Repeated measures analyses of variance comparing between-group behavioral differences were not conducted in this first study given the large effect sizes (*d*>0.8) needed to observe significant group differences with adequate power.

To investigate the relationship between outcome gains and training engagement, behavioral changes in the WellMind training group were correlated with the number of training sessions completed by participants using Spearman correlations.

Neurocognitive Data Analyses

We applied a uniform processing pipeline to all EEG data published in several of our studies [43,45,46,48-53]. This included (1) EEG channel data processing and (2) cortical source localization of the EEG data to estimate source-level neural activity. Details of this analysis are provided in the Multimedia Appendix 1 [51,53-68].

Alpha band EEG data were trial averaged for high versus low consistency (ie, attended vs distracted) breath monitoring trials on the interoceptive attention task. These trials were compared for within-group pre- versus postintervention activity differences in the fronto-parietal network (FPN), cingulo-opercular network (CON), and DMN using paired *t* tests. Effect sizes were also calculated for neural data, reported as Cohen *d*: 0.2=small, 0.5=medium, and 0.8=large [69]. Given that we have observed large effect size neural outcomes (*d*>0.8) in our previous digital training studies [22,23,70], repeated measures ANOVA was conducted to analyze between-group post- versus preintervention network effects; the Greenhouse-Geisser significance correction was applied to adjust for lack of sphericity. Finally, Spearman correlations were used to analyze neurobehavioral associations.

Ethical Considerations

Each participant gave written informed consent in accordance with the Declaration of Helsinki before participating in the experiment. All the experimental procedures were approved by the institutional review board of the University of California San Diego (protocol #180140). All data is de-identified, and up to US $150 in compensation was provided as an e-Gift card to all participants for completing all aspects of the study, including assessments and intervention procedures.

Results

Baseline Group Comparisons

There were no demographic differences between the 2 groups for age, gender, and ethnicity (Table 1). Age comparisons were made using the Wilcoxon sum rank test, and gender and ethnicity comparisons were made using chi-square tests.

Table 1. Summary of participant demographics and baseline behaviors. Data were checked for normality, and WellMind and control groups were appropriately compared using the *t* test if normal, or else using the nonparametric Wilcoxon sum rank test. Gender and ethnicity variables were compared using chi-square tests.

Demographics and baseline behaviors	WellMind (n=22)	Control (n=21)	Group difference *P* value
Age (years), mean (SD)	27.91 (3.15)	29.67 (4.96)	.32
Gender n (%)			.55
Men	8 (36)	12 (57)	
Women	14 (64)	8 (38)	
Ethnicity n (%)			.19
Asian	8 (36)	8 (38)	
Black or African American	0 (0)	1 (5)	
American Indian or Alaska Native	0 (0)	0 (0)	
White	10 (46)	9 (43)	
More than 1 ethnicity	3 (14)	2 (10)	
Other	1 (5)	1 (5)	
Trait mindfulness, mean (SD)	3.18 (0.66)	3.14 (0.77)	.84
Self-compassion, mean (SD)	2.69 (0.57)	3.09 (0.73)	.05
MBI[a] emotional exhaustion, mean (SD)	16.27 (4.45)	24.05 (10.22)	.02[b]
MBI personal accomplishment, mean (SD)	21.41 (3.63)	32.86 (6.73)	<.001[c]
MBI depersonalization, mean (SD)	5.36 (3.33)	9.29 (7.93)	.23

[a]MBI: Maslach Burnout Inventory.

[b]*P*<.05.

[c]*P*<.001.

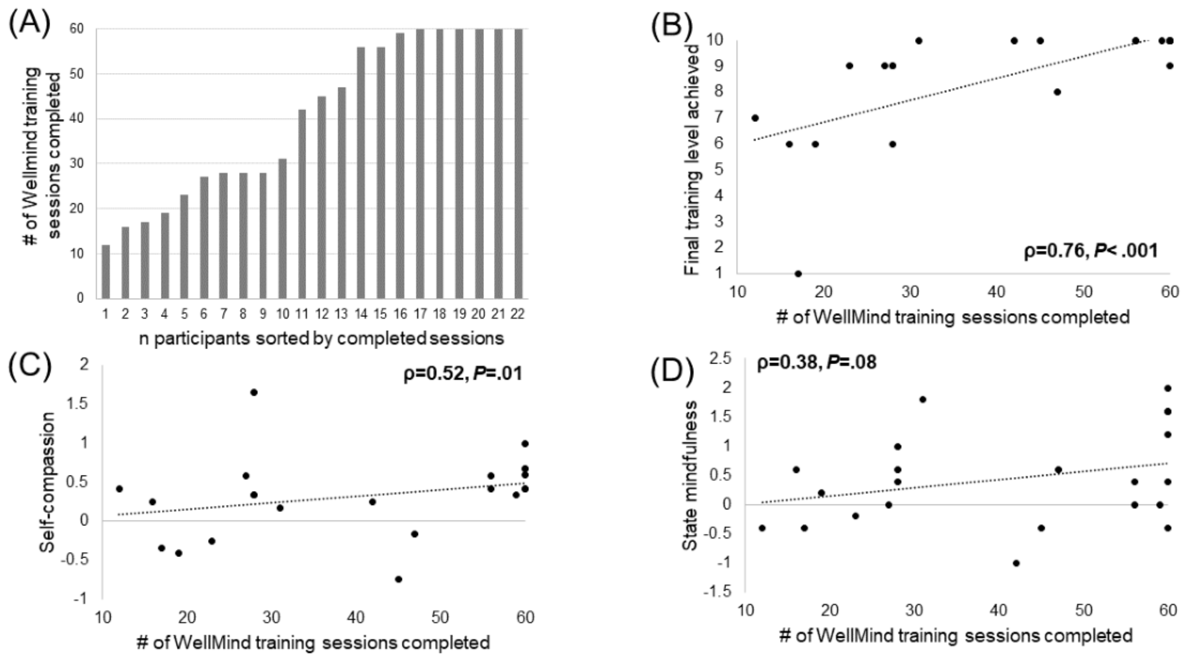
Our main outcome variables of mindfulness and self-compassion also did not show significant group differences at baseline. Burnout measures on the MBI showed significantly less emotional exhaustion and sense of personal accomplishment in the WellMind group relative to the control group at baseline, but no differences in MBI depersonalization. Overall, burnout levels in the WellMind group were low, that is, less than a scale midscore of 18 for emotional exhaustion, greater than a scale midscore of 16 for sense of personal accomplishment, and less than a scale midscore of 10 for depersonalization.

All behavioral measures had consistently high Cronbach α measured across all human participants at baseline (mindfulness: α=.81, self-compassion: α=.86, and MBI: α=.88).

Behavioral Results

Participants in the WellMind group completed 40.64 (SD 17.79) training sessions on average, or 68% (40.64/60) of the total 60 sessions; the number of training sessions completed by different participants is shown in Figure 2A. The number of training sessions completed directly related to the final level achieved in training across participants (ρ=0.76, *P*<.001; Figure 2B). Additionally, we found that the number of training sessions was also significantly related to post- versus preintervention increase in self-compassion (ρ=0.52, *P*=.01; Figure 2C) and trended toward significance in relation to post- versus preintervention increase in mindfulness (ρ=0.38, *P*=.08; Figure 2D), but this was not statistically significant.

Figure 2. WellMind training sessions completed and relationship with outcomes. (A) The number of training sessions completed by each WellMind participant. (B) The final level of breath training is significantly related to the number of completed training sessions. (C) Post- versus prechange in self-compassion is significantly related to number of completed training sessions. (D) Post- versus prechange in state mindfulness showed a trend in its relationship to the number of completed training sessions. The Spearman correlation results are shown.



The within-group changes in primary measures of self-compassion and state mindfulness and exploratory measures of burnout (MBI scores) between post- versus preintervention and follow-up versus preintervention sessions are shown in Tables 2 and 3 for the WellMind and control groups. A significant increase in self-compassion and mindfulness was observed at post- versus preintervention sessions only in the WellMind group, and the significant increase in self-compassion was sustained at follow-up (Figure 3). As the self-compassion scale has 2 components for compassionate self-responding (CSR) and uncompassionate self-responding (USR) [71], we also analyzed whether either CSR or USR undergo significant training-related change. CSR reflects self-kindness, common humanity, and mindfulness, while USR reflects the opposite constructs of self-judgment, isolation, and overidentification.

We found that the WellMind training-related improvement in self-compassion was exclusively driven by a reduction in USR (post vs preintervention change: –0.42, SD 0.55; *P*=.002; and follow-up vs preintervention change: –0.82, SD 1.27; *P*=.009), but there was no significant change in CSR at post (*P*=.17) or follow-up (*P*=.07). Baseline burnout in our sample was low, and there were no significant post- versus preintervention changes in MBI scores in both groups.

Cohen *d* effect sizes for within-group differences were calculated as the mean difference between post- versus preintervention (or follow-up vs preintervention) outcomes expressed in pooled SD units. Cohen *d* effect sizes were also calculated for between-group differences. Effect sizes for significant self-compassion and mindfulness outcomes were in the medium range (Tables 2 and 3).

Table 2. Summary of behavioral outcomes obtained at baseline (T1) and 3 months (postintervention; T2 after baseline. The primary outcome measures for the study, self-compassion and state mindfulness, showed significant changes in the WellMind group at postintervention time point and are noted. The control group did not show any significant changes. Data were checked for normality, and pre- versus postintervention within-group differences were appropriately compared using a paired *t* test if normal, or else using the nonparametric Wilcoxon sign rank test. Both within-group and between-group Cohen *d* effect sizes were calculated.

Pre- versus postintervention outcomes	WellMind				Control				Between-group effect size
	T1, mean (SD)	T2, mean (SD)	Effect size	Group difference *P* value	T1, mean (SD)	T2, mean (SD)	Effect size	Group difference *P* value	
Self-compassion	2.69 (0.57)	3.02 (0.58)	0.57	.007[a]	3.09 (0.73)	3.15 (0.67)	0.09	.61	0.49
State mindfulness	2.94 (0.79)	3.37 (0.85)	0.52	.02[b]	3.22 (0.94)	3.25 (0.82)	0.03	.86	0.53
MBI[c] emotional exhaustion	16.27 (4.45)	15 (5.86)	–0.24	.15	24.05 (10.22)	23.05 (12.45)	–0.09	.89	–0.04
MBI personal accomplishment	21.41 (3.63)	21.64 (4.41)	0.06	.61	32.86 (6.73)	34.67 (7.77)	0.25	.14	–0.35
MBI depersonalization	5.36 (3.33)	4.5 (3.71)	–0.24	.23	9.29 (7.93)	7.57 (7.37)	–0.22	.39	0.18

[a]*P*<.001.

[b]*P*<.05.

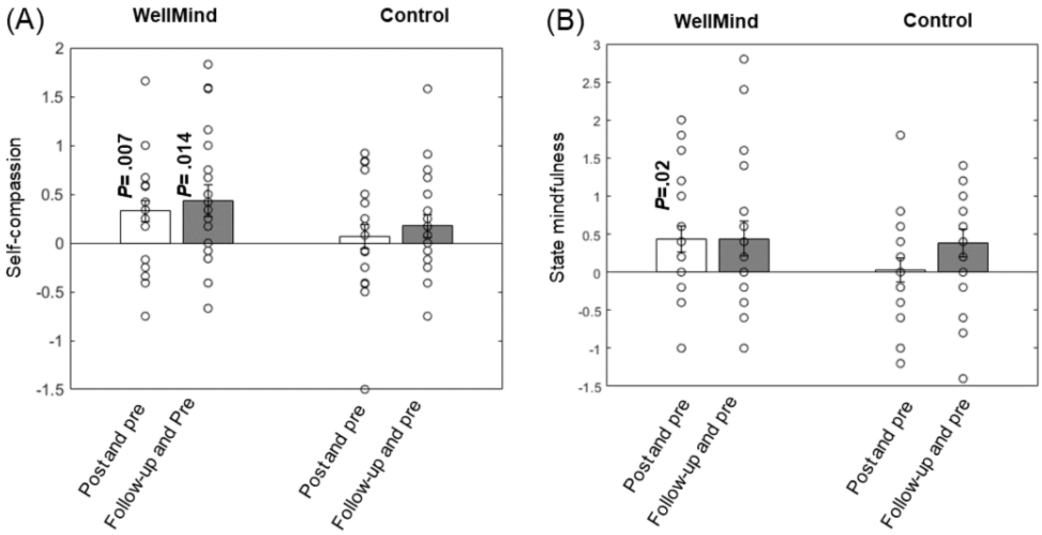
[c]MBI: Maslach Burnout Inventory.

Table 3. Summary of behavioral outcomes obtained at baseline (T1) and 6 months after baseline (follow-up; T3). Data were checked for normality, and pre- versus postintervention within-group differences were appropriately compared using a paired *t* test if normal, or else using the nonparametric Wilcoxon sign rank test. Both within-group and between-group Cohen *d* effect sizes were calculated.

Pre- versus postintervention outcomes	WellMind				Control				Between-group effect size
	T1, mean (SD)	T3, mean (SD)	Effect size	Group difference *P* value	T1, mean (SD)	T3, mean (SD)	Effect size	Group difference *P* value	
Self-compassion	2.69 (0.57)	3.14 (0.55)	0.80	.014[a]	3.09 (0.73)	3.29 (0.54)	0.31	.16	0.41
State mindfulness	2.94 (0.79)	3.41 (1.08)	0.50	.07	3.22 (0.94)	3.55 (0.78)	0.38	.05	0.06

[a]*P*<.001.

Figure 3. Post- versus preintervention (T2 vs T1) and follow-up versus preintervention (T3 vs T1) self-compassion and state mindfulness outcomes in the WellMind (n=22) and control (n=21) group; data at follow-up were missing for 2 participants in the WellMind group and for 1 participant in the control group. Both (A) self-compassion and (B) state mindfulness scores significantly improved in the WellMind group. Bar plots show the change in score mean and standard error about the mean for error bars, with the actual distribution of scores shown as scatter points. As self-compassion and state mindfulness measures had normal distributions, **P* value results are from paired t tests between pre- and postintervention or preintervention and follow-up assessments.



Neurocognitive Results

We analyzed post- versus preintervention modulation of source-localized neural activity on the interoceptive attention to breathing assessment in brain regions of interest collated within canonical cognitive control networks, specifically the FPN, CON, and DMN. There were no between-group differences in neural network activity at baseline (*P*>.05). The within-group changes in network activity between post- versus preintervention sessions are shown in Table 4 for the WellMind and control groups. Per our hypothesis, a significant decrease in activity in the mind-wandering and rumination-associated DMN was observed at post versus pre sessions only in the WellMind group (Figure 4). Besides this, the only other significant change observed was a decrease in post- versus preintervention FPN activity in the control group.

Cohen *d* effect sizes were calculated for both within-group and between-group differences (Table 4). Notably, the reduction in post- versus preintervention DMN activity in the WellMind group showed both large within-group and between-group effect sizes (DMN between group repeated measures ANOVA: session interaction: $F_{1,32}$=10.11; *P*=.003; η^2=0.24; main effects of group or session were not significant, *P*>.30; also, the other networks, FPN and CON, did not show any significant between-group effects, *P*>.10).

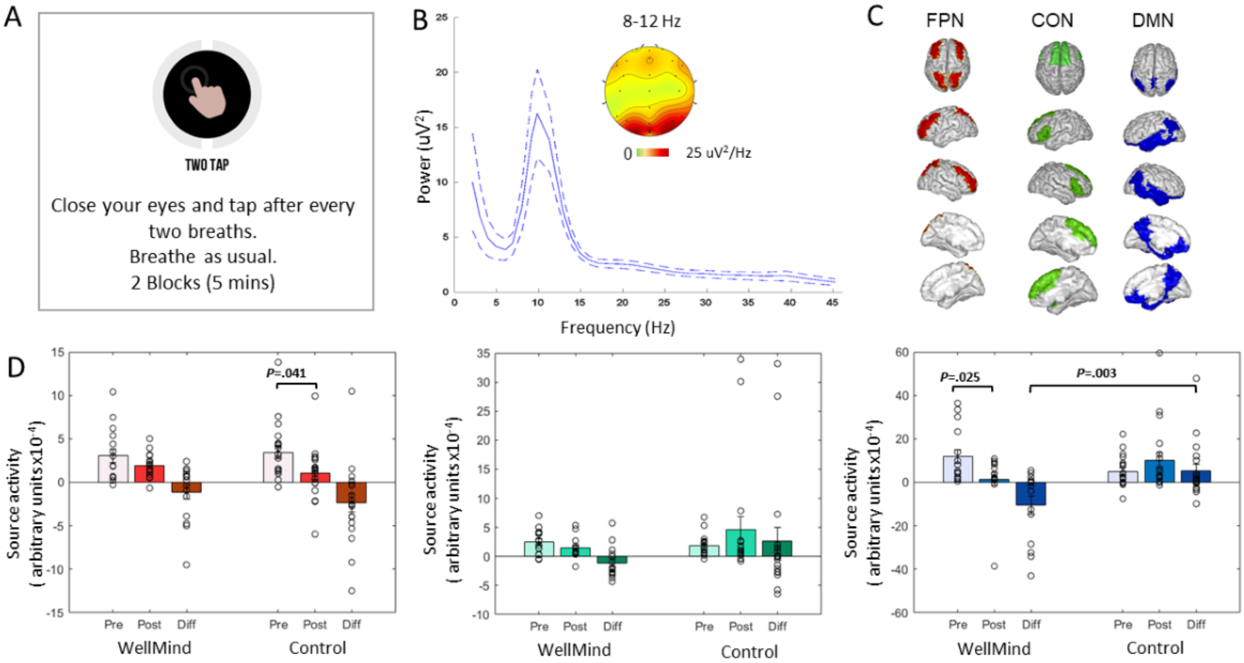
Table 4. Summary of neural activity outcomes obtained at baseline (T1) and at 3 months (postintervention; T2). The primary network of interest for the study, the default mode network (DMN), showed significant change in the WellMind group at post- versus preintervention and is noted. Besides, the control group showed a significant change in fronto-parietal network (FPN) activity. Mean (SD) data are shown in 10-4 cortical source arbitrary units for all variables. Post- versus preintervention within-group and between-group Cohen *d* effect sizes are calculated.

Pre- versus postintervention outcomes	WellMind				Control				Between-group effect size
	T1, mean (SD)	T2, mean (SD)	Effect size	Group difference *P* value	T1, mean (SD)	T2, mean (SD)	Effect size	Group difference *P* value	
FPN	3.08 (3.14)	1.96 (1.45)	–0.46	.21	3.45 (3.39)	1.08 (3.08)	–0.73	.04[a]	0.30
CON[b]	2.55 (2.18)	1.44 (1.76)	–0.56	.14	1.88 (1.77)	4.57 (9.86)	0.38	.27	–0.48
DMN	11.92 (12.55)	1.44 (11.66)	–0.87	.03[a]	4.94 (7.10)	10.26 (15.43)	0.44	.09	–1.10

[a]*P*<.05.

[b]CON: cingulo-opercular network.

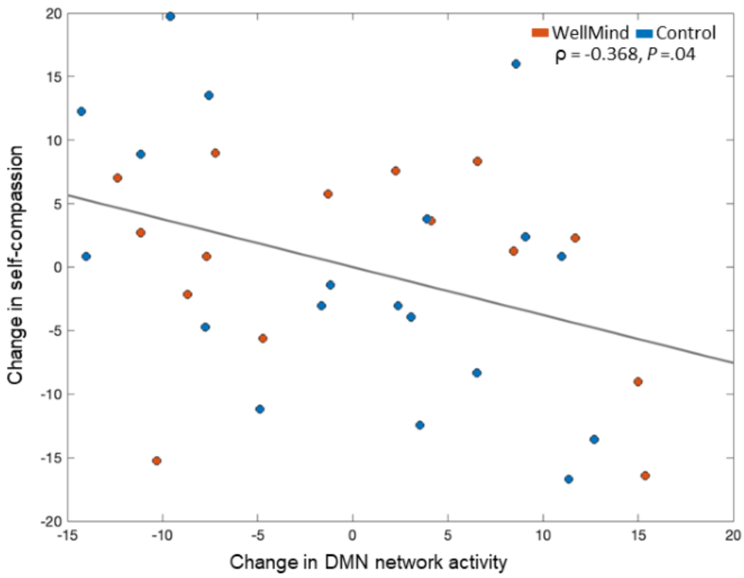
Figure 4. Training-related neurophysiological changes evaluated on the attention-to-breath monitoring assessment. (A) Schematic of task instructions. (B) A power frequency plot of scalp channel data across all participants and sessions showed peak processing in the α frequency band (8-12 Hz). (C) Source-reconstructed electroencephalography data were analyzed for 3 networks: frontoparietal network (FPN), the cingulo-opercular network (CON), and the default mode network (DMN); regions of interest averaged within each network are shown. (D) Comparisons of the WellMind versus control group network activity showed significant reduction in activity only for the WellMind group in the DMN; bar plots show mean and standard error about the mean α band activity (y-axes: 10–4 cortical source activity arbitrary units) within the 0- to 4-second epoch before breath responses, as well as the relative response on low versus high consistency (ie, distracted versus attended) trials at pre- and postintervention time points and the post-pre difference. Actual activity distributions are shown as scatter points.



Neurobehavioral Associations

We conducted neurobehavioral correlation analyses between neural measures of post- versus preintervention change in DMN activity and change in primary behavioral outcomes of self-compassion or state mindfulness. Spearman partial correlations were implemented that accounted for participant groups and their baseline burnout score differences (MBI; Table 1). We found a significant relationship between change in DMN network activity and change in self-compassion (ρ=–0.368; *P*=.04; Figure 5) but not state mindfulness (*P*>.50). Specifically, individuals who showed the largest improvement in self-compassion also showed the greatest DMN suppression. We also verified that these effects were driven by the relationship between DMN suppression and reduction in USR (ρ=0.37; *P*=.04), but there was no significant relationship with change in CSR (*P*=.35).

Figure 5. Relationship between post- versus preintervention change in self-compassion scores and change in default mode network (DMN) network activity. Residuals are plotted using Spearman partial correlation, taking into account group assignment to the WellMind or control group and baseline burnout score differences between groups.



We also found that the extent of training-related DMN suppression in the WellMind group was significantly related to their baseline DMN activity (ρ=–0.74; *P*=.002); this result did not change when controlling for the number of WellMind training sessions completed, and no such relationship was observed for the control group (*P*=.71). These results suggest that baseline DMN activity could be a predictive marker for the extent of training-related neural plasticity in this network, which in turn relates to improvement in self-compassion.

Discussion

In this study, we developed WellMind, a digital mindfulness and compassion training, and implemented it with health care professionals. Respecting their real-world time constraints, the training was brief (5-10 minutes per session) and available for up to 60 sessions. The novelty of the training lay in its closed-loop, that is, performance-adaptive mechanics and quantitative feedback design applied to attentive breathing, with levels of compassion cultivation instructions. To account for the practice effects of repeat assessments, we included a no-contact control group in the study design. We found that the WellMind intervention significantly improved the primary behavioral outcomes of self-compassion and mindfulness, with the improvements in self-compassion sustained at follow-up. No such behavioral effects were observed in the control group, and overall effect sizes were in the medium range. Concomitantly, we also found training-related neurophysiological suppression of the DMN, which is implicated in mind-wandering and rumination, and the extent of DMN suppression related to significant improvements in self-compassion.

A recent meta-analytic review of 27 studies [72] and randomized controlled trials found that smartphone apps can be used to enhance mindfulness and compassion skills, as well as reduce stress [73,74]. Additionally, a recent scoping review concluded that it is feasible to deliver compassionate care within digital health care, particularly telemedicine [75]. Yet, previous digital interventions in this field have reported small effect sizes [72]. In this study, we found medium effect size behavioral changes that further correlated with the extent of digital training completed across human participants, highlighting the advantages of this closed-loop digital training. It is also notable that the medium effect size improvements in self-compassion were sustained at follow-up; this finding is aligned with evidence from our past digital intervention studies showing robust long-term behavioral effects [22,76].

The core element of our digital WellMind training is the attention to breathing practice, which is also a foundational element of traditional meditation; indeed, meditation practice has been shown to enhance state-mindfulness [77,78]. Yet, traditional practice lacks real-time feedback and performance-based level progressions and can have variable outcomes depending on the teacher [79,80]. WellMind is a teacher-independent digital app within which user feedback is key to maintaining engagement over multiple sessions [24]. Notably, as users engaged with more sessions, we also observed them progress to higher levels of breath monitoring (maximum monitoring of 10 breaths at a time)—a significant correlation was found between breath monitoring level and training sessions completed. At higher training levels, progressively more sophisticated compassion instructions were also unveiled, which may be driving the correlation between the number of training sessions completed and improvements in self-compassion. Furthermore, when we investigated the positive (CSR) and negative (USR) subcomponents of the self-compassion scale [71], we found that reduction in USR (ie, reduction in self-judgment, isolation, and overidentification) exclusively underlies the training-related improvement in total self-compassion. To the best of our knowledge, previous research has not addressed the differential intervention-related plasticity of CSR versus USR components but has shown that these are distinct constructs that differentially relate to well-being and cognitive responses to daily life problems [81,82].

With regard to real-time feedback, WellMind integrates breath monitoring consistency-based auditory feedback as a gentle chime that signals the user to return to attentive breath monitoring when distracted. Indeed, auditory feedback during focused attention meditation has been shown to improve state mindfulness [83]. Notably, in large sample surveys of perceived barriers to meditation across hundreds of participants, a lack of individualized feedback and progress tracking have been prominently cited as important hindrances [84,85]. WellMind was designed to remove these barriers and, hence, motivate training. It is also distinct from previous closed-loop digital meditation approaches that prompt the user to retrospectively and subjectively report whether they were attentive or distracted during their meditation practice and did not integrate compassion cultivation [22,23].

In addition to behavioral outcomes, we also investigated EEG-based neurophysiological outcomes evaluated on an interoceptive breath monitoring task. The eyes-closed interoceptive task required monitoring of 2 breaths at a time and was devoid of feedback or performance-adaptive levels, and thus served as a pure assessment of breath-focused interoception that was targeted by the WellMind training. Consistent with our hypothesis based on previous evidence [28,29,31], we observed training-related suppression of cortical source-localized DMN activity, with no such changes found in the control group. The DMN outcome had a large between-group effect size. In contrast, the FPN and CON executive control networks did not show training-related changes; the FPN showed a significant within-group post versus pre activity reduction in the control group that could be related to attentional lapses or boredom during repeat assessments in this group [22,86,87]. Finally, neurobehavioral correlations showed that training-related DMN activity suppression was significantly related to improvement in self-compassion, specifically reduction in the USR component. Also, individuals with high DMN activity at baseline experienced greater DMN suppression with training, suggesting baseline DMN activity as an individual-specific neural marker to determine who may benefit most from such training. Overall, it is worth noting that we obtained these results using EEG, a scalable approach for measuring neural markers relative to other neuroimaging modalities such as functional magnetic resonance imaging. We also note that in our previous work, we have shown high test-retest reliability of such EEG data collected within the context of cognitive tasks [43].

As per the limitations of this study, the sample size and nonactive control group are the primary limitations, although it is worth noting that this is the first ever implementation of the digital WellMind intervention in health care professionals. We also acknowledge other limitations of our intervention design, such as that participants were not blind to the intervention condition and that WellMind participants received a greater frequency of contact from the experimenters during the intervention (email reminders to encourage adherence), while the control group had no such contact. These differences may lead to differences in expectations, resulting in positive outcomes observed in the WellMind group. Also, burnout in our sample was low, and we did not observe any post versus pre changes in burnout. Yet, previous work has suggested that organization-directed workplace interventions could be more effective at addressing burnout [25] and a recent study showed that physical exercise facilitated by mobile app technology also benefits burnout [88]. In this context, our research presents the possibility that individual interventions focused on self-compassion and mindfulness delivered before the development of burnout may be helpful in burnout prevention. Our data, showing long-term benefits in self-compassion with moderate to large effect sizes, suggests this approach may be useful in preventing burnout. Yet, this field has acknowledged that there are challenges to implementing behavioral interventions at the organizational level, especially in health care [8]. These challenges include a clear lack of time for physicians as well as concerns regarding confidentiality and discrimination within this workforce. In this context, WellMind offers brief training sessions (5-10 minutes in duration) that can be engaged flexibly at any time of day. Its closed-loop features promote adherence, although we acknowledge that not everyone is similarly motivated, as reflected by the variable number of sessions completed by the training participants. Security and confidentiality should be a paramount concern for any digital intervention [89,90]. Hence, WellMind access is enabled on a secure, password-protected HIPAA-compliant platform, with only deidentified data available for review and analyses. These features can facilitate future adoption. Overall, it is recommended that such interventions are embedded within the early stages of physician education as a foundational resource, when there may be more availability of time, to enhance receptivity, adherence, and build resilience to future burnout.

In conclusion, this study showcases an accessible and closed-loop approach to foster compassion and mindfulness among health care professionals. We found significant behavioral effects as well as robust brain plasticity related to these effects. These findings encourage future scale-up of the digital intervention to promote physician well-being and prevent burnout.

Acknowledgments

This work was supported by grants from the Sanford Institute for Empathy and Compassion at the University of California, San Diego (JM) and Sanford Institute for Empathy and Compassion Postdoctoral Fellowship Award (SJ). We thank Alankar Misra for the software development of the BrainE software and several University of California San Diego undergraduate students who assisted with data collection. We also thank Pragathi Balasubramani and Gillian Grennan for their assistance with the initial project set-up. The BrainE software is copyrighted for commercial use (Regents of the University of California Copyright #SD2018-816) and free for research and educational purposes.

Data Availability

The data set generated and analyzed in this study are available from the Data Dryad repository [91].

The WellMind digital intervention on the BrainE platform app is available for health care education and research on the website and is copyrighted for commercial use (Regents of the University of California Copyright #SD2024-132) [92].

Authors' Contributions

DR and JM contributed to the conception and design of the study. SJ conducted data management and data analytics and wrote the first draft of the manuscript. SRP and JKM collected and managed the data; JN and NA managed and analyzed the data. JM supervised all data collection, data management, analytics, and manuscript writing. All authors contributed to manuscript edits and revisions and approved the submitted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary methods.
[DOCX File , 22 KB - mental_v11i1e49467_app1.docx]

Multimedia Appendix 2
CONSORT-EHEALTH checklist (V 1.6.1).
[PDF File (Adobe PDF File), 1079 KB - mental_v11i1e49467_app2.pdf]

References

1. Soares JP, Lopes RH, de Souza Mendonça PB, Silva CRDV, Rodrigues CCFM, de Castro JL. Use of the Maslach Burnout Inventory among public health care professionals: scoping review. JMIR Ment Health 2023;10:e44195 [FREE Full text] [doi: 10.2196/44195] [Medline: 37477960]
2. Tajirian T, Jankowicz D, Lo B, Sequeira L, Strudwick G, Almilaji K, et al. Tackling the burden of electronic health record use among physicians in a mental health setting: physician engagement strategy. J Med Internet Res 2022;24(3):e32800 [FREE Full text] [doi: 10.2196/32800] [Medline: 35258473]
3. Shanafelt TD, West CP, Sinsky C, Trockel M, Tutty M, Satele DV, et al. Changes in burnout and satisfaction with work-life integration in physicians and the general US working population between 2011 and 2017. Mayo Clin Proc 2019;94(9):1681-1694 [FREE Full text] [doi: 10.1016/j.mayocp.2018.10.023] [Medline: 30803733]
4. Shanafelt TD, West CP, Sinsky C, Trockel M, Tutty M, Wang H, et al. Changes in burnout and satisfaction with work-life integration in physicians and the general US working population between 2011 and 2020. Mayo Clin Proc 2022;97(3):491-506 [FREE Full text] [doi: 10.1016/j.mayocp.2021.11.021] [Medline: 35246286]
5. West CP, Shanafelt TD. The influence of personal and environmental factors on professionalism in medical education. BMC Med Educ 2007;7(1):29 [FREE Full text] [doi: 10.1186/1472-6920-7-29] [Medline: 17760986]
6. Kim E, Mallett R, Hrabok M, Yang YA, Moreau C, Nwachukwu I, et al. Reducing burnout and promoting health and wellness among medical students, residents, and physicians in Alberta: protocol for a cross-sectional questionnaire study. JMIR Res Protoc 2020;9(4):e16285 [FREE Full text] [doi: 10.2196/16285] [Medline: 32301742]
7. Dyrbye LN, Massie FS, Eacker A, Harper W, Power D, Durning SJ, et al. Relationship between burnout and professional conduct and attitudes among US medical students. JAMA 2010;304(11):1173-1180 [FREE Full text] [doi: 10.1001/jama.2010.1318] [Medline: 20841530]
8. Harvey SB, Epstein RM, Glozier N, Petrie K, Strudwick J, Gayed A, et al. Mental illness and suicide among physicians. Lancet 2021;398(10303):920-930 [FREE Full text] [doi: 10.1016/S0140-6736(21)01596-8] [Medline: 34481571]
9. Lai J, Ma S, Wang Y, Cai Z, Hu J, Wei N, et al. Factors associated with mental health outcomes among health care workers exposed to coronavirus disease 2019. JAMA Netw Open 2020;3(3):e203976 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.3976] [Medline: 32202646]
10. Tement S, Ketiš ZK, Miroševič Š, Selič-Zupančič P. The impact of Psychological Interventions with Elements of Mindfulness (PIM) on empathy, well-being, and reduction of burnout in physicians: a systematic review. Int J Environ Res Public Health 2021;18(21):11181 [FREE Full text] [doi: 10.3390/ijerph182111181] [Medline: 34769700]
11. West CP, Dyrbye LN, Erwin PJ, Shanafelt TD. Interventions to prevent and reduce physician burnout: a systematic review and meta-analysis. Lancet 2016;388(10057):2272-2281. [doi: 10.1016/S0140-6736(16)31279-X] [Medline: 27692469]
12. Panagioti M, Panagopoulou E, Bower P, Lewith G, Kontopantelis E, Chew-Graham C, et al. Controlled interventions to reduce burnout in physicians: a systematic review and meta-analysis. JAMA Intern Med 2017;177(2):195-205 [FREE Full text] [doi: 10.1001/jamainternmed.2016.7674] [Medline: 27918798]
13. Hashem Z, Zeinoun P. Self-compassion explains less burnout among healthcare professionals. Mindfulness (N Y) 2020;11(11):2542-2551 [FREE Full text] [doi: 10.1007/s12671-020-01469-5] [Medline: 32929384]
14. Dev V, Fernando AT, Lim AG, Consedine NS. Does self-compassion mitigate the relationship between burnout and barriers to compassion? a cross-sectional quantitative study of 799 nurses. Int J Nurs Stud 2018;81:81-88. [doi: 10.1016/j.ijnurstu.2018.02.003] [Medline: 29518624]
15. Kemper KJ, McClafferty H, Wilson PM, Serwint JR, Batra M, Mahan JD, et al. Do mindfulness and self-compassion predict burnout in pediatric residents? Acad Med 2019;94(6):876-884 [FREE Full text] [doi: 10.1097/ACM.0000000000002546] [Medline: 30520809]
16. Krasner MS, Epstein RM, Beckman H, Suchman AL, Chapman B, Mooney CJ, et al. Association of an educational program in mindful communication with burnout, empathy, and attitudes among primary care physicians. JAMA 2009;302(12):1284-1293 [FREE Full text] [doi: 10.1001/jama.2009.1384] [Medline: 19773563]
17. Eckstein J, Rana ZH, Caravan S, Sharma R, Potters L, Parashar B. Burnout in radiation oncology physician workforce: the effect of mindfulness and fulfillment. Adv Radiat Oncol 2022;7(6):100971 [FREE Full text] [doi: 10.1016/j.adro.2022.100971] [Medline: 35662794]
18. Mascaro JS, Kelley S, Darcher A, Negi LT, Worthman C, Miller A, et al. Meditation buffers medical student compassion from the deleterious effects of depression. J Posit Psychol 2016;13(2):133-142. [doi: 10.1080/17439760.2016.1233348]
19. Epstein RM, Krasner MS. Physician resilience: what it means, why it matters, and how to promote it. Acad Med 2013;88(3):301-303 [FREE Full text] [doi: 10.1097/ACM.0b013e318280cff0] [Medline: 23442430]
20. Wen L, Sweeney TE, Welton L, Trockel M, Katznelson L. Encouraging mindfulness in medical house staff via smartphone app: a pilot study. Acad Psychiatry 2017;41(5):646-650. [doi: 10.1007/s40596-017-0768-3] [Medline: 28795335]

21. Smith JL, Allen JW, Haack CI, Wehrmeyer KL, Alden KG, Lund MB, et al. Impact of app-delivered mindfulness meditation on functional connectivity, mental health, and sleep disturbances among physician assistant students: randomized, wait-list controlled pilot study. JMIR Form Res 2021;5(10):e24208 [FREE Full text] [doi: 10.2196/24208] [Medline: 34665153]
22. Mishra J, Sagar R, Parveen S, Kumaran S, Modi K, Maric V, et al. Closed-loop digital meditation for neurocognitive and behavioral development in adolescents with childhood neglect. Transl Psychiatry 2020;10(1):153 [FREE Full text] [doi: 10.1038/s41398-020-0820-z] [Medline: 32424253]
23. Ziegler DA, Simon AJ, Gallen CL, Skinner S, Janowich JR, Volponi JJ, et al. Closed-loop digital meditation improves sustained attention in young adults. Nat Hum Behav 2019;3(7):746-757. [doi: 10.1038/s41562-019-0611-9] [Medline: 31160812]
24. Mishra J, Anguera JA, Gazzaley A. Video games for neuro-cognitive optimization. Neuron 2016;90(2):214-218 [FREE Full text] [doi: 10.1016/j.neuron.2016.04.010] [Medline: 27100194]
25. DeChant PF, Acs A, Rhee KB, Boulanger TS, Snowdon JL, Tutty MA, et al. Effect of organization-directed workplace interventions on physician burnout: a systematic review. Mayo Clin Proc Innov Qual Outcomes 2019;3(4):384-408 [FREE Full text] [doi: 10.1016/j.mayocpiqo.2019.07.006] [Medline: 31993558]
26. Anālayo B. Meditation on the breath: mindfulness and focused attention. Mindfulness 2019;10(8):1684-1691. [doi: 10.1007/s12671-019-01169-9]
27. Maric V, Ramanathan D, Mishra J. Respiratory regulation and interactions with neuro-cognitive circuitry. Neurosci Biobehav Rev 2020;112:95-106 [FREE Full text] [doi: 10.1016/j.neubiorev.2020.02.001] [Medline: 32027875]
28. Brewer JA, Worhunsky PD, Gray JR, Tang YY, Weber J, Kober H. Meditation experience is associated with differences in default mode network activity and connectivity. Proc Natl Acad Sci U S A 2011;108(50):20254-20259 [FREE Full text] [doi: 10.1073/pnas.1112029108] [Medline: 22114193]
29. Garrison KA, Zeffiro TA, Scheinost D, Constable RT, Brewer JA. Meditation leads to reduced default mode network activity beyond an active task. Cogn Affect Behav Neurosci 2015;15(3):712-720 [FREE Full text] [doi: 10.3758/s13415-015-0358-3] [Medline: 25904238]
30. Scheibner HJ, Bogler C, Gleich T, Haynes JD, Bermpohl F. Internal and external attention and the default mode network. Neuroimage 2017;148:381-389 [FREE Full text] [doi: 10.1016/j.neuroimage.2017.01.044] [Medline: 28110087]
31. Simon R, Engström M. The default mode network as a biomarker for monitoring the therapeutic effects of meditation. Front Psychol 2015;6:776 [FREE Full text] [doi: 10.3389/fpsyg.2015.00776] [Medline: 26106351]
32. Harrison BJ, Pujol J, López-Solà M, Hernández-Ribas R, Deus J, Ortiz H, et al. Consistency and functional specialization in the default mode brain network. Proc Natl Acad Sci U S A 2008;105(28):9781-9786 [FREE Full text] [doi: 10.1073/pnas.0711791105] [Medline: 18621692]
33. Koban L, Gianaros PJ, Kober H, Wager TD. The self in context: brain systems linking mental and physical health. Nat Rev Neurosci 2021;22(5):309-322 [FREE Full text] [doi: 10.1038/s41583-021-00446-8] [Medline: 33790441]
34. Kucyi A, Esterman M, Riley CS, Valera EM. Spontaneous default network activity reflects behavioral variability independent of mind-wandering. Proc Natl Acad Sci U S A 2016;113(48):13899-13904 [FREE Full text] [doi: 10.1073/pnas.1611743113] [Medline: 27856733]
35. Malinowski P. Neural mechanisms of attentional control in mindfulness meditation. Front Neurosci 2013;7:8 [FREE Full text] [doi: 10.3389/fnins.2013.00008] [Medline: 23382709]
36. Poerio GL, Sormaz M, Wang HT, Margulies D, Jefferies E, Smallwood J. The role of the default mode network in component processes underlying the wandering mind. Soc Cogn Affect Neurosci 2017;12(7):1047-1062 [FREE Full text] [doi: 10.1093/scan/nsx041] [Medline: 28402561]
37. Zhou HX, Chen X, Shen YQ, Li L, Chen NX, Zhu ZC, et al. Rumination and the default mode network: meta-analysis of brain imaging studies and implications for depression. Neuroimage 2020;206:116287 [FREE Full text] [doi: 10.1016/j.neuroimage.2019.116287] [Medline: 31655111]
38. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009;41(4):1149-1160 [FREE Full text] [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
39. Misra A, Ojeda A, Mishra J. BrainE: a digital platform for evaluating, engaging and enhancing brain function. Regents of the University of California Copyright SD2018-816. 2018. URL: https://escholarship.org/content/qt7pm2t2vp/qt7pm2t2vp_noSplash_d271bb684ce452e1d2ebc0429a1b7e1a.pdf [accessed 2023-12-20]
40. Jinpa T. Compassion Cultivation Training© (CCTTM). Compassion Inst. 2009. URL: https://www.compassioninstitute.com/cct/ [accessed 2023-04-20]
41. Raes F, Pommier E, Neff KD, Van Gucht D. Construction and factorial validation of a short form of the self-compassion scale. Clin Psychol Psychother 2011;18(3):250-255. [doi: 10.1002/cpp.702] [Medline: 21584907]
42. Brown KW, Ryan RM. The benefits of being present: mindfulness and its role in psychological well-being. J Pers Soc Psychol 2003;84(4):822-848. [doi: 10.1037/0022-3514.84.4.822] [Medline: 12703651]
43. Balasubramani PP, Ojeda A, Grennan G, Maric V, Le H, Alim F, et al. Mapping cognitive brain functions at scale. Neuroimage 2021;231:117641 [FREE Full text] [doi: 10.1016/j.neuroimage.2020.117641] [Medline: 33338609]
44. Kothe C, Medine D, Boulay C, Grivich M, Stenner T. Lab streaming layer. Copyright. 2019. URL: https://labstreaminglayer.org/#/ [accessed 2023-12-20]

45. Grennan G, Balasubramani P, Maric V, Ramanathan D, Mishra J. Default mode and frontal executive network interactions enable interoceptive attention and mindfulness. Research Square. Preprint posted online on October 4, 2021 [FREE Full text] [doi: 10.21203/rs.3.rs-951928/v1]
46. Shah RV, Grennan G, Zafar-Khan M, Alim F, Dey S, Ramanathan D, et al. Personalized machine learning of depressed mood using wearables. Transl Psychiatry 2021;11(1):338 [FREE Full text] [doi: 10.1038/s41398-021-01445-0] [Medline: 34103481]
47. Scientific scales and measures. Greater Good Science Center B. 2003. URL: https://ggsc.berkeley.edu/who_we_serve/researchers/scientific_scales_and_measures [accessed 2023-04-20]
48. Balasubramani PP, Diaz-Delgado J, Grennan G, Alim F, Zafar-Khan M, Maric V, et al. Distinct neural activations correlate with maximization of reward magnitude versus frequency. Cereb Cortex 2023;33(10):6038-6050. [doi: 10.1093/cercor/bhac482] [Medline: 36573422]
49. Grennan G, Balasubramani PP, Vahidi N, Ramanathan D, Jeste DV, Mishra J. Dissociable neural mechanisms of cognition and well-being in youth versus healthy aging. Psychol Aging 2022;37(7):827-842 [FREE Full text] [doi: 10.1037/pag0000710] [Medline: 36107693]
50. Kato R, Balasubramani PP, Ramanathan D, Mishra J. Utility of cognitive neural features for predicting mental health behaviors. Sensors (Basel) 2022;22(9):3116 [FREE Full text] [doi: 10.3390/s22093116] [Medline: 35590804]
51. Nan J, Balasubramani PP, Ramanathan D, Mishra J. Neural dynamics during emotional video engagement relate to anxiety. Front Hum Neurosci 2022;16:993606 [FREE Full text] [doi: 10.3389/fnhum.2022.993606] [Medline: 36438632]
52. Ojeda A, Wagner M, Maric V, Ramanathan D, Mishra J. EEG source derived salience network coupling supports real-world attention switching. Neuropsychologia 2023;178:108445 [FREE Full text] [doi: 10.1016/j.neuropsychologia.2022.108445] [Medline: 36502931]
53. Ojeda A, Kreutz-Delgado K, Mishra J. Bridging M/EEG source imaging and independent component analysis frameworks using biologically inspired sparsity priors. Neural Comput 2021;33(9):2408-2438 [FREE Full text] [doi: 10.1162/neco_a_01415] [Medline: 34412115]
54. Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods 2004 Mar 15;134(1):9-21. [doi: 10.1016/j.jneumeth.2003.10.009] [Medline: 15102499]
55. Ojeda A, Kreutz-Delgado K, Mullen T. Fast and robust block-sparse Bayesian learning for EEG source imaging. Neuroimage 2018 Jul 01;174:449-462. [doi: 10.1016/j.neuroimage.2018.03.048] [Medline: 29596978]
56. Barry RJ, Clarke AR, Johnstone SJ, Magee CA, Rushby JA. EEG differences between eyes-closed and eyes-open resting conditions. Clin Neurophysiol 2007 Dec;118(12):2765-2773. [doi: 10.1016/j.clinph.2007.07.028] [Medline: 17911042]
57. Kan DPX, Croarkin PE, Phang CK, Lee PF. EEG differences between eyes-closed and eyes-open conditions at the resting stage for euthymic participants. Neurophysiology 2018 Mar 26;49(6):432-440. [doi: 10.1007/s11062-018-9706-6]
58. Li L. The differences among eyes-closed, eyes-open and attention states: an EEG study. 2010 Presented at: 2010 6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM); September 23-25, 2010; Chengdu, China p. 1-4. [doi: 10.1109/WICOM.2010.5600726]
59. Valipour S, Shaligram AD, Kulkarni GR. Detection of an alpha rhythm of EEG signal based on EEGLAB. International Journal of Engineering Research and Applications 2014 Jan;4(1):154-159 [FREE Full text]
60. Pascual-Marqui R, Michel C, Lehmann D. Low resolution electromagnetic tomography: a new method for localizing electrical activity in the brain. Int J Psychophysiol 1994 Oct;18(1):49-65. [doi: 10.1016/0167-8760(84)90014-x] [Medline: 7876038]
61. Ding L, He B. Sparse source imaging in electroencephalography with accurate field modeling. Hum Brain Mapp 2008 Sep;29(9):1053-1067 [FREE Full text] [doi: 10.1002/hbm.20448] [Medline: 17894400]
62. Stopczynski A, Stahlhut C, Larsen JE, Petersen MK, Hansen LK. The smartphone brain scanner: a portable real-time neuroimaging system. PLoS One 2014;9(2):e86733 [FREE Full text] [doi: 10.1371/journal.pone.0086733] [Medline: 24505263]
63. Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 2006 Jul 01;31(3):968-980. [doi: 10.1016/j.neuroimage.2006.01.021] [Medline: 16530430]
64. Holmes CJ, Hoge R, Collins L, Woods R, Toga AW, Evans AC. Enhancement of MR images using registration for signal averaging. J Comput Assist Tomogr 1998;22(2):324-333. [doi: 10.1097/00004728-199803000-00032] [Medline: 9530404]
65. Dosenbach NUF, Visscher KM, Palmer ED, Miezin FM, Wenger KK, Kang HC, et al. A core system for the implementation of task sets. Neuron 2006 Jun 01;50(5):799-812 [FREE Full text] [doi: 10.1016/j.neuron.2006.04.031] [Medline: 16731517]
66. Dosenbach NUF, Fair DA, Miezin FM, Cohen AL, Wenger KK, Dosenbach RAT, et al. Distinct brain networks for adaptive and stable task control in humans. Proc Natl Acad Sci U S A 2007 Jun 26;104(26):11073-11078 [FREE Full text] [doi: 10.1073/pnas.0704320104] [Medline: 17576922]
67. Menon V, Uddin LQ. Saliency, switching, attention and control: a network model of insula function. Brain Struct Funct 2010 Jun;214(5-6):655-667 [FREE Full text] [doi: 10.1007/s00429-010-0262-0] [Medline: 20512370]

68. Uddin LQ, Yeo BTT, Spreng RN. Towards a universal taxonomy of macro-scale functional human brain networks. Brain Topogr 2019 Nov;32(6):926-942 [FREE Full text] [doi: 10.1007/s10548-019-00744-6] [Medline: 31707621]
69. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd Edition. Milton Park, Abingdon-on-Thames, Oxfordshire, England, UK: Routledge; 1988.
70. Mishra J, de Villers-Sidani E, Merzenich M, Gazzaley A. Adaptive training diminishes distractibility in aging across species. Neuron 2014;84(5):1091-1103 [FREE Full text] [doi: 10.1016/j.neuron.2014.10.034] [Medline: 25467987]
71. Neff KD. The self-compassion scale is a valid and theoretically coherent measure of self-compassion. Mindfulness 2015;7(1):264-274. [doi: 10.1007/s12671-015-0479-3]
72. Linardon J. Can acceptance, mindfulness, and self-compassion be learned by smartphone apps? a systematic and meta-analytic review of randomized controlled trials. Behav Ther 2020;51(4):646-658. [doi: 10.1016/j.beth.2019.10.002] [Medline: 32586436]
73. Min B, Park H, Kim JI, Lee S, Back S, Lee E, et al. The effectiveness of a neurofeedback-assisted mindfulness training program using a mobile app on stress reduction in employees: randomized controlled trial. JMIR Mhealth Uhealth 2023;11:e42851 [FREE Full text] [doi: 10.2196/42851] [Medline: 37788060]
74. Kirk U, Staiano W, Hu E, Ngnoumen C, Kunkle S, Shih E, et al. App-based mindfulness for attenuation of subjective and physiological stress reactivity in a population with elevated stress: randomized controlled trial. JMIR Mhealth Uhealth 2023;11:e47371 [FREE Full text] [doi: 10.2196/47371] [Medline: 37831493]
75. Kemp J, Zhang T, Inglis F, Wiljer D, Sockalingam S, Crawford A, et al. Delivery of compassionate mental health care in a digital technology-driven age: scoping review. J Med Internet Res 2020;22(3):e16263 [FREE Full text] [doi: 10.2196/16263] [Medline: 32141833]
76. Mishra J, Sagar R, Joseph AA, Gazzaley A, Merzenich MM. Training sensory signal-to-noise resolution in children with ADHD in a global mental health setting. Transl Psychiatry 2016;6(4):e781 [FREE Full text] [doi: 10.1038/tp.2016.45] [Medline: 27070409]
77. Kiken LG, Garland EL, Bluth K, Palsson OS, Gaylord SA. From a state to a trait: trajectories of state mindfulness in meditation during intervention predict changes in trait mindfulness. Pers Individ Dif 2015;81:41-46 [FREE Full text] [doi: 10.1016/j.paid.2014.12.044] [Medline: 25914434]
78. Tang YY, Hölzel BK, Posner MI. The neuroscience of mindfulness meditation. Nat Rev Neurosci 2015;16(4):213-225. [doi: 10.1038/nrn3916] [Medline: 25783612]
79. Lindahl JR, Fisher NE, Cooper DJ, Rosen RK, Britton WB. The varieties of contemplative experience: a mixed-methods study of meditation-related challenges in Western Buddhists. PLoS One 2017;12(5):e0176239 [FREE Full text] [doi: 10.1371/journal.pone.0176239] [Medline: 28542181]
80. Ruijgrok-Lupton PE, Crane RS, Dorjee D. Impact of mindfulness-based teacher training on MBSR participant well-being outcomes and course satisfaction. Mindfulness (N Y) 2018;9(1):117-128 [FREE Full text] [doi: 10.1007/s12671-017-0750-x] [Medline: 29387265]
81. Muris P, Otgaar H, Meesters C, Heutz A, van den Hombergh M. Self-compassion and adolescents' positive and negative cognitive reactions to daily life problems. J Child Fam Stud 2019;28(5):1433-1444 [FREE Full text] [doi: 10.1007/s10826-019-01353-4]
82. Pandey R, Tiwari GK, Parihar P, Rai PK. Positive, not negative, self-compassion mediates the relationship between self-esteem and well-being. Psychol Psychother 2021;94(1):1-15. [doi: 10.1111/papt.12259] [Medline: 31750614]
83. Hunkin H, King DL, Zajac IT. EEG neurofeedback during focused attention meditation: effects on state mindfulness and meditation experiences. Mindfulness 2020;12(4):841-851. [doi: 10.1007/s12671-020-01541-0]
84. Hunt CA, Hoffman MA, Mohr JJ, Williams AL. Assessing perceived barriers to meditation: the Determinants of Meditation Practice Inventory-Revised (DMPI-R). Mindfulness (N Y) 2020;11(5):1139-1149 [FREE Full text] [doi: 10.1007/s12671-020-01308-7] [Medline: 33664878]
85. Russ SL, Maruyama G, Sease TB, Jellema S. Do early experiences matter? development of an early meditation hindrances scale linked to novice meditators; intention to persist. Psychol Conscious Theory Res Pract US 2017;4(3):274-287. [doi: 10.1037/cns0000129]
86. Mahmoud MHZA. Neuroimaging of Endogenous Lapses of Responsiveness. 2022. URL: https://ir.canterbury.ac.nz/items/b74a7b1e-7125-4a5b-a550-16869ec5f8d7 [accessed 2023-12-20]
87. Unsworth N, Robison MK. A locus coeruleus-norepinephrine account of individual differences in working memory capacity and attention control. Psychon Bull Rev 2017;24(4):1282-1311 [FREE Full text] [doi: 10.3758/s13423-016-1220-5] [Medline: 28108977]
88. Boucher VG, Haight BL, Hives BA, Zumbo BD, Merali-Dewji A, Hutton S, et al. Effects of 12 weeks of at-home, application-based exercise on health care workers' depressive symptoms, burnout, and absenteeism: a randomized clinical trial. JAMA Psychiatry 2023;80(11):1101-1109. [doi: 10.1001/jamapsychiatry.2023.2706] [Medline: 37556150]
89. Venegas MD, Brooks JM, Myers AL, Storm M, Fortuna KL. Peer support specialists and service users' perspectives on privacy, confidentiality, and security of digital mental health. IEEE Pervasive Comput 2022;21(2):41-50 [FREE Full text] [doi: 10.1109/MPRV.2022.3141986] [Medline: 35814864]

90. World Health Organization. Recommendations on digital interventions for health system. WHO. 2019. URL: https://apps.who.int/iris/bitstream/handle/10665/324998/WHO-RHR-19.7-eng.pdf [accessed 2023-12-20]
91. Dryad. URL: https://datadryad.org/stash/share/ofaFuG5MA_9dNn6xeV4yX3OMoTJhXZp1KuHPj2ZS_o0 [accessed 2024-01-11]
92. WellMind for the Healthcare Community. Powered by REDCap. URL: https://tinyurl.com/wellmind-ucsdhealth [accessed 2023-12-30]

Abbreviations

CON: cingulo-opercular network
CSR: compassionate self-responding
DMN: default mode network
EEG: electroencephalography
FPN: fronto-parietal network
HIPAA: Health Insurance Portability and Accountability Act
MBI: Maslach Burnout Inventory
RT: response time
UCSD: University of California San Diego
USR: uncompassionate self-responding

Edited by J Torous; submitted 30.05.23; peer-reviewed by N Gupta, D Gratzer; comments to author 21.10.23; revised version received 09.11.23; accepted 01.12.23; published 22.01.24.

Please cite as:
Jaiswal S, Purpura SR, Manchanda JK, Nan J, Azeez N, Ramanathan D, Mishra J
Design and Implementation of a Brief Digital Mindfulness and Compassion Training App for Health Care Professionals: Cluster Randomized Controlled Trial
JMIR Ment Health 2024;11:e49467
URL: https://mental.jmir.org/2024/1/e49467
doi:10.2196/49467
PMID:38252479

Corrigenda and Addenda

Correction: Digital Mental Health Interventions for Alleviating Depression and Anxiety During Psychotherapy Waiting Lists: Systematic Review

Sijia Huang[1], MA; Yiyue Wang[1], BS; Gen Li[2], PhD; Brian J Hall[2,3,4], PhD; Thomas J Nyman[1,2,5], PhD

[1]Faculty of Arts and Sciences, New York University Shanghai, Shanghai, China
[2]Center for Global Health Equity, New York University Shanghai, Shanghai, China
[3]School of Global Public Health, New York University, New York, NY, United States
[4]Health, Behavior and Society, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, United States
[5]School of Psychology and Clinical Language Sciences, University of Reading, Reading, United Kingdom

Corresponding Author:
Thomas J Nyman, PhD
School of Psychology and Clinical Language Sciences
University of Reading
Earley
Reading, RG6 6ET
United Kingdom
Phone: 44 1189875123
Email: t.nyman@reading.ac.uk

Related Article:

Correction of: https://mental.jmir.org/2024/1/e56650

(JMIR Ment Health 2024;11:e67281) doi:10.2196/67281

In "Digital Mental Health Interventions for Alleviating Depression and Anxiety During Psychotherapy Waiting Lists: Systematic Review" (JMIR Ment Health 2024;11:e56650) 2 errors have been noted:

In Table 3, row "Twomey et al (2014)", column 4:

> *A brief introductory session and five 2- to 40-minute sessions*

has been revised to:

> *A brief introductory session and five 20- to 40-minute sessions*

In Table 6, row "Whitfield et al (2006)", column 4:

> *Six sessions of 4-60 minutes*

has been revised to:

> *Six sessions of 45-60 minutes*

The correction will appear in the online version of the paper on the JMIR Publications website on October 16, 2024 together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 07.10.24; this is a non–peer-reviewed article;accepted 07.10.24; published 16.10.24.

Please cite as:
Huang S, Wang Y, Li G, Hall BJ, Nyman TJ
Correction: Digital Mental Health Interventions for Alleviating Depression and Anxiety During Psychotherapy Waiting Lists: Systematic Review
JMIR Ment Health 2024;11:e67281
URL: https://mental.jmir.org/2024/1/e67281
doi:10.2196/67281
PMID:

Corrigenda and Addenda

Correction: Data Integrity Issues With Web-Based Studies: An Institutional Example of a Widespread Challenge

Blandine French[1], PhD; Camilla Babbage[1], PhD; Katherine Bird[1], MA; Lauren Marsh[1], PhD; Mirabel Pelton[1], PhD; Shireen Patel[1], MA; Sarah Cassidy[1], PhD; Stefan Rennick-Egglestone[1], PhD

School of Psychology, University of Nottingham, Nottingham, United Kingdom

Corresponding Author:
Blandine French, PhD
School of Psychology
University of Nottingham
University Park Campus
Psychology, Pharmacy, Life Sciences, East Dr
Nottingham, NG7 2RD
United Kingdom
Phone: 44 0115 748 6970
Email: blandine.french@nottingham.ac.uk

Related Article:

Correction of: https://mental.jmir.org/2024/1/e58432/

(JMIR Ment Health 2024;11:e67286) doi:10.2196/67286

In "Data Integrity Issues With Web-Based Studies: An Institutional Example of a Widespread Challenge" (JMIR Ment Health 2024;11:e58432) 4 errors have been noted:

1. In *Introduction*, the reference range in the sentence "These issues have also been reported in qualitative web-based interviews [2-26]..." has been revised to:

> *These issues have also been reported in qualitative web-based interviews [24-26]...*

2. In the subsection *Case Study 4: Web-Based Survey to Assess the Acceptability of SPARX, What Happened?*, the age range in the sentence "Our target was to reach 100-200, 1- to 19-year-olds" has been revised to:

> *Our target was to reach 100-200, 11- to 19-year-olds.*

3. In the subsection *Methodological or Research Challenges and Strategies, Overview*, the reference range in the sentence "This includes many strategies that have been summarized in publications [3-38]..." has been revised to:

> *This includes many strategies that have been summarized in publications [36-38]...*

4. Also in the subsection *Methodological or Research Challenges and Strategies, Overview*, the reference range in the sentence "There are many noteworthy strategies put forward in this literature [3-38]…" has been revised to:

> *There are many noteworthy strategies put forward in this literature [36-38]...*

The correction will appear in the online version of the paper on the JMIR Publications website on October 17, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 07.10.24; this is a non–peer-reviewed article;accepted 07.10.24; published 17.10.24.

Please cite as:
French B, Babbage C, Bird K, Marsh L, Pelton M, Patel S, Cassidy S, Rennick-Egglestone S
Correction: Data Integrity Issues With Web-Based Studies: An Institutional Example of a Widespread Challenge
JMIR Ment Health 2024;11:e67286
URL: https://mental.jmir.org/2024/1/e67286
doi:10.2196/67286
PMID:

Use of AI in Mental Health Care: Community and Mental Health Professionals Survey

Shane Cross[1,2], MPsych, PhD; Imogen Bell[1,2], BSc, PhD; Jennifer Nicholas[1,2], PhD; Lee Valentine[1,2], PhD; Shaminka Mangelsdorf[1,2], DPsych; Simon Baker[1], PhD; Nick Titov[3,4], PhD; Mario Alvarez-Jimenez[1,2], PhD

1
2
3
4

Corresponding Author:
Shane Cross, MPsych, PhD

Abstract

Background: Artificial intelligence (AI) has been increasingly recognized as a potential solution to address mental health service challenges by automating tasks and providing new forms of support.

Objective: This study is the first in a series which aims to estimate the current rates of AI technology use as well as perceived benefits, harms, and risks experienced by community members (CMs) and mental health professionals (MHPs).

Methods: This study involved 2 web-based surveys conducted in Australia. The surveys collected data on demographics, technology comfort, attitudes toward AI, specific AI use cases, and experiences of benefits and harms from AI use. Descriptive statistics were calculated, and thematic analysis of open-ended responses were conducted.

Results: The final sample consisted of 107 CMs and 86 MHPs. General attitudes toward AI varied, with CMs reporting neutral and MHPs reporting more positive attitudes. Regarding AI usage, 28% (30/108) of CMs used AI, primarily for quick support (18/30, 60%) and as a personal therapist (14/30, 47%). Among MHPs, 43% (37/86) used AI; mostly for research (24/37, 65%) and report writing (20/37, 54%). While the majority found AI to be generally beneficial (23/30, 77% of CMs and 34/37, 92% of MHPs), specific harms and concerns were experienced by 47% (14/30) of CMs and 51% (19/37) of MHPs. There was an equal mix of positive and negative sentiment toward the future of AI in mental health care in open feedback.

Conclusions: Commercial AI tools are increasingly being used by CMs and MHPs. Respondents believe AI will offer future advantages for mental health care in terms of accessibility, cost reduction, personalization, and work efficiency. However, they were equally concerned about reducing human connection, ethics, privacy and regulation, medical errors, potential for misuse, and data security. Despite the immense potential, integration into mental health systems must be approached with caution, addressing legal and ethical concerns while developing safeguards to mitigate potential harms. Future surveys are planned to track use and acceptability of AI and associated issues over time.

(JMIR Ment Health 2024;11:e60589) doi:10.2196/60589

KEYWORDS

mental health; health care; AI; community members; mental health professional; web-based survey; Australia; descriptive statistic; thematic analysis; cost reduction; data security; digital health; digital intervention; artificial intelligence

Introduction

Mental ill health is the leading cause of disability worldwide [1], yet fewer than half of all people with a mental health condition seek or receive evidence-based treatment [2-4]. Among the key structural barriers to effective care is that the demand outstrips the supply of qualified mental health professionals (MHPs), resulting in severely limited access and excessive wait times [5]. Moreover, MHPs are frequently burdened by substantial time-intensive administrative responsibilities and tasks, such as note-taking, detailed report writing, and planning for therapeutic sessions, limiting their availability to provide clinical care [6].

As digital technology becomes commonplace in society, tasks and services that were once performed manually, and often slowly, are now accomplished more efficiently via automated technology systems. However, despite many industries embracing new technologies for enhanced efficiency and responsiveness, the same progress has not been made in mental health care. Mental health care remains inaccessible, cumbersome to navigate, reactive, and slow to deliver, leaving mental health consumers frustrated and care providers burnt out [7].

People now have much greater access to information, including medical information and their own health data, than ever before [8]. In a contemporary landscape, where the prevalence of "on-demand" services is increasing, both mental health consumers and MHPs may expect comparable responsiveness. As a result, many are turning to digital products and services that aim to immediately address their needs. Young people, for example, are open and interested in using a range of digital technologies for mental health support, and many clinicians are already using these tools as part of routine care [9]. Wide-scale adoption of telehealth through the COVID-19 pandemic demonstrated the capacity for services to shift in response to changing demands, resulting in MHPs strongly endorsing the ongoing provision of technology-enhanced services [10]. This shift in attitude toward digital technology reflects an acknowledgment of the potential it holds for addressing barriers to providing effective and accessible care.

Recent advances in artificial intelligence (AI) have raised both excitement and debate concerning opportunities to harness this technology for mental health care [11,12]. AI encompasses a range of computer-based digital techniques and methodologies that perform cognitive processes characteristic of humans; such as learning, problem solving, pattern recognition, generalization, and predictive inference [13,14]. The recent advancement in natural language processing (NLP), a specialized branch of AI, has enabled chatbots and other language-driven systems to address requests, respond to queries, and provide advice autonomously, without human intervention [15]. Commercial tools such as ChatGPT enable users to enter any kind of query and obtain real-time responses significantly faster than traditional methods, such as internet searches.

A wide range of AI enabled products and services have been trialled in mental health care and health care more broadly [14]. AI has been used by health professionals to help solve complex problems such as identifying and diagnosing anomalies in medical images and genetic testing, predicting medical risk and disease prognosis, facilitating diagnostic and treatment decisions and recording and classifying clinical progress notes to name but a few [14]. For people with mental health difficulties, platforms such as Woebot [16] have been developed that use chatbots to deliver cognitive-behavioural therapy. More recently, tools such as ChatGPT have become freely available to the public, and with over 100 million users in its first few months, it was the fastest growing commercial application in history [17]. According to one community survey of over 1000 people in Australia, just under a half (48%) of Australians had heard of ChatGPT and almost a quarter (23%) had used it, with millennials (born between 1981 and 1996) and those with bachelor degrees and higher making up the majority of users [18]. Another youth survey found that 70% of young people 14 - 17 have used ChatGPT, with 59% using it for study and 42% for completing school assignments [19].

Large language model technologies like ChatGPT, are increasingly used by certain groups of consumers as an alternative to seeing a qualified MHP, and by some MHPs to assist with burdensome administrative tasks [20]. A recent global survey of approximately 800 psychiatrists [21] found that 75% thought it likely that AI would provide medical documentation, 54% to synthesize patient information to reach a diagnosis, 51% to analyze patient information to establish prognosis, and 47% to formulate personalized medication or therapy treatment plans for patients. In total, 36% felt that the benefits of AI would outweigh the risks, 25% felt that the risks would outweigh the benefits, and the rest were uncertain. These findings indicate that the segments of the mental health workforce anticipate that AI will be involved in care provision in some way, and that there are clearly risks and benefits which must be better understood.

The use of AI to support mental health care does come with potential harms [12,20]. For people using AI for mental health support there is the risk of misdiagnosis or misinformation, stemming from AI's potential for error. There are also questions about the role of empathy in AI systems, although a recent study highlighted that AI can outperform physicians in empathy measures [22]. Data privacy emerges as another salient issue, given the sensitive nature of mental health information and the potential for data breaches or misuse [23]. Biases inherent in nonrepresentative training data can lead to issues of inequity in diagnoses or treatments while the often-opaque decision-making processes of AI systems raise concerns about how complex decisions were made. These potential risks can lead to adverse consequences, and currently, there is limited information regarding the potentially harmful effects of these systems. Consequently, there is a dearth of legislation to safeguard users against such detrimental outcomes [23].

With the widescale popularity of AI technologies such as ChatGPT, it is highly likely that many of these applications are already being used in various ways for mental health care. As society continues to debate the various benefits and risks that this brings, it is critical to understand how and why these technologies are currently being used in the context of mental health care. This study is the first in a series of planned surveys which aims to estimate the current rates of use of AI technology by CMs for mental health and well-being purposes, as well as MHPs for professional purposes, to better understand the scale of use as well as the experienced benefits, harms, and risks associated with its use.

Methods

Study Design and Setting

Community members (CMs) and MHPs were invited to complete one of two web-based surveys. The CM survey was advertised to the general population of people aged 16 years and older who reside in Australia. The MHP survey was advertised to MHPs who reside in Australia. The survey was advertised on social media platforms including LinkedIn, Instagram, and Facebook using a snowballing method for 8 weeks between mid-February and mid-April 2024.

Procedure

The web-based survey was administered using Qualtrics XM (Qualtrics). After accessing the survey link, interested potential participants were screened for eligibility (aged older than 16 years and residing in Australia).

Measures

The survey included questions regarding the following topics:

- Participant characteristics: demographics for both surveys and clinical service use, and Kessler 10-Item Scale (K10) [24] mental health measure for the CMs survey only. The K10 is scored between 10 and 50. The score ranges are normal (10-19), mild distress (20-24), moderate distress (25-30), and severe distress (>30).
- Technology comfort, attitudes, and use: MHPs and CMs comfort with technology, their attitudes to AI using the AI attitudes scale [25], their interest in AI and their intention for future AI use. The AI attitudes scale is a 4-item scale which asks about general attitudes to AI. Each item is scored between 1 and 10, and the total score is the average score of the 4 items. It has good internal consistency with a Cronbach α of 0.82.
- AI use cases: a number of exemplar AI use cases (see Multimedia Appendix 1) to (1) support CMs mental health and well-being, and (2) to support their MHPs in performing their work duties. CMs and MHPs were presented with a series of potential use cases where AI could assist with specific tasks. Respondents were asked to rate on a scale of 0 to 10 how likely it was that they would use AI for the specific use case.
- Use, experienced benefits and harms: questions for the subset of those who have used AI tools pertaining to their experiences of benefit, harm, and risk.
- Free-text, open-ended responses to both groups about what excites or concerns them regarding the use of AI for mental health care.

The survey is available from the authors upon request.

Statistical Analyses

Quantitative data were analyzed using descriptive statistics in SPSS (version 22.0; IBM Corp). The sentiment and thematic elements in free-text responses within free-text responses were manually sorted by classifying responses into positive, negative, or neutral categories based on specific keywords and context. Each response was read in its entirety, and sentiments were categorized as positive, negative, or neutral based on the presence of specific keywords and the overall context of the response. Positive sentiments were identified by words such as "hope," "benefit," and "optimistic," among others. Negative sentiments were indicated by terms like "concern," "risk," and "fear." Responses lacking clear sentiment indicators or expressing ambiguous feelings were classified as neutral. Concurrently, thematic analysis [26] was performed to identify and interpret patterns within the data. This involved an iterative process of reading and coding the data, generating initial codes, and collating codes into potential themes. Themes were reviewed and refined through discussions among the research team to ensure they accurately represented the data set.

Ethical Considerations

Ethical approval for the use of the data was obtained from the University of Melbourne Human Research Ethics Committee (reference 2024-27805-48669-4). This study complied with the Declaration of Helsinki.

Results

Sample Characteristics

The final sample consisted of 107 CMs and 86 MHPs. Demographic characteristics of both samples are presented in Table 1. The mean age of both groups was equivalent. The majority of CMs were employed or studying. MHPs tended to have higher levels of formal education than CMs. For MHPs, the majority were either clinical or general psychologists (21/86, 40%). For CMs, the majority had a previous mental health diagnosis or significant difficulties with their mental health or emotional well-being (76/108, 70%), and the majority had also seen a professional for these difficulties (74/108, 69%). The mean K10 score was 22.8 (SD 8.9), indicating mild levels of psychological distress.

Table . Characteristics of CMs[a] and MHPs[b].

Demographic characteristics		CMs (n=108)	MHPs (n=86)
Age (years), mean (SD)		36.9 (16.2)	41.7 (10.9)
Gender, n (%)			
	Male	29 (26.9)	27 (31.4)
	Female	71 (65.7)	59 (67)
	Nonbinary, gender diverse, or nonconforming	5 (4.6)	0 (0)
	Prefer not to say	3 (2.8)	0 (0)
Aboriginal or Torres Strait Islander, n (%)			
	Yes	4 (3.7)	0 (0)
	No	103 (95.4)	84 (98)
	Prefer not to say	1 (0.9)	2 (2)
Employment, n (%)			
	Employed	69 (63.9)	—[c]
	Student	20 (18.5)	—
	Not in the labor force (looking for work, volunteer work, pensioner, or home duties)	19 (17.6)	—
Income after tax (Aus $; conversion rate of Aus $1=US $0.68 is applicable), n (%)			
	<$45,000	41 (38.3)	—
	$45,001–$120,000	48 (44.9)	—
	>$120,001	19 (16.8)	—
Highest level of education, n (%)			
	High school or equivalent	21 (19.6)	0 (0)
	Technical and further education or associate degree	18 (16.8)	3 (3.5)
	Bachelor degree	29 (27.1)	18 (21)
	Postgraduate diploma or graduate certificate	15 (13.1)	8 (9)
	Masters degree	17 (15.9)	44 (51)
	Doctoral degree or Doctor of Philosophy	8 (7.5)	13 (15)
Profession, n (%)			
	Clinical psychologist	—	21 (24)
	General practitioner	—	1 (1)
	Generalist psychologist	—	13 (15)
	Mental health management	—	2 (2)
	Mental health nurse	—	15 (17)
	Occupational therapist	—	4 (5)
	Peer or lived experience worker	—	3 (3.5)
	Psychiatrist or psychiatry registrar	—	4 (5)
	Social worker	—	19 (22)
	Therapist or counselor	—	4 (5)
Clinical and service use characteristics			
	Ever had a previous diagnosis or had significant difficulties with your mental health or emotional well-being? n (%)		
	Yes	76 (70.4)	—
	No	32 (29.6)	—

Demographic characteristics		CMs (n=108)	MHPs (n=86)
Have you ever seen a health professional for mental health concerns? n (%)			
	Yes	74 (68.5)	—
	No	29 (26.9)	—
	Missing	5 (4.6)	—
K10, mean (SD)		22.8 (8.9)	—

[a]CM: community member.

[b]MHP: mental health professional.

[c]Not applicable.

Technology Comfort, AI Attitudes, and AI Use Intention

In terms of comfort with using digital technology, 79% (68/86) of MHPs and 82% (89/108) of CMs rated themselves as being very comfortable, somewhat comfortable, or comfortable, whereas 22% (19/86) of MHPs and 18% (19/108) of CMs described themselves as being somewhat or very uncomfortable.

Table 2 shows responses to the AI attitudes scale. CMs had neutral attitudes and MHPs tended to have more positive attitudes toward AI across all measured dimensions. Tables 3 and 4 show that MHPs also tend to be more interested in using AI and more are more likely to use it in the future for work purposes than CMs are to use AI to manage emotional and mental well-being.

Table . AI[a] attitudes scale for CMs[b] and MHPs[c].

AI attitudes scale (1=not at all; 10=completely agree)	CMs (n=95), mean (SD)	MHPs (n=82), mean (SD)
I believe that AI will improve my life	5.15 (2.7)	6.62 (2.5)
I believe that AI will improve my work	5.52 (3.0)	6.70 (2.8)
I think I will use AI technology in the future	6.79 (3.0)	7.63 (2.4)
I think AI technology is positive for humanity	5.05 (2.7)	6.00 (2.4)
Average score	5.63 (2.5)	6.74 (2.3)

[a]AI: artificial intelligence.

[b]CM: community member.

[c]MHP: mental health professional.

Table . Community member interest in the use of artificial intelligence.

Questions		Values
How interested are you in using AI to support your mental health and emotional well-being? (n=95), n (%)		
	Not interested at all	26 (27)
	Slightly interested	16 (17)
	Somewhat interested	23 (24)
	Moderately interested	13 (14)
	Extremely interested	17 (18)
How likely are you to use AI tools in future to support your mental health and emotional well-being? (n=89), n (%)		
	Very unlikely	15 (17)
	Unlikely	9 (10)
	Somewhat unlikely	9 (10)
	Neither likely no unlikely	13 (15)
	Somewhat likely	21 (24)
	Likely	15 (17)
	Very likely	7 (8)
How likely are you to use AI for the following (0 - 10)? mean (SD)		
	Mood tracking	5.44 (3)
	Therapeutic chatbots	4.48 (3.3)
	Personalized recommendations	5.72 (3)
	Early detection and monitoring	5.28 (3.2)
	Crisis intervention support	4.47 (3.1)

Table . Mental health professional interest in the use of artificial intelligence.

Questions		Values
How interested are you in using AI to assist with tasks in your role as a mental health professional? (n=82), n (%)		
	Not interested at all	8 (10)
	Slightly interested	9 (11)
	Somewhat interested	13 (16)
	Moderately interested	23 (28)
	Extremely interested	29 (35)
How likely are you to use these and other AI tools in future to support your work? (n=74), n (%)		
	Very unlikely	3 (4)
	Unlikely	3 (4)
	Somewhat unlikely	5 (7)
	Neither likely no unlikely	10 (14)
	Somewhat likely	17 (23)
	Likely	13 (18)
	Very likely	23 (31)
How likely are you to use AI for the following (0‐10), mean (SD)		
	Assessment and diagnosis	6.12 (3)
	Provide personalized treatment recommendations to clients	6.14 (2.9)
	Track and guide client progress	7.00 (2.6)
	Enhancing client engagement	5.94 (3)
	Administrative assistance	8.16 (2)
	Literature and research analysis	8.07 (2.2)
	Training and simulation	7.43 (2.5)

AI Use Cases

Table 3 shows that CMs tended to rate their likelihood of using AI for a range of tasks associated with managing their emotional and mental well-being midway between unlikely and likely. The use cases that were most to least popular were (1) providing personal recommendations, (2) mood tracking, (3) detecting early warning signs, (4) use of therapeutic chatbots, and (5) crisis or suicide prevention support. MHPs on the other hand rated themselves as more likely to use AI in use cases across the board. The use cases that were most to least popular were (1) administrative tasks support, (2) synthesizing the latest clinical evidence, (3) training and simulation, (4) tracking consumer progress, (5) personalized recommendations for clients, (6) assisting with assessment and diagnosis, and (7) enhancing consumer engagement with treatment.

Use of AI

In total, 30 CMs (30/108, 28%) and 37 MHPs (37/86, 43%) reported use of AI in the previous 6 months. Of those, ChatGPT was the most common AI tool used by both CMs (16/30, 52%) and MHPs (20/37, 54%). Table S1 in Multimedia Appendix 1 outlines the reasons respondents provided for using these tools, as well as their experienced benefits, harms, and concerns. Further, 77% (23/30) of CMs and 92% (34/37) of MHPs reported AI to be very beneficial, somewhat beneficial, or beneficial, whereas 10% (3/30) of CMs and 3% (1/37) of MHPs found AI to be very harmful, somewhat harmful, or harmful. CMs mainly used these tools to obtain quick advice when emotionally distressed (18/30, 60%) or as a personal therapist or coach they could converse with to help manage their emotional and mental health (14/30, 47%). The most reported benefits were their availability (20/30, 67%), their low cost compared to therapy (18/30, 60%), and their privacy (16/30, 53%). About half of CMs (16/30, 53%) reported that they did not experience harms or concerns. The rest reported a range of concerns, such as responses being too general or not personalized (11/30, 37%) being unsure where their data was going (11/30, 37%). MHPs primarily used these tools to research mental health topics (24/37, 65%) and to assist with report and letter writing (20/37, 54%). Most reported it being helpful (25/37, 68%) and time saving (25/37, 68%). No harms or concerns were experienced by 49% (18/37); however, the rest reported concerns such as the outputs being too general (12/37, 32%), outputs being inaccurate (10/37, 27%), and being uncertain about the ethics of using these tools for these professional purposes (9/37, 24%).

Themes and Subthemes of Content Analysis in Free-Text Responses

Respondents were invited to share any concerns or interests they had regarding the use of AI for their specific purposes. A total of 66 responses were received from CMs, and 50 responses were received from MHPs. Among CMs, sentiment was rated as positive in 13 (20%) comments, negative in 17 (26%) comments, and neutral in 38 (58%) comments. Of those with positive sentiment, most were excited about AI making mental health care more accessible and efficient, more personalized, and better integrated with other technologies. The content of their negative sentiment was the lack of human support, errors and misdiagnosis, and ethical or data privacy concerns.

For MHPs of the 50 responses sentiment was rated as positive in 12 (24%) comments, negative in 13 (26%) comments, and neutral in 25 (50%) comments. Positive sentiment themes involved increase in efficiency and therefore increased accessibility of mental health care, as well as advanced diagnostics and treatment outcomes. Negative sentiment comments involved concerns about data governance and security, misuse by clinicians, and regulatory challenges (Tables 5 and 6).

Table . Analysis of community memebers' positive and negative sentiment themes on the future of AI[a] use in mental health care.

Theme		Description	Quote	Community members, n (%)
Positive sentiment				13 (20)
	Optimism about accessibility and efficiency	Many are excited about AI's potential to make mental health care more accessible and efficient, especially in underserved or remote areas.	"AI can provide constant, instant, and affordable support for everyone who needs it."	12 (92)
	Excitement about technological advancements	Some express a general excitement about the integration of cutting-edge technology in mental health and its potential to revolutionize care.	"There are so many possibilities and it can be revolutionary for mental health diagnosis and treatment."	10 (77)
	Potential for personalized care	There is enthusiasm for how AI can personalize treatment plans based on individual needs and historical data	"Tailored support for young people to be supported in a way that suits them."	7 (54)
Negative sentiment				17 (26)
	Concerns about lack of human connection	Respondents express concern that AI might not provide the empathetic and nuanced interaction that a human therapist offers	"Lack of human connection increasing the issues that harm mental health in the first place."	10 (59)
	Ethical and privacy concerns	Concerns regarding the ethical use of AI and data privacy issues are significant, with worries about how sensitive data is handled.	"Concerned about the privacy of therapy sessions when AI is involved."	9 (53)
	Worries about misdiagnosis or lack of sensitivity	Some fear that AI may not correctly interpret complex human emotions and could lead to misdiagnosis or inappropriate treatment suggestions.	"AI not being able to pick up on serious distress signals that a human would notice."	8 (47)

[a]AI: artificial intelligence.

Table . Analysis of mental health professionals' positive and negative sentiment themes on the future of AI[a] use in mental health care for mental health professionals.

Theme		Description	Quote	Mental health professionals, n (%)
Positive sentiment				12 (24)
	Technological potential and benefits	Positive views on how AI can enhance the efficiency, accessibility, and quality of mental health care.	"The potential to deliver quality, timely, relevant health care information that allows patient to make more informed choices for their treatment."	12 (100)
	Technological advancements	Excitement about specific AI technologies that may improve mental health diagnostics and treatment.	"Big data simulations of neural processing; simulations of neurolinguistic indicators of treatment engagement and response, LLM[b]-based mental health co-pilots for both clinicians and patients, no more bloody referral letters!"	5 (42)
Negative sentiment				13 (26)
	Risks and misuse	Concerns over potential negative impacts of AI, including risks of misuse by clinicians.	"I have some concern that clinicians may overly rely on AI decisions or outputs that they do not critically analyse the outputs when they make clinical decisions."	13 (100)
	Ethical and regulatory challenges	Concerns about the lack of adequate ethical guidelines and regulations for AI in health care and the interaction with registered professionals	" …there need to be enough guardrails to make it safe." "there is no clinical judgement in AI and the use of this to replace things only clinicians should be practicing."	5 (38.5)
	Data governance and security	Concerns about how data is managed and protected, focusing on issues like privacy, security, and confidentiality.	"Data governance will be tricky." "…AI can be used and abused by companies" "data can be shared and sold and then used to manipulate people." "Confidentiality and only as good as the data in the internet- reflects status quo not creative potential."	4 (31)

[a]AI: artificial intelligence.

[b]LLM: large language model.

Discussion

Principal Findings

This study is to our knowledge the first to survey both CMs and MHPs on their patterns of use, experiences, and perceived benefits and harms associated with the application of AI technologies in mental health care. This analysis provides a critical insight into how AI is currently being used to support mental health care from the perspective of CMs and MHPs, which may inform technological development and guide ethical, professional, and policy initiatives.

Attitudes to AI between the groups varied. CMs scored similarly to published community norms on the AI Attitudes Scale [25] (full scale average score 5.63, SD 2.5 vs 5.54, SD 1.78), while MHPs scored significantly higher (full scale average score 6.74, SD 2.3 vs 5.54, SD 1.78). AI use cases for CMs also had lower levels of endorsement than AI use cases for MHPs. Of note, all CM use cases involved scenarios where AI would be used directly to support personal mental health, whereas MHP use cases were split between indirect or administrative professional tasks and direct client mental health support tasks, the former being more likely to be endorsed. Potentially this is due to direct client use cases conceivably carrying more risk, making CMs and MHPs alike wary of using AI in this way. The intended purpose of commercial AI tools is also more aligned with professional support functionalities than direct mental health care applications. The difference highlights a significant area for future development and the necessity of balancing technological advancement with the training and education of both MHPs and CMs in safe use. The difference in use case endorsements also demonstrates that MHPs and CMs experience different pain points in their day-to-day lives. In the challenging context of embedding new technologies into mental health practice, equal consideration should be given to how AI technology can address these pain points for both groups.

Regarding actual use, we found that AI tools, most commonly ChatGPT, were used by around a third of CMs and 40% of MHPs. CMs tended to use these tools to obtain quick mental health advice or to receive emotional support, and nearly half used them as a personal coach or therapist, reporting the benefits being accessibility, privacy, and low-cost. These tools and associated AI techniques have been recognized for their potential to make mental health care more accessible, accurate, and efficient [14,27]. It is important to note, however, that these commercially available AI tools are not intended for such purposes, and as such, may present predictable and unpredictable risks [27]. About half reported experiencing harms or concerns as a result of use, noting that responses were too general, nonpersonalized, inaccurate, or unhelpful. Further, a lack of clarity regarding data security and the ethics of using AI tools in this way was also reported. These kinds of issues can create harm in some cases. For example, Tessa, an integrated rule-based and large language model AI chatbot designed to support patients with eating disorders [28] had to be withdrawn when it started to provide weight loss advice that ran counter to eating disorder clinical guidelines, sparking calls for greater regulatory measures for the safety of these tools in those and other contexts [29]. Survey respondents also expressed concerns regarding the lack of human support and potential for misdiagnosis, which reflects a need for cautious and informed integration of AI into mental health practices, particularly where AI is used to predict illness or risk states [14].

MHPs report using AI tools to support research, administrative, and training tasks, with a substantial number reporting time-saving benefits. Nonetheless, approximately a third of MHPs indicated concerns about the generality and potential inaccuracy of AI outputs, which emphasizes the need for ongoing scrutiny of the quality and application of AI in diverse settings for a wide range of purposes [27,30]. These findings align with the broader discourse in general health care delivery on the integration of AI into care, where efficiency and productivity gains must be balanced with accuracy, reliability, and ethical considerations [20,31].

The expressed concerns by all respondents regarding data governance, security, and the ethical implications of using AI tools in mental health care were notable. As AI technologies continue to advance, it is paramount that data security and ethical use are prioritized to protect both consumers and professionals, and to maintain trust in these tools [20]. As outlined by Luxton [32] a decade ago, psychologists and other mental health care professionals have an essential part to play in the development, evaluation, and ethical use of AI technologies. In a field still grappling to retrofit regulation for non-AI digital health tools, the rapid development of AI health technology and the associated avalanche of personal health data, “blackbox” processing, and data sharing requires swift action to put the necessary safeguard structures in place.

This study has some limitations. First, the online recruitment strategy may have attracted more respondents familiar with technology, although approximately 1 in 5 reported some discomfort with technology use. Second, the relatively small sample size and recruitment method means that the results may not be fully representative of the broader population and therefore limit the generalizability of the findings. This limitation is expected to be addressed in subsequent similar surveys, which will track the acceptability and perceived concerns and issues of using AI in mental health care, over time. Third, the reporting of benefits and harms may be an underestimate as use may have hidden or time-delayed effects. Nevertheless, the findings provide a useful insight into how AI is both currently perceived and experienced by users. Future research could build on these preliminary findings with larger and more diverse samples, potentially through cross-jurisdictional studies that can provide a more comprehensive view of the impact of AI on mental health care.

Conclusions

Our study underscores the promise and challenges of AI in mental health care. As AI tools evolve, it is essential that they are developed with ethics, inclusivity, accuracy, safety and the genuine needs of end users in mind. This will not only guide technological advancement but also ensure that AI serves as a valuable complement to overwhelmed traditional mental health services, ultimately improving outcomes and efficiencies for all stakeholders involved.

Acknowledgments

The authors would like to thank Ali Morrow, BBusJourn, Fern Nicholls, BA, and Prof Blake Dear, PhD, for their assistance in disseminating the survey. We are also grateful to the people who gave their time to complete the survey.

Data Availability

Deidentified data can be made available upon request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Tables for artificial intelligence (AI) tool experience for the subset of the community members sample who used AI and AI tool experience for the subset of the mental health professionals sample who used AI tools.

[DOCX File, 23 KB - mental_v11i1e60589_app1.docx]

References

1. Gore FM, Bloem PJ, Patton GC, et al. Global burden of disease in young people aged 10–24 years: a systematic analysis. Lancet 2011 Jun;377(9783):2093-2102. [doi: 10.1016/S0140-6736(11)60512-6]
2. Wang PS, Angermeyer M, Borges G, et al. Delay and failure in treatment seeking after first onset of mental disorders in the World Health Organization's World Mental Health Survey initiative. World Psychiatry 2007 Oct;6(3):177-185. [Medline: 18188443]
3. Whiteford HA, Buckingham WJ, Harris MG, et al. Estimating treatment rates for mental disorders in Australia. Aust Health Rev 2014 Feb;38(1):80-85. [doi: 10.1071/AH13142] [Medline: 24308925]
4. National study of mental health and wellbeing. Australian Bureau of Statistics. URL: https://www.abs.gov.au/statistics/health/mental-health/national-study-mental-health-and-wellbeing/latest-release [accessed 2023-05-22]
5. Kavanagh BE, Corney KB, Beks H, Williams LJ, Quirk SE, Versace VL. A scoping review of the barriers and facilitators to accessing and utilising mental health services across regional, rural, and remote Australia. BMC Health Serv Res 2023 Oct 4;23(1):1060. [doi: 10.1186/s12913-023-10034-4] [Medline: 37794469]
6. Green AE, Albanese BJ, Shapiro NM, Aarons GA. The roles of individual and organizational factors in burnout among community-based mental health service providers. Psychol Serv 2014 Feb;11(1):41-49. [doi: 10.1037/a0035299] [Medline: 24564442]
7. O'Connor K, Muller Neff D, Pitman S. Burnout in mental health professionals: a systematic review and meta-analysis of prevalence and determinants. Eur Psychiatry 2018 Sep;53:74-99. [doi: 10.1016/j.eurpsy.2018.06.003] [Medline: 29957371]
8. Lupton D. The digitally engaged patient: self-monitoring and self-care in the digital health era. Soc Theory Health 2013 Aug;11(3):256-270. [doi: 10.1057/sth.2013.10]
9. Bell IH, Thompson A, Valentine L, Adams S, Alvarez-Jimenez M, Nicholas J. Ownership, use of, and interest in digital mental health technologies among clinicians and young people across a spectrum of clinical care needs: cross-sectional survey. JMIR Ment Health 2022 May 11;9(5):e30716. [doi: 10.2196/30716] [Medline: 35544295]
10. Nicholas J, Bell IH, Thompson A, et al. Implementation lessons from the transition to telehealth during COVID-19: a survey of clinicians and young people from youth mental health services. Psychiatry Res 2021 May;299:113848. [doi: 10.1016/j.psychres.2021.113848] [Medline: 33725578]
11. Valentine L, D'Alfonso S, Lederman R. Recommender systems for mental health apps: advantages and ethical challenges. AI Soc 2023 Aug;38(4):1627-1638. [doi: 10.1007/s00146-021-01322-w]
12. McCradden M, Hui K, Buchman DZ. Evidence, ethics and the promise of artificial intelligence in psychiatry. J Med Ethics 2023 Aug;49(8):573-579. [doi: 10.1136/jme-2022-108447] [Medline: 36581457]
13. D'Alfonso S. AI in mental health. Curr Opin Psychol 2020 Dec;36:112-117. [doi: 10.1016/j.copsyc.2020.04.005] [Medline: 32604065]
14. Graham S, Depp C, Lee EE, et al. Artificial intelligence for mental health and mental illnesses: an overview. Curr Psychiatry Rep 2019 Nov 7;21(11):116. [doi: 10.1007/s11920-019-1094-0] [Medline: 31701320]
15. Hirschberg J, Manning CD. Advances in natural language processing. Science 2015 Jul 17;349(6245):261-266. [doi: 10.1126/science.aaa8685] [Medline: 26185244]
16. Durden E, Pirner MC, Rapoport SJ, Williams A, Robinson A, Forman-Hoffman VL. Changes in stress, burnout, and resilience associated with an 8-week intervention with relational agent "Woebot.". Internet Interv 2023 Sep;33:100637. [doi: 10.1016/j.invent.2023.100637] [Medline: 37635948]
17. Malik A. OpenAI's ChatGPT now has 100 million weekly active users. TechCrunch. 2023 Nov 6. URL: https://techcrunch.com/2023/11/06/openais-chatgpt-now-has-100-million-weekly-active-users/ [accessed 2024-08-30]
18. Tan S. Awareness versus usage of ChatGPT in Australia: how do they vary demographically? YouGov. 2023 Jun 8. URL: https://business.yougov.com/content/8400-awareness-versus-usage-of-chatgpt-in-australia-how-do-they-vary-demographically [accessed 2023-08-12]
19. Denejkina A. Young people's perception and use of generative AI. Youth Insight. 2023 Jun 27. URL: https://youthinsight.com.au/education/young-peoples-perception-and-use-of-generative-ai/ [accessed 2023-08-12]
20. Blease C, Torous J. ChatGPT and mental healthcare: balancing benefits with risks of harms. BMJ Ment Health 2023 Nov;26(1):e300884. [doi: 10.1136/bmjment-2023-300884] [Medline: 37949485]
21. Doraiswamy PM, Blease C, Bodner K. Artificial intelligence and the future of psychiatry: insights from a global physician survey. Artif Intell Med 2020 Jan;102:101753. [doi: 10.1016/j.artmed.2019.101753] [Medline: 31980092]
22. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]

23. Gooding P. Mapping the rise of digital mental health technologies: emerging issues for law and society. Int J Law Psychiatry 2019;67:101498. [doi: 10.1016/j.ijlp.2019.101498] [Medline: 31785726]
24. Kessler RC, Andrews G, Colpe LJ, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002 Aug;32(6):959-976. [doi: 10.1017/s0033291702006074] [Medline: 12214795]
25. Grassini S. Development and validation of the AI attitude scale (AIAS-4): a brief measure of general attitude toward artificial intelligence. Front Psychol 2023 Jul 24;14:1191628. [doi: 10.3389/fpsyg.2023.1191628] [Medline: 37554139]
26. Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008 Jul 10;8:45. [doi: 10.1186/1471-2288-8-45] [Medline: 18616818]
27. Lee EE, Torous J, De Choudhury M, et al. Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. Biol Psychiatry Cogn Neurosci Neuroimaging 2021 Sep;6(9):856-864. [doi: 10.1016/j.bpsc.2021.02.001] [Medline: 33571718]
28. Chan WW, Fitzsimmons-Craft EE, Smith AC, et al. The challenges in designing a prevention chatbot for eating disorders: observational study. JMIR Form Res 2022 Jan 19;6(1):e28003. [doi: 10.2196/28003] [Medline: 35044314]
29. Sharp G, Torous J, West ML. Ethical challenges in AI approaches to eating disorders. J Med Internet Res 2023 Aug 14;25:e50696. [doi: 10.2196/50696] [Medline: 37578836]
30. Russell RG, Lovett Novak L, Patel M, et al. Competencies for the use of artificial intelligence-based tools by health care professionals. Acad Med 2023 Mar 1;98(3):348-356. [doi: 10.1097/ACM.0000000000004963] [Medline: 36731054]
31. Esmaeilzadeh P. Use of AI-based tools for healthcare purposes: a survey study from consumers' perspectives. BMC Med Inform Decis Mak 2020 Jul 22;20(1):170. [doi: 10.1186/s12911-020-01191-1] [Medline: 32698869]
32. Luxton DD. Artificial intelligence in psychological practice: current and future applications and implications. Prof Psychol Res Pr 2014;45(5):332-339. [doi: 10.1037/a0034559]

Abbreviations

AI: artificial intelligence
CM: community member
K10: Kessler 10-Item Scale
MHP: mental health professional

Edited by J Torous; submitted 16.05.24; peer-reviewed by M Pulier, P Velmovitsky; accepted 30.07.24; published 11.10.24.

Please cite as:
Cross S, Bell I, Nicholas J, Valentine L, Mangelsdorf S, Baker S, Titov N, Alvarez-Jimenez M
Use of AI in Mental Health Care: Community and Mental Health Professionals Survey
JMIR Ment Health 2024;11:e60589
URL: https://mental.jmir.org/2024/1/e60589
doi:10.2196/60589

Original Paper

In-Person and Teleconsultation Services at a National Hospital in Peru: Time Series Analysis of General and Psychiatric Care Amid the COVID-19 Pandemic

David Villarreal-Zegarra[1,2], MPH; Jackeline García-Serna[2], BSc; Piero Segovia-Bacilio[2], BSc; Nikol Mayo-Puchoc[2], BSc; Alba Navarro-Flores[3], MSc; Jeff Huarcaya-Victoria[4,5], MSc, MD

[1]Escuela de Medicina Humana, Universidad César Vallejo, Trujillo, Peru

[2]Instituto Peruano de Orientación Psicológica, Lima, Peru

[3]International Max Planck Research School for Translational Psychiatry (IMPRS-TP), Munich, Germany

[4]Escuela Profesional de Medicina Humana, Universidad Privada San Juan Bautista, Ica, Peru

[5]Unidad de Investigación de Psiquiatría, Departamento de Psiquiatría, Hospital Nacional Guillermo Almenara Irigoyen, Lima, Peru

Corresponding Author:
Jeff Huarcaya-Victoria, MSc, MD
Escuela Profesional de Medicina Humana, Universidad Privada San Juan Bautista
Carretera Panamericana Sur N° 103, 113 y 123
Ica
Peru
Phone: 51 950322888
Email: jeff.huarcaya@upsjb.edu.pe

Abstract

Background: The COVID-19 pandemic led to a global reduction in health care accessibility for both infected and noninfected patients, posing a particular burden on those with chronic conditions, including mental health issues. Peru experienced significant devastation from the pandemic, resulting in a collapsed health care system and leading to the world's highest per capita mortality rate as a result of COVID-19. Understanding the trends in health care utilization, particularly in mental health care, is crucial for informing pandemic response efforts and guiding future recovery strategies.

Objective: This study aims to analyze the trends of outpatient medical and psychiatric consultations during the COVID-19 pandemic in a national hospital in Peru.

Methods: This observational study was conducted at a national hospital in Lima, Peru. We analyzed data on user care across all services, including psychiatric services, from May 2019 to December 2022. The data were calculated for users served per month, including the number of users seen monthly in mental health services. Sociodemographic variables such as sex (female or male), age (≥0 years), type of medical appointment (regular or additional), and modality of care (in-person or teleconsultations) were taken into account. An interrupted time series regression model was conducted to assess the number of outpatient medical and psychiatric consultations. Subgroup analyses were performed based on service modality, including overall consultations, telemonitoring/teleconsultations only, or face-to-face only, for all service users and for mental health service users.

Results: A total of 1,515,439 participants were included, with females comprising 275,444/484,994 (56.80%) of the samples. Only 345,605/1,515,439 (22.81%) visits involved telemedicine. The total monthly outpatient visits were significantly reduced compared with the expected projection (*P*<.001) at the beginning of the pandemic, followed by a later monthly increment of 298.7 users. Face-to-face interventions experienced a significant reduction at the beginning of the pandemic (*P*<.001), gradually recovering in the following months. By contrast, telemedicine use initially increased but subsequently declined toward the end of the pandemic. A similar trend was observed in mental health units.

Conclusions: During the pandemic years, health care utilization in both general and psychiatric services experienced a significant decrease, particularly at the beginning of the pandemic (March 2020). However, no significant trends were observed in either case throughout the pandemic period. Telemedicine consultations witnessed a significant increase overall during this period, particularly among mental health users.

(JMIR Ment Health 2024;11:e53980) doi:10.2196/53980

KEYWORDS

health care utilization; mental health use; COVID-19; mental health; health care; psychiatric care; teleconsultation; hospital; Peru; chronic; patient; patients; telemonitoring

Introduction

During the COVID-19 pandemic, many medical consultations unrelated to the virus were significantly reduced to focus on containing the spread of the infection [1]. Health care systems faced the significant challenge of adapting to the increasing demand for treating patients with COVID-19 while continuing to provide care for those with other conditions, despite having limited resources available [2]. Outpatient clinics, in particular, experienced a reduction in utilization during this time, primarily as a result of the decrease in scheduled medical visits and the temporary suspension of some services in many health centers [3]. This situation significantly impacted patients' health, especially those with chronic conditions requiring continuous follow-up sessions, and led to substantial economic losses [4].

In low- and middle-income countries, where access to medical care was already limited, the pandemic posed a nearly insurmountable challenge [5,6]. The prevalence of other public health issues, such as malnutrition, malaria, tuberculosis, and HIV/AIDS, rose to unprecedented levels [7,8]. The case of Peru exemplifies unmet health care needs during the pandemic in the context of a severely unprepared, yet upper-middle-income country. With hospitals suffering from poor infrastructure, limited medical supplies, and outdated technology for decades [9], Peru's health care system collapsed during the pandemic. This led to the highest mortality rate per capita, low immunization rates, and a high frequency of infections among health care providers [10,11]. As a consequence of this situation, the outcomes for patients with chronic and noncommunicable disorders worsened [12].

An important subset of chronic patients included those with psychiatric disorders. For them, various aspects of the pandemic—such as fear of contagion, new diagnoses, and the loss of loved ones—combined with the inadequate health care response led to an increase in the incidence and severity of cases [13]. In Peru, mental health and substance abuse disorders are among the most burdensome conditions [14]. Following the COVID-19 pandemic, there has been an increase in the prevalence of moderate depressive symptoms and in the proportion of cases being treated for mild depressive symptoms [15-17].

An alternative to managing the high demand of patients was provided by telemedicine and other digital technologies, which improved accessibility and the quality of care [18,19]. However, a significant problem remains: many patients lack access to the necessary technology for telemedicine and other forms of virtual medical care utilized during the pandemic.

Hence, it is crucial to examine how the utilization of outpatient medical consultations has been affected by the COVID-19 pandemic in these countries. This exploration can help identify effective solutions to tackle present and future challenges in health care delivery. In this regard, this study will concentrate on analyzing the utilization of outpatient medical and psychiatric consultations during the COVID-19 pandemic at a social security hospital in Peru. It will use a time series analysis methodology to scrutinize trends and patterns of service utilization from 2019 to 2022.

Methods

Study Design

Our study adopted an observational design and utilized data from the Hospital Nacional Guillermo Almenara Irigoyen (HNGAI) in Lima, Peru.

Setting

The study was conducted at the HNGAI, a tertiary referral center and highly complex health care facility situated in Lima, Peru. The HNGAI occupies a prominent position within Peru's health care system, being one of the largest hospitals under the social security system in terms of bed capacity, boasting a total of 960 hospital beds. It serves as a comprehensive medical institution catering to a broad range of medical specialties, including psychiatry.

The hospital's significance and scope of services are evident from its designation as a Specialized Health Institute III-2, the highest level bestowed by the Ministry of Health of Peru on hospital establishments. With a catchment population of 1,547,840 individuals covered by social insurance, the HNGAI plays a vital role in providing health care services to a substantial portion of the population.

As a tertiary referral center, the HNGAI assumes the responsibility for managing a diverse range of medical conditions, from routine ailments to highly complex and specialized cases. Its comprehensive capabilities and expertise make it a preferred choice for patients seeking specialized medical care.

Data on user care across all services provided by the HNGAI, including psychiatric services such as adult psychiatry, child and adolescent psychiatry, addictive behaviors, and day hospital services provided by the psychiatry department, were utilized. The data were extracted from the "ExplotaDatos system" generated by the Social Health Insurance EsSalud, covering the period from May 1, 2019, to December 31, 2022.

Participants

The number of consultations considered for the study included all hospital services, also encompassing psychiatric services. Among these consultations, women accounted for 275,444/484,964 (56.80%) and 584,618/1,030,445 (56.73%) of the sample before and during the pandemic, respectively. The age range of participants in the study ranged from 12 to over 65 years. The number of participants with their own health insurance was 341,299/484,964 (70.38%) and 749,626/1,030,445 (72.75%) before and during the pandemic, respectively. Additionally, 391,604/484,964 (80.75%) and 911,384/1,030,445 (88.45%) participants accessed care through a regular medical appointment, while the remaining participants accessed care through an additional one. Mental health care services constituted 26,916/484,964 (5.55%) and 68,385/1,030,445 (6.64%) of the total consultations before and after the pandemic, respectively. Finally, telemonitoring and teleconsultation care accounted for 30/484,964 (0.01%) and 345,575/1,030,445 (33.54%) of the total consultations before and during the pandemic, respectively. Variables considered are listed in Textbox 1.

Textbox 1. Variables considered in this study.

- 1. Users Served Per Month

Defined as the number of users served in all services during a month by means of in-person consultations as well as telemonitoring or teleconsultation.

- 2. Number of Users Seen Monthly in Mental Health Services

Defined as the number of users attended to in the mental health services, including adult psychiatry, child and adolescent psychiatry, addictive behaviors, and day hospital, through in-person consultations as well as telemonitoring or teleconsultation.

- 3. Sociodemographic Covariates

The sociodemographic variables were sex (female or male), age (≥0 years), type of medical appointment (normal or additional), and modality of care (telemonitoring or teleconsultations).

Statistical Analysis

Descriptive analyses were conducted on users seen from May 2019 to December 2022, with monthly data collection. Sociodemographic variables were analyzed for all medical consultations, including those carried out by psychiatric services. We used interrupted time series regression models, a quasi-experimental approach, to assess the number of outpatient medical and psychiatric consultations following the onset of the COVID-19 pandemic. This involved analyzing trends and patterns of service utilization from May 2019 to December 2022. The analysis did not encompass care provided by the psychology service. The estimated impact of the pandemic on total outpatient medical and psychiatric consultations was assessed in terms of changes in level (intercept) and changes in the slope of prevalence across the time series before, during, and after the pandemic. Subgroup analyses were conducted according to service modality, that is, overall, telemonitoring/teleconsultations only, or face-to-face only, for all service users and psychiatric service users.

Ethics Approval

We did not access any individual personal data, nor did we have contact with participants, as the data were secondary and anonymous. The data were accessed upon request, and we did not collect primary data, thus eliminating any ethical risk. Our study was approved by the Institutional Research Ethics Committee of the HNGAI (approval number 245 CIEI-IOIyD-GRPA-ESSALUD-2023).

Results

Participants

A total of 1,515,439 participants were attended to as outpatients between May 2019 and December 2022. During this period, the majority of participants were females (860,062/1,515,409, 56.75%), working-age adults (777,236/1,515,409, 51.29%), health insurance holders (1,090,925/1,515,409, 71.99%), and had scheduled regular medical appointments (1,302,988/1,515,409, 85.98%). Only 345,605/1,515,409 (22.81%) outpatient visits involved telemonitoring or teleconsultation. Table 1 presents the characteristics of the participants before and during the pandemic for all users and users in mental health services.

Table 1. Sociodemographic characteristics of the participants in all services.

Sociodemographic characteristics	All users served (n=1,515,439)		Psychiatric services (n=95,301)	
	Before the pandemic (n=484,964), n (%)	During the pandemic (n=1,030,445), n (%)[a]	Before the pandemic (n=26,916), n (%)	During the pandemic (n=68,385), n (%)[a]
Sex				
Female	275,444 (56.80)	584,618 (56.73)	12,747 (47.36)	34,534 (50.50)
Male	209,550 (43.21)	445,827 (43.27)	14,169 (52.64)	33,851 (49.50)
Age categories (years)				
0-12	53,984 (11.13)	94,521 (9.17)	4756 (17.67)	9424 (13.78)
13-17	14,500 (2.99)	30,970 (3.01)	2911 (10.82)	8509 (12.44)
18-64	234,214 (48.30)	543,022 (52.70)	13,741 (51.05)	37,067 (54.20)
65 or older	182,296 (37.59)	361,932 (35.12)	5508 (20.46)	13,385 (19.57)
Family relationship to the insured				
Others	143,695 (29.63)	280,819 (27.25)	12,929 (48.03)	33,838 (49.48)
Holder	341,299 (70.38)	749,626 (72.75)	13,987 (51.97)	34,547 (50.52)
Type of medical appointment				
Normal	391,604 (80.75)	911,384 (88.45)	20,180 (74.97)	60,565 (88.56)
Additional	93,390 (19.26)	119,061 (11.55)	6736 (25.03)	7820 (11.44)
Telemonitoring/teleconsultations				
No	484,964 (100)	684,870 (66.46)	26,916 (100)	39,300 (57.47)
Yes	30 (0.01)	345,575 (33.54)	0 (0)	29,085 (42.53)

[a]The lockdown started on March 16, 2020.

The most common mental health problem diagnoses in users of psychiatric services were schizophrenia, schizotypal, and delusional disorders (F20-F29); affective disorders (F30-F39); and anxiety, stress-related, and somatoform disorders (F40-F48). An extended list of the mental health diagnoses is provided in Table 2.

Table 2. Psychiatric diagnoses in psychiatric services (N=95,301).

ICD-10[a] diagnostic code: diagnosis	Before the pandemic (n=26,916), n (%)	During the pandemic (n=68,385), n (%)[b]
F00-F09: Organic mental disorders including symptomatic disorders	2548 (9.47)	5622 (8.22)
F10-F19: Mental and behavioral disorders due to psychoactive substance use	2051 (7.62)	4635 (6.78)
F20-F29: Schizophrenia, schizotypal, and delusional disorders	5262 (19.55)	14,553 (21.28)
F30-F39: Mood (affective) disorders	4797 (17.82)	13,132 (19.20)
F40-F48: Neurotic, stress-related, and somatoform disorders	4297 (15.96)	11,017 (16.11)
F50-F59: Behavioral syndromes associated with physiological disturbances and physical factors	236 (0.88)	812 (1.19)
F60-F69: Adult personality and behavioral disorders	656 (2.44)	1637 (2.39)
F70-F79: Mental retardation	504 (1.87)	1197 (1.75)
F80-F89: Developmental psychological disorders	2306 (8.57)	5471 (8.00)
F90-F98: Behavioral and emotional disorders often occurring in childhood and adolescence	3485 (12.95)	5639 (8.25)
F99-F99: Mental disorder not specified	0 (0)	0 (0)
Other diagnoses	774 (2.88)	4670 (6.83)

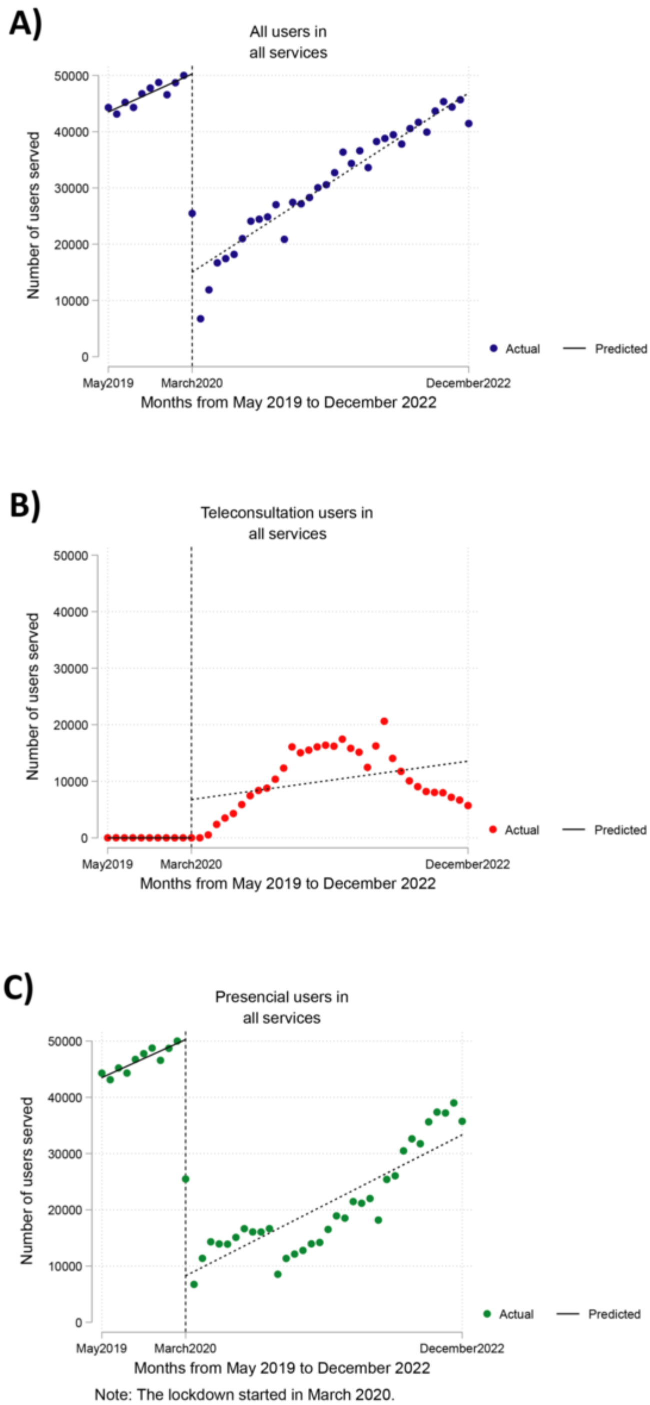
[a]ICD-10: 10th revision of the International Statistical Classification of Diseases and Related Health Problems.

[b]The lockdown started on March 16, 2020.

Number of Users in All Services

Figure 1A illustrates that the total number of users seen per month in outpatient care across all services experienced a significant decrease at the onset of the pandemic (March 2020), with a reduction of –35,226.4 users seen compared with what was expected for that month (95% CI –38,558.3 to –31,894.6; *P*<.001). However, an upward trend was observed during the pandemic, with a monthly increase of 289.7 users across all services (95% CI 67.2-512.2; *P*=.01). By contrast, the number of users served at the beginning of the pandemic remained the same as that measured in the month before the start of the pandemic (February 2020).

Figure 1. Interrupted time series analysis for the number of users served per month in all services between May 2019 and December 2022. (A) Interrupted time series analysis for users in all services. (B) Interrupted time series analysis for teleconsultation users in all services. (C) Interrupted time series analysis for face-to-face users in all services.



The number of users served by telemonitoring and teleconsultation was nearly 0 before the pandemic (Figure 1B). Subsequently, there was an increase in the middle of the pandemic followed by a decrease in the last months of the evaluation period. However, no significant trend was found in the number of users visited per month by telemonitoring and teleconsultation alone (*P*=.13).

The number of users served by face-to-face care per month experienced a significant reduction at the beginning of the pandemic (Figure 1C), with a decrease of –42,001.9 users served compared with what was expected for that month (95% CI –46,934.2 to –37,069.6; *P*<.001). However, no significant trend was observed in the number of users seen per month for face-to-face care alone (*P*=.58). Table 3 presents the coefficients of the time series analysis.

Table 3. Interrupted time series regression analysis for the number of users served.

User service	Coefficients[a]	*P* value[a]	95% CI
All users served			
Overall			
Preintervention slope[b]	675.5	<.001	531.5 to 819.5
Change in intercept[c]	*–35,226.4*	*<.001*	–38,558.3 to –31,894.6
Change in slope (interaction)[d]	*289.7*	*.01*	67.2 to 512.2
Intercept[e]	43,513.7	<.001	42,827.2 to 44,200.2
Postintervention linear trend[f]	965.2	<.001	812.0 to 1118.5
Only telemonitoring/teleconsultations			
Preintervention slope	–0.8	.03	–1.6 to –0.1
Change in intercept	*6775.5*	*.009*	1777.9 to 11,773.0
Change in slope (interaction)	206.3	.13	–64.8 to 477.5
Intercept	6.7	.01	1.5 to 11.9
Postintervention linear trend	205.5	.13	–65.7 to 476.7
Only in-person			
Preintervention slope	676.4	<.001	532.1 to 820.6
Change in intercept	*–42,001.9*	*<.001*	–46,934.2 to –37,069.6
Change in slope (interaction)	83.4	.585	–222.7 to 389.4
Intercept	43,507.0	<.001	42,818.3 to 44,195.8
Postintervention linear trend	759.7	<.001	506.5 to 1013.0
Psychiatric services			
Overall			
Preintervention slope	23.2	.17	–10.5 to 56.8
Change in intercept	*–1550.3*	*<.001*	–1818.2 to –1282.3
Change in slope (interaction)	31.2	.09	–5.3 to 67.6
Intercept	2471.0	<.001	2326.6 to 2615.5
Postintervention linear trend	54.4	<.001	42.8 to 65.9
Only telemonitoring/teleconsultations			
Preintervention slope	0.0	—[g]	—
Change in intercept	204.9	—	—
Change in slope (interaction)	39.4	—	—
Intercept	0.0	—	—
Postintervention linear trend	39.4	—	—
Only in-person			
Preintervention slope	23.2	.17	–10.5 to 56.8
Change in intercept	*–1755.2*	*<.001*	–2139.2 to –1371.2
Change in slope (interaction)	–8.2	.67	–46.9 to 30.5
Intercept	2471.0	<.001	2326.6 to 2615.5
Postintervention linear trend	14.9	.07	–1.4 to 31.3

[a]Values in italics are significant (*P*<.05).

[b]Preintervention slope corresponds to the previous trend of the number of users served.

[c]Change in intercept refers to the change in the number of users served at the beginning of the COVID-19 lockdown.

[d]Change in slope (interaction) refers to the change in the trend of the number of users served over time after March 2020.

[e]Intercept represents the number of users served at the beginning of the study period.

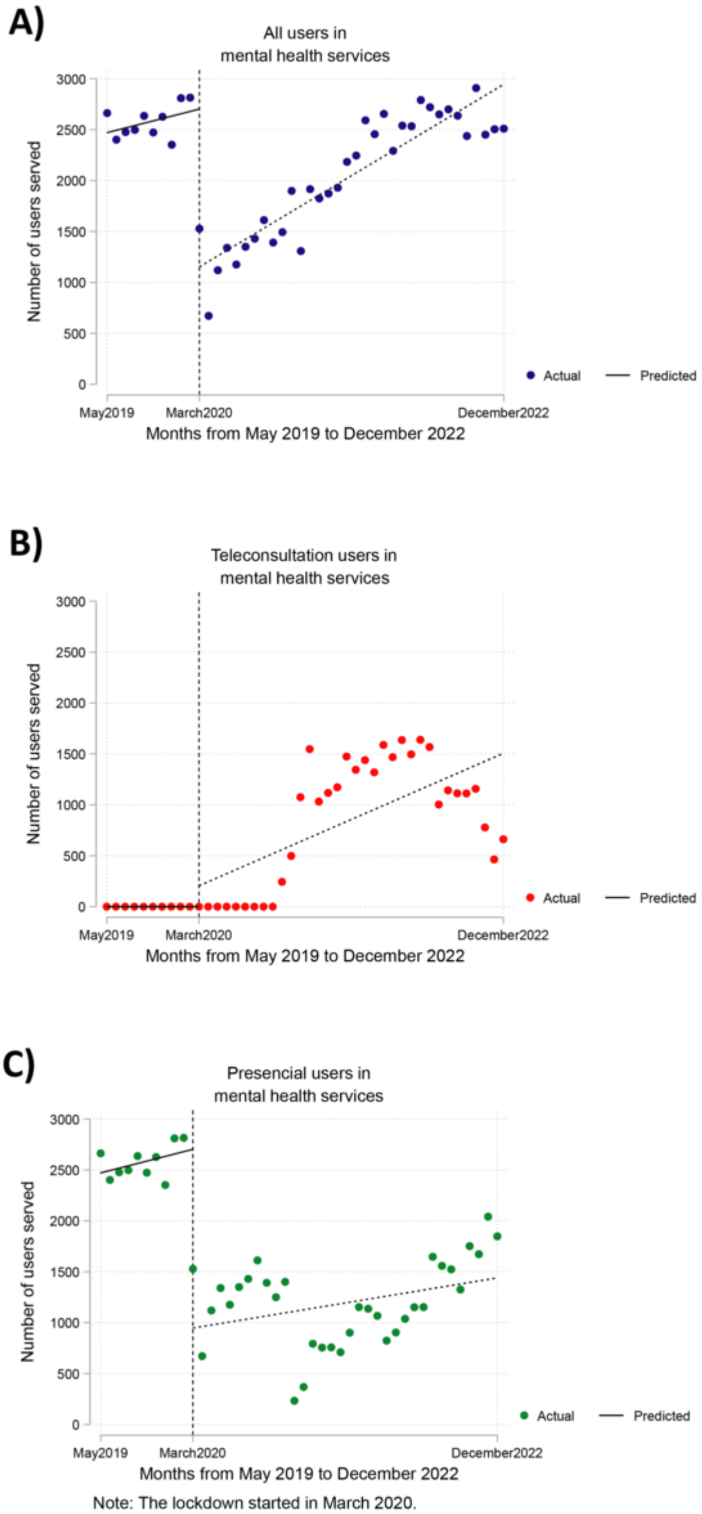
[f]Postintervention linear trend represents the trend in the number of users served after the onset of the pandemic. Autocorrelation at lag(1) was considered.

[g]The model did not converge so the analysis as such could not be performed.

Number of Users in Psychiatric Services

Figure 2A illustrates that the number of outpatients seen per month in the 4 psychiatric services experienced a significant reduction at the onset of the pandemic (March 2020). There were –1550.3 fewer users seen than expected for that month (95% CI –1818.2 to –1282.3; *P*<.001). However, although the number of users seen per month increased during the pandemic, no significant trend was found (*P*=.09).

Figure 2. Interrupted time series analysis for the number of users served per month in all psychiatric services between May 2019 and December 2022. (A) Interrupted time series analysis for users in mental health services. (B) Interrupted time series analysis for teleconsultation users in mental health services. (C) Interrupted time series analysis for face-to-face users in mental health services.



The number of users receiving telemonitoring and teleconsultation services in the 4 psychiatric services per month is indicated in Figure 2B. However, the time series analysis did not converge, making it impossible to assess whether there were significant changes (Table 3).

There was a significant reduction in the number of face-to-face user visits per month in the 4 psychiatric services at the beginning of the pandemic (Figure 2C). The reduction amounted to –1755.2 users compared with the expected number for that month (95% CI –2139.2 to –1371.2; *P*<.001). However, no significant trend was found during the pandemic for the number of face-to-face visits per month only (*P*=.67).

Discussion

Principal Findings

During the initial months of the pandemic, there was a decrease in the number of users accessing health services, which gradually increased over time. Before the pandemic, medical care through telemonitoring and teleconsultation was almost nonexistent, but it increased during the pandemic before decreasing again in the last months of the evaluated period. By contrast, there were no significant differences in the number of users attending face-to-face, telemonitoring, and teleconsultation modalities during the evaluation months. Regarding patients seen in psychiatric services, we observed a significant decrease in the number of outpatients at the beginning of the pandemic, followed by a progressive increase over time. Face-to-face visits per month were significantly reduced in all 4 mental health services.

Comparison With Other Studies

At the beginning of the pandemic, we observed an overall reduction in health care utilization. A systematic review conducted until August 2020, encompassing 20 countries, found that health care utilization was reduced in one-third of them. Eighty-one primary studies reported 143 estimates and concluded a mean reduction of 28% in admissions and 31% in new diagnoses [1]. In Peru, mandatory confinement was established in March 2020, lasting approximately 3 months [20]. During this period, the population was not allowed to leave their homes except for emergencies.

Amid the pandemic, we observed a gradual but slow increase in care utilization. A study using a time series design to assess the effect of the pandemic on 31 health services across 10 countries (low income=2, lower middle income=3, upper middle income=3, and high income=2) found that total outpatient visits decreased between 9% and 40% and remained below expectations by the end of 2020 [21]. A scoping review analyzing changes in medical care access found that among 38 studies from Europe, Asia, and Africa, 33 reported a statistically significant reduction in service use. Similarly, the studies reported possible barriers to health care access, including limited supplies and personnel to care for patients with non–COVID-19 having other medical issues, as well as increased waiting times. A potential explanation for this slow increment in Peru, as well as in other Latin American countries, could be provided by a survey reporting that around 66% of Peruvian participants attempted self-medication during the pandemic [22]. Additionally, part of the population faced barriers such as fear of contagion and the stigma of being diagnosed with COVID-19, as well as financial difficulties [23]. Although the number of cases was not increasing during and after the first phase of restriction, the maintenance of low rates of utilization was suggested to be a result of factors such as fear of contagion, sanctions related to outside mobilization, reduced access to medical centers, and prioritization of COVID-19 cases, among others [24]. We observed that by the end of the pandemic, overall health care utilization rates had risen to levels comparable to those observed 1 year before the pandemic's onset. However, a notable addition was the emergence of a new category of consultations conducted via telemedicine. Similarly, a study in the United Kingdom found that after the release of social restriction measures, the frequency of health care utilization returned to levels that were not significantly different from prepandemic utilization [25].

Regarding psychiatric services utilization, we observed a reduction at the beginning of the pandemic that progressively improved toward the end of this period. Amidst the pandemic, Peru was submerged in chaos because of various factors, including a debilitated health care system, political mismanagement, and poor adherence to social restrictions [11]. This scenario significantly impacted the mental health of the population, as evidenced by high levels of perceived stress in the general Peruvian population [26]. A time series analysis of depression diagnoses in Peru showed that the number of new patients increased by 0.17% per month after the beginning of the pandemic [15].

In the United States, during the initial months of the pandemic, there was a 50% reduction in face-to-face mental health encounters. However, the utilization of telehealth services played a pivotal role in the swift restoration of service delivery. This uptake of telehealth, accounting for approximately 47.9% of average monthly encounters, facilitated the timely provision of mental health care despite the limitations imposed by the pandemic [27]. In another study, it has been reported that the COVID-19 pandemic resulted in a significant decrease of more than 50% in in-person mental health care utilization rates among commercially insured adults. Rates of various mental health disorders, including anxiety disorders, bipolar disorder, and adjustment disorders, declined during the pandemic. By contrast, telehealth service utilization increased substantially, by 16-20 times. Combining both in-person and telehealth services, there was an overall increase in care observed for anxiety and adjustment disorders. These findings highlight the impact of the pandemic on mental health care delivery and the potential of telehealth in providing accessible services during times of crisis [28].

In the domain of psychiatric services, there was a discernible decline of 12% per week in visits to psychiatric emergency wards during the initial phase of social distancing measures implemented from March to May 2020 [29]. In France, an analysis of the number of medical admissions to a psychiatric ward found that, compared with 2019, admissions were reduced by 18% in 2020, with this reduction increasing to 42% during the first lockdown. Similarly, the number of patients admitted to the emergency ward was reduced by 20% and 56%, respectively [30]. A study in Italy showed that 25% of their community mental health centers reduced their access hours [31]. Conversely, a Peruvian study found that community mental health centers were able to regain service capacity and fill the service gap created by the health crisis 9 months after the COVID-19 pandemic [32].

In Peru, telemedicine and in-person health care showed a similar trend compared with other countries in the region. In many Latin American countries, efforts to establish a telemedicine system produced positive results, as they reduced the overload of hospitals, decreased waiting times, and provided access to patients living in remote areas [33]. A study conducted in an emergency department in Argentina showed that telemedicine consultations peaked during the first lockdown period of the COVID-19 pandemic in March 2020. Subsequently, telemedicine consultations reduced progressively and were replaced by face-to-face interventions, but they remained at a stable value above the prepandemic trend [34]. It is worth noting that most of the teleconsultations were conducted without technical complications. Telemedicine in our context has been supported by a legal framework since 2017, and after the pandemic outbreak, it was included as part of several regulations of the Ministry of Health and the Peruvian College of Medicine, demonstrating the interest in the establishment and refinement of these technologies in health care [35]. The transition of mental health services to telemedicine in response to high patient demand has demonstrated that synchronous digital interventions facilitate the continuity of care [36]. These interventions not only reduce geographic limitations for patients and therapists but also eliminate the long hours of travel common in Lima [36].

Additionally, these interventions reduce stigma because they occur in a more private setting, can be delivered by trained health care providers (not only physicians), and could reduce costs [37]. A survey of psychologists in the United States showed an 85% increase in the use of telemedicine after the pandemic, and they projected that at least 35% of their work would involve telemedicine after the pandemic [38]. In addition, they mentioned that telemedicine use by psychologists was positively influenced by (1) being female; (2) having training in that field and treating patients with anxiety or partners' or women's complaints; and (3) working in a rural area, treating patients with antisocial personality disorder, or doing rehabilitation or psychometry work. Although telemedicine offers several advantages, it is crucial to consider the values and preferences of both patients and therapists to improve utilization rates and ease the demand for in-person care.

Our study revealed a decrease in mental health services during the latter months of 2022, potentially attributable to reduced utilization of teleconsultations within the evaluated health care system following the easing of COVID-19 pandemic restrictions. This could be attributed to the inconsistent implementation of mental health teleconsultations within the Peruvian context. Professionals often lacked dedicated equipment for these consultations, and there was an overwhelming demand for care, resulting in difficulties securing teleconsultation appointments [39]. Furthermore, patients with mental health issues expressed a preference for face-to-face care over teleconsultation care [40]. These factors likely contributed to the observed decrease in mental health teleconsultations following the resolution of the health emergency.

Public Health Implications

These results indicate that the decline in health care utilization during the pandemic underscores the fragility of the health care system. It was ill-prepared not only for managing a pandemic but also for delivering care in routine circumstances [41]. This calls for an improvement of the structure of the health care system, such as enhancing the surveillance system, training professionals for emergencies and disasters, strengthening the relationships between scientific and medical institutions, and fostering community engagement [42].

The use of telemedicine as a potential intervention to alleviate the strain on hospitals and health care centers has been a notable aspect of the pandemic response, underscoring the importance of implementing strategies for its adoption. This includes providing proper training to relevant stakeholders on these topics [43]. Future studies could explore the trends in health care utilization, particularly the significant uptake of telemedicine, which was virtually nonexistent before and during the peak of the pandemic but showed a progressive reduction (without disappearing entirely) toward the latter stages of analysis.

Furthermore, we encourage policy makers and health intelligence teams to utilize our findings as a foundation for developing health policies and regulations aimed at enhancing teleconsultation in mental health and health care overall. In Peru, the implementation of teleconsultation posed significant challenges. Many health professionals lacked the necessary equipment for conducting consultations, there was a lack of training in teleconsultation platform usage, and care centers themselves had limited access to the internet [39,44]. Moreover, telemedicine users persisted even after the pandemic, underscoring the ongoing demand for such services in resource-constrained settings such as Peru. Authors from countries sharing similar contexts could draw lessons from our positive telemedicine experience. We advocate for the implementation of these technologies within a framework that considers both barriers and facilitators [37]. Our findings can also be used to assess the postpandemic recovery of the health system.

Limitations and Strengths

This study highlights several strengths. First, it analyzes data from a leading national reference hospital in Lima, Peru, using a time series analysis to track utilization trends. We encompassed all patients receiving care over an extensive period, spanning most of the pandemic's duration. The sample size exceeded 1.5 million patients, and we also acquired additional data on mental health utilization, acknowledging its substantial burden in Peru. Advanced statistical methods were rigorously used in the methodology to analyze trends in health care utilization over these years. However, it is important to note that these results are confined to a national hospital in Lima and may not apply to other areas of the country or region. Further exploration of determinants affecting the variability in health care utilization, such as demographic, epidemiological, or clinical factors, is warranted. Therefore, we recommend further studies to assess whether care varied among specific age groups, genders, educational levels, or clinical variables, which may have hindered participants from accessing care.

Conclusions

In this time series analysis conducted at a national hospital in Peru during the pandemic years (March 2020 to December 2022), data from 1.5 million patients were analyzed. We observed a significant decrease in face-to-face health care utilization at the onset of the pandemic (March 2020) in both the outpatient clinic and mental health care services, indicating a notable impact on patient attendance. No significant trends were observed in both groups over the course of the pandemic. However, there was a noticeable increase in the number of users accessing mental health services each month. Overall, there was a trend of increased monthly utilization of the aforesaid services among all users during the pandemic period. Notably, telemedicine interventions were virtually nonexistent before the pandemic, but consultations during this period increased significantly, both overall and particularly among mental health users. At the conclusion of the study period, telemedicine services in mental health continued to attract users, indicating sustained demand for these interventions that were previously unavailable before the pandemic. Further follow-up evaluations could assess the long-term feasibility of this intervention as a means to enhance mental health access in the country. The utilization notably decreased in the subsequent months, as evidenced by the decline in the number of users accessing health care services at the hospital and in the number of teleconsultation users across all services. During times of emergency, the availability of easily accessible care services and the implementation of digital health services have proven to be of utmost importance. These results serve as a demonstration of the profound impact of the national COVID-19 lockdown on health care utilization in Peru. Governmental and key decision makers are encouraged to respond promptly to health needs and to draw lessons from these experiences, thus aiding the population in facing future health emergencies effectively.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Availability of Data and Materials

The data used or analyzed in this study are available from the corresponding author on reasonable request.

Authors' Contributions

DVZ was responsible for conceptualization, methodology, software, formal analysis, data curation, writing (review and editing), visualization, and supervision. JGS contributed to the conceptualization, methodology, formal analysis, and writing the original draft. PSB participated in conceptualization, formal analysis, and writing the original draft. NMP and ANF were involved in conceptualization and writing (review and editing). JHV handled conceptualization, investigation, data curation, writing (review and editing), and supervision. However, the author's involvement in the next steps of publication was limited. DVZ wrote the final version of the manuscript, restructured the analysis, and updated the search.

Conflicts of Interest

None declared.

References

1. Moynihan R, Sanders S, Michaleff ZA, Scott AM, Clark J, To EJ, et al. Impact of COVID-19 pandemic on utilisation of healthcare services: a systematic review. BMJ Open 2021 Mar 16;11(3):e045343 [FREE Full text] [doi: 10.1136/bmjopen-2020-045343] [Medline: 33727273]
2. Hangartner N, Di Gangi S, Elbl C, Senn O, Bisatz F, Fehr T. Impact of the COVID-19 pandemic on emergency outpatient consultations and admissions of non-COVID-19 patients (ECCO)--a cross-sectional study. PLoS One 2022;17(6):e0269724 [FREE Full text] [doi: 10.1371/journal.pone.0269724] [Medline: 35687575]
3. Sim B, Nam EW. The impact of COVID-19 pandemic on outpatient visits for all-cause and chronic diseases in Korea: a nationwide population-based study. Int J Environ Res Public Health 2022 May 06;19(9):5674 [FREE Full text] [doi: 10.3390/ijerph19095674] [Medline: 35565068]
4. Provenzano DA, Sitzman BT, Florentino SA, Buterbaugh GA. Clinical and economic strategies in outpatient medical care during the COVID-19 pandemic. Reg Anesth Pain Med 2020 Aug;45(8):579-585 [FREE Full text] [doi: 10.1136/rapm-2020-101640] [Medline: 32475827]
5. Malik MA. Fragility and challenges of health systems in pandemic: lessons from India's second wave of coronavirus disease 2019 (COVID-19). Glob Health J 2022 Mar;6(1):44-49 [FREE Full text] [doi: 10.1016/j.glohj.2022.01.006] [Medline: 35070474]
6. Bong C, Brasher C, Chikumba E, McDougall R, Mellin-Olsen J, Enright A. The COVID-19 pandemic: effects on low- and middle-income countries. Anesth Analg 2020 Jul;131(1):86-92 [FREE Full text] [doi: 10.1213/ANE.0000000000004846] [Medline: 32243287]
7. Di Gennaro F, Occa E, Ramirez L, Marotta C, Segala FV, Santana J, et al. Malaria, HIV and malnutrition among internally displaced people in Mozambique during COVID-19 pandemic: results from a community-based intervention. Ann Glob Health 2022;88(1):106 [FREE Full text] [doi: 10.5334/aogh.3969] [Medline: 36561539]

8. Foo CD, Shrestha P, Wang L, Du Q, García-Basteiro A, Abdullah AS, et al. Integrating tuberculosis and noncommunicable diseases care in low- and middle-income countries (LMICs): a systematic review. PLoS Med 2022 Jan;19(1):e1003899 [FREE Full text] [doi: 10.1371/journal.pmed.1003899] [Medline: 35041654]
9. Soto A. Barriers to effective care in the referral hospitals of Peru's Ministry of Health: serving patients in the 21st century with 20th century resources. Rev Peru Med Exp Salud Publica 2019;36(2):304-311. [doi: 10.17843/rpmesp.2019.362.4425] [Medline: 31460645]
10. Chafloque-Vásquez RA, Pampa-Espinoza L, Celis Salinas JC. Seroprevalencia de COVID-19 en trabajadores de un hospital de la Amazonía peruana. Acta Med Peru 2020 Sep 01;37(3):390-392. [doi: 10.35663/amp.2020.373.1050]
11. Herrera-Añazco P, Uyen-Cateriano A, Mezones-Holguin E, Taype-Rondan A, Mayta-Tristan P, Malaga G, et al. Some lessons that Peru did not learn before the second wave of COVID-19. Int J Health Plann Manage 2021 May;36(3):995-998 [FREE Full text] [doi: 10.1002/hpm.3135] [Medline: 33595137]
12. Pesantes MA, Lazo-Porras M, Cárdenas M, Diez-Canseco F, Tanaka-Zafra JH, Carrillo-Larco RM, et al. Healthcare challenges for people with diabetes during the national state of emergency due to COVID-19 in Lima, Peru: primary healthcare recommendations. Rev Peru Med Exp Salud Publica 2020 Dec 02;37(3):541-546. [doi: 10.17843/rpmesp.2020.373.5980] [Medline: 33295559]
13. Lindert J, Jakubauskiene M, Bilsen J. The COVID-19 disaster and mental health-assessing, responding and recovering. Eur J Public Health 2021 Nov 09;31(Supplement_4):iv31-iv35 [FREE Full text] [doi: 10.1093/eurpub/ckab153] [Medline: 34751367]
14. Alva-Diaz C, Huerta-Rosario A, Molina RA, Pacheco-Barrios K, Aguirre-Quispe W, Navarro-Flores A, et al. Mental and substance use disorders in Peru: a systematic analysis of the Global Burden of Disease study. J Public Health (Berl) 2020 Jun 19;30(3):629-638. [doi: 10.1007/s10389-020-01341-9]
15. Villarreal-Zegarra D, Reátegui-Rivera CM, Otazú-Alfaro S, Yantas-Alcantara G, Soto-Becerra P, Melendez-Torres GJ. Estimated impact of the COVID-19 pandemic on the prevalence and treatment of depressive symptoms in Peru: an interrupted time series analysis in 2014-2021. Soc Psychiatry Psychiatr Epidemiol 2023 Sep 08;58(9):1375-1385 [FREE Full text] [doi: 10.1007/s00127-023-02446-8] [Medline: 36884090]
16. Huarcaya-Victoria J, Barreto J, Aire L, Podestá A, Caqui M, Guija-Igreda R, et al. Mental health in COVID-2019 survivors from a general hospital in Peru: sociodemographic, clinical, and inflammatory variable associations. Int J Ment Health Addict 2023;21(2):1264-1285 [FREE Full text] [doi: 10.1007/s11469-021-00659-z] [Medline: 34602915]
17. Huarcaya-Victoria J, Alarcon-Ruiz CA, Barzola-Farfán W, Cruzalegui-Bazán C, Cabrejos-Espinoza M, Aspilcueta-Montoya G, et al. One-year follow-up of depression, anxiety, and quality of life of Peruvian patients who survived COVID-19. Qual Life Res 2023 Jan;32(1):139-149 [FREE Full text] [doi: 10.1007/s11136-022-03208-w] [Medline: 35939252]
18. Aragona M, Barbato A, Cavani A, Costanzo G, Mirisola C. Negative impacts of COVID-19 lockdown on mental health service access and follow-up adherence for immigrants and individuals in socio-economic difficulties. Public Health 2020 Sep;186:52-56 [FREE Full text] [doi: 10.1016/j.puhe.2020.06.055] [Medline: 32771661]
19. Sharma MS, De Maio M, Young K, Santopietro J. Transformation of outpatient psychiatry. Psychiatr Clin North Am 2022 Mar;45(1):57-69. [doi: 10.1016/j.psc.2021.11.003] [Medline: 35219442]
20. Giraldo EB. COVID-19 in Peru. Indian J Psychiatry 2020 Sep;62(Suppl 3):S498-S501 [FREE Full text] [doi: 10.4103/psychiatry.IndianJPsychiatry_1045_20] [Medline: 33227059]
21. Arsenault C, Gage A, Kim MK, Kapoor NR, Akweongo P, Amponsah F, et al. COVID-19 and resilience of healthcare systems in ten countries. Nat Med 2022 Jun;28(6):1314-1324 [FREE Full text] [doi: 10.1038/s41591-022-01750-1] [Medline: 35288697]
22. Mejia CR, Ruiz-Aquino M, Garcia-Arteta G, Ramos G, Astete-Pérez I, Pineda-Cáceres M, et al. Factors associated with the use of medicines and self-medication during the COVID-19 pandemic in 12 Latin American countries. Heliyon 2024 Jan 15;10(1):e22299 [FREE Full text] [doi: 10.1016/j.heliyon.2023.e22299] [Medline: 38173506]
23. Pujolar G, Oliver-Anglès A, Vargas I, Vázquez M. Changes in access to health services during the COVID-19 pandemic: a scoping review. Int J Environ Res Public Health 2022 Feb 03;19(3):1749 [FREE Full text] [doi: 10.3390/ijerph19031749] [Medline: 35162772]
24. Mental Health and Substance Use (MSD) Team, World Health Organization (WHO). The Impact of COVID-19 on Mental, Neurological and Substance Use Services. Geneva, Switzerland: World Health Organization; Oct 5, 2020:1-49.
25. Chen S, She R, Qin P, Kershenbaum A, Fernandez-Egea E, Nelder JR, et al. The medium-term impact of COVID-19 lockdown on referrals to secondary care mental health services: a controlled interrupted time series study. Front Psychiatry 2020;11:585915 [FREE Full text] [doi: 10.3389/fpsyt.2020.585915] [Medline: 33324258]
26. Boluarte-Carbajal A, Navarro-Flores A, Villarreal-Zegarra D. Explanatory model of perceived stress in the general population: a cross-sectional study in Peru during the COVID-19 context. Front Psychol 2021;12:673945. [doi: 10.3389/fpsyg.2021.673945] [Medline: 34248770]
27. Zhu JM, Myers R, McConnell KJ, Levander X, Lin SC. Trends in outpatient mental health services use before and during the COVID-19 pandemic. Health Aff (Millwood) 2022 Apr;41(4):573-580 [FREE Full text] [doi: 10.1377/hlthaff.2021.01297] [Medline: 35377763]

28. McBain RK, Cantor J, Pera MF, Breslau J, Bravata DM, Whaley CM. Mental health service utilization rates among commercially insured adults in the US during the first year of the COVID-19 pandemic. JAMA Health Forum 2023 Jan 06;4(1):e224936 [FREE Full text] [doi: 10.1001/jamahealthforum.2022.4936] [Medline: 36607697]
29. Bruckner TA, Huo S, Huynh M, Du S, Young A, Ro A. Psychiatric emergencies in Los Angeles county during, and after, initial COVID-19 societal restrictions: an interrupted time-series analysis. Community Ment Health J 2023 May;59(4):622-630 [FREE Full text] [doi: 10.1007/s10597-022-01043-4] [Medline: 36509936]
30. Perozziello A, Sousa D, Aubriot B, Dauriac-Le Masson V. Use of mental health services in the aftermath of COVID-19 waves: a retrospective study conducted in a French Psychiatric and Neurosciences University Hospital. BMJ Open 2023 Feb 23;13(2):e064305 [FREE Full text] [doi: 10.1136/bmjopen-2022-064305] [Medline: 36822810]
31. Carpiniello B, Tusconi M, Zanalda E, Di Sciascio G, Di Giannantonio M, Executive Committee of The Italian Society of Psychiatry. Psychiatry during the Covid-19 pandemic: a survey on mental health departments in Italy. BMC Psychiatry 2020 Dec 16;20(1):593 [FREE Full text] [doi: 10.1186/s12888-020-02997-z] [Medline: 33327940]
32. Villarreal-Zegarra D, Segovia-Bacilio P, Paredes-Angeles R, Vilela-Estrada AL, Cavero V, Diez-Canseco F. Provision of community mental health care before and during the COVID-19 pandemic: a time series analysis in Peru. Int J Soc Psychiatry 2023 Dec;69(8):1996-2006 [FREE Full text] [doi: 10.1177/00207640231185026] [Medline: 37449754]
33. Camacho-Leon G, Faytong-Haro M, Carrera K, Molero M, Melean F, Reyes Y, et al. A narrative review of telemedicine in Latin America during the COVID-19 pandemic. Healthcare (Basel) 2022 Jul 22;10(8):1361 [FREE Full text] [doi: 10.3390/healthcare10081361] [Medline: 35893183]
34. Grande Ratti MF, Frid SA, Rodriguez P, Pollan JA, Rovegno L, Esteban JA, et al. Impact of COVID-19 pandemic on telehealth and in-person visits: implications from an emergency department in Argentina. Stud Health Technol Inform 2022 Jun 06;290:377-379. [doi: 10.3233/SHTI220100] [Medline: 35673039]
35. Curioso WH, Galán-Rodas E. El rol de la telesalud en la lucha contra el COVID-19 y la evolución del marco normativo peruano. Acta Med Peru 2020 Aug 18;37(3):366-375. [doi: 10.35663/amp.2020.373.1004]
36. Li H, Glecia A, Kent-Wilkinson A, Leidl D, Kleib M, Risling T. Transition of mental health service delivery to telepsychiatry in response to COVID-19: a literature review. Psychiatr Q 2022 Mar;93(1):181-197 [FREE Full text] [doi: 10.1007/s11126-021-09926-7] [Medline: 34101075]
37. Villarreal-Zegarra D, Alarcon-Ruiz CA, Melendez-Torres GJ, Torres-Puente R, Navarro-Flores A, Cavero V, et al. Development of a framework for the implementation of synchronous digital mental health: realist synthesis of systematic reviews. JMIR Ment Health 2022 Mar 29;9(3):e34760 [FREE Full text] [doi: 10.2196/34760] [Medline: 35348469]
38. Pierce BS, Perrin PB, Tyler CM, McKee GB, Watson JD. The COVID-19 telepsychology revolution: a national study of pandemic-based changes in U.S. mental health care delivery. Am Psychol 2021 Jan;76(1):14-25. [doi: 10.1037/amp0000722] [Medline: 32816503]
39. Paredes-Angeles R, Cavero V, Vilela-Estrada A, Cusihuaman-Lope N, Villarreal-Zegarra D, Diez-Canseco F. Telehealth in community mental health centers during the COVID-19 pandemic in Peru: a qualitative study with key stakeholders. SSM - Mental Health 2024 Jun;5:100287 [FREE Full text] [doi: 10.1016/j.ssmmh.2023.100287]
40. Cusihuaman-Lope N, Vilela-Estrada AL, Cavero V, Villarreal-Zegarra D, Diez-Canseco F. Experiences of users and family members on the care received at community mental health centers in Lima and Callao during the COVID-19 pandemic. Rev Peru Med Exp Salud Publica 2023;40(3):278-286. [doi: 10.17843/rpmesp.2023.403.12717] [Medline: 37991031]
41. Vargas A, Cubas Benavides F, León-Jiménez F, Cubas S. La nueva pandemia y la verdadera crisis del sistema de salud en el Perú. Rev Med Hered 2021 Jul 19;32(2):132-133. [doi: 10.20453/rmh.v32i2.3988]
42. Castro Baca M, Villena Pacheco A. La pandemia del COVID-19 y su repercusión en la salud pública en Perú. Acta Med Peru 2021 Oct 24;38(3):161-162. [doi: 10.35663/amp.2021.383.2227]
43. Appleton R, Barnett P, Vera San Juan N, Tuudah E, Lyons N, Parker J, et al. Implementation strategies for telemental health: a systematic review. BMC Health Serv Res 2023 Jan 25;23(1):78 [FREE Full text] [doi: 10.1186/s12913-022-08993-1] [Medline: 36694164]
44. Mayo-Puchoc N, Bejarano-Carranza J, Paredes-Angeles R, Vilela-Estrada AL, García-Serna J, Cusihuaman-Lope N, et al. Paper promises: Peruvian frontline health workers' perspectives on mental health policies during COVID-19. Health Policy Plan 2023 Nov 16;38(Supplement_2):ii3-ii13 [FREE Full text] [doi: 10.1093/heapol/czad055] [Medline: 37995267]

Abbreviations

HNGAI: Hospital Nacional Guillermo Almenara Irigoyen

Edited by J Torous; submitted 25.10.23; peer-reviewed by A AL-Asadi, TAR Sure; comments to author 30.01.24; revised version received 10.02.24; accepted 12.02.24; published 08.07.24.

Please cite as:
Villarreal-Zegarra D, García-Serna J, Segovia-Bacilio P, Mayo-Puchoc N, Navarro-Flores A, Huarcaya-Victoria J
In-Person and Teleconsultation Services at a National Hospital in Peru: Time Series Analysis of General and Psychiatric Care Amid the COVID-19 Pandemic
JMIR Ment Health 2024;11:e53980
URL: https://mental.jmir.org/2024/1/e53980
doi:10.2196/53980
PMID:38976320

Original Paper

Empathy Toward Artificial Intelligence Versus Human Experiences and the Role of Transparency in Mental Health and Social Support Chatbot Design: Comparative Study

Jocelyn Shen[1], MS; Daniella DiPaola[1], MS; Safinah Ali[1], MS; Maarten Sap[2], PhD; Hae Won Park[1], PhD; Cynthia Breazeal[1], PhD

[1]MIT Media Lab, Cambridge, MA, United States

[2]Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, United States

Corresponding Author:
Jocelyn Shen, MS
MIT Media Lab
75 Amherst Street
Cambridge, MA, 02139
United States
Phone: 1 3109802254
Email: joceshen@mit.edu

Abstract

Background: Empathy is a driving force in our connection to others, our mental well-being, and resilience to challenges. With the rise of generative artificial intelligence (AI) systems, mental health chatbots, and AI social support companions, it is important to understand how empathy unfolds toward stories from human versus AI narrators and how transparency plays a role in user emotions.

Objective: We aim to understand how empathy shifts across human-written versus AI-written stories, and how these findings inform ethical implications and human-centered design of using mental health chatbots as objects of empathy.

Methods: We conducted crowd-sourced studies with 985 participants who each wrote a personal story and then rated empathy toward 2 retrieved stories, where one was written by a language model, and another was written by a human. Our studies varied disclosing whether a story was written by a human or an AI system to see how transparent author information affects empathy toward the narrator. We conducted mixed methods analyses: through statistical tests, we compared user's self-reported state empathy toward the stories across different conditions. In addition, we qualitatively coded open-ended feedback about reactions to the stories to understand how and why transparency affects empathy toward human versus AI storytellers.

Results: We found that participants significantly empathized with human-written over AI-written stories in almost all conditions, regardless of whether they are aware (t_{196}=7.07, *P*<.001, Cohen *d*=0.60) or not aware (t_{298}=3.46, *P*<.001, Cohen *d*=0.24) that an AI system wrote the story. We also found that participants reported greater willingness to empathize with AI-written stories when there was transparency about the story author (t_{494}=–5.49, *P*<.001, Cohen *d*=0.36).

Conclusions: Our work sheds light on how empathy toward AI or human narrators is tied to the way the text is presented, thus informing ethical considerations of empathetic artificial social support or mental health chatbots.

(JMIR Ment Health 2024;11:e62679) doi:10.2196/62679

KEYWORDS

empathy; large language models; ethics; transparency; crowdsourcing; human-computer interaction

Introduction

Empathy, the sharing of emotions with a social other, is foundational in developing strong interpersonal ties [1,2] and mental well-being [3]. With the rise of large language models (LLMs) and increase in chatbots for social companionship [4] and mental health [5,6], it is crucial to understand how empathy toward artificially intelligent agents manifests and what the social implications of this phenomenon are. In particular, commercial chatbots often display anthropomorphism by adopting their own identities or experiences. Current artificial intelligence (AI) systems hold the ability to express social and emotional influences through the mechanisms of empathy, which can lead to downstream impacts in the real world. For example, in AI service applications, increased empathy improves service acceptance and user compliance [7,8]. Such empathetic relationships with synthetic agents can unfold in the following ways: (1) in the behavior of the agent, where the agent behaves in an empathic way toward other agents and toward the user, and (2) in the relation the agent establishes with the user, where the agent looks like and acts in a way that leads the user to establish an empathic relation with it [9,10]. Prior works indicate that acceptance and trust toward AI devices is directly related to how much people empathize with an AI system [9]. Further works validate that the more personal information AI agents disclose, the more empathetic human conversation partners are toward these agents [11]. As machines are increasingly capable of telling human-like stories in daily life, this raises important questions about how people might empathize with AI-written stories in real-world, user-facing contexts, comparing them with how people perceive stories created by AI, and how transparency of the system modulates these psychological responses [12,13].

In this work, we study how much people empathize with stories created by AI compared with stories created by other humans as well as how author disclosure affects perceived empathy. Humans can breathe life into inanimate or artificial systems [9,14-16] and are able to relate to fictional experiences when they are human-like or realistic in the scope of one's own life [17,18]. As such, this work calls for ethical and philosophical concerns about differences in empathy toward humans and AI—machines have no lived experiences yet can produce stories as their "own" [19,20]. If the machine uses these fabricated experiences to elicit a particular behavior from the user, is this considered manipulation? How are behaviors shifted if the user is aware that the experiences are fabricated? How might empathy toward AI agents, such as social companion or mental health chatbots, impact downstream human outcomes?

Because outputs from generative AI are not an artificial agent's actual experiences [20], but rather a probabilistically sampled sequence of text from human experiences, it is important to be precise and nuanced when communicating results from generated text to ensure ethical deployment of such systems in the mental health domain [21-23]. In the field of social psychology, researchers have explored how nudges, small subtle changes that can inspire big changes in actions, can modulate empathy [24]. Yet, in the AI domain, few works have explored how subtle design changes in the presentation of AI-written stories can significantly shift attitudes and empathy toward AI systems [25].

Prior works generally indicate that perceptions of AI can change depending on transparency. Most works find that knowledge of AI involvement reduces the perception of the agent or quality of interaction and that there are fundamental qualities of "humanness" in texts written by people [25-27], but that fostering trust and acceptance can lead to more empathy toward an AI agent [9]. Grounded by these works, we hypothesize that empathy toward AI-written stories, both generated and retrieved in response to a user's own personal story, will be significantly lower than empathy toward human-written stories whether the author is disclosed [H1]. We also hypothesize that people will be more willing to empathize with AI stories when the author of the story is made transparent, as the output could be perceived as more trustworthy [H2]. To test our hypotheses, in this work, we investigate the following:

1. How does empathy change when stories, human- or AI-written, are retrieved versus generated directly by a language model?
2. How does transparency about the author of a story play a role in empathy toward human versus AI narrators?

In summary, we aim to answer the following research questions:

- How does empathy toward human- versus AI-written stories differ?
- What qualities of human- versus AI-written stories affect people's empathetic responses?
- How do the aforementioned results change when the narrator of a story is made transparent to users?
- What are the ethical implications of empathy toward AI stories in social support and mental health chatbots, and how are these implications influenced by transparency of the story's author?

Methods

Study Procedure

We conducted 4 crowd-sourced studies with a total of 985 participants to assess the effects of author origin on empathy. Within each session, participants wrote their own personal stories and rated empathy toward stories written by people or by ChatGPT.

The retrieved stories were matched based on similarity of the embeddings of stories, and generated stories were generated on the fly, given the user's story as a prompt. We used ChatGPT to generate a set of stories using seed stories from the EmpathicStories data set [28]. Stories generated by ChatGPT (gpt-3.5-turbo) were prompted with a context story and the following instruction: Write a story from your own life that the narrator would empathize with. Do not refer to the narrator explicitly.

The study's 4 comparisons are as follows:

- H-CR: we compared empathy toward the narrator across *human-written retrieved* stories and *ChatGPT-retrieved* stories
- H-CR+T: we compared empathy toward the narrator across *human-written retrieved* stories and *ChatGPT-retrieved* stories, making transparent to the user whether the story they read was written by a human or AI before they rated their empathy (repeat H-CR with transparency)
- H-CG: we compared empathy toward the narrator across *human-written retrieved* stories and *ChatGPT-generated* stories (in response to the user's story as a context).
- H-CG+T: we compared empathy toward the narrator across *human-written retrieved* stories and *ChatGPT-generated* stories, making transparent to the user whether the story they read was written by a human or AI before they rated their empathy (repeat H-CG with transparency).

Finally, in all studies, participants reported how their empathy toward the stories would change if the stories were written by AI. We define human-written retrieved stories as personal stories that are written by people and retrieved using a model that takes the user's self-written story as a query. ChatGPT-retrieved stories come from our pregenerated set of 1568 stories and are retrieved in the same manner as human-written retrieved stories. Finally, ChatGPT-generated stories are stories that are generated on the fly, given the user's story as input. Examples of stories across conditions are shown in Textbox 1.

Textbox 1. Examples of a user story (written by a participant) and the corresponding retrieved human-written story, pregenerated ChatGPT story retrieved based on the user's story, and on-the-fly generated ChatGPT story.

User story

"I started to experience a nagging pain in my right upper abdomen...The GP sent me for an ultrasound which detected hydronephrosis and referred me to a consultant...Following more scans I was found to have multiple spinal cysts in the whole spine...I felt so grateful and validated that someone had finally helped me...I am starting to finally feel at peace that I have a reason and I am not crazy."

Human-written retrieved story

"A week ago, I was having pain in my back. I thought it was just normal wear and tear until the pain got worse...When I was waiting for the doctor, a nurse came in the room...She said that I had kidney stones...I was so relieved after that. I felt like I had a new lease on life, and decided to commit myself to eating healthier and drinking much more water."

ChatGPT-written retrieved story

"I remember the first time I felt a strange pain in my chest...At first, I thought it might be a heart attack, but I was young and healthy,...Finally, after months of searching, I found a specialist who was able to give me some answers...It turned out that I had a rare condition...It was a relief to finally have a diagnosis, but it was also a shock to realize that my instincts had been right all along."

ChatGPT-written story

"I remember the first time I felt a strange pain in my chest...At first, I thought it might be a heart attack, but I was young and healthy,...Finally, after months of searching, I found a specialist who was able to give me some answers...It turned out that I had a rare condition...It was a relief to finally have a diagnosis, but it was also a shock to realize that my instincts had been right all along."

Ethical Considerations

Our study was approved by the Massachusetts Institute of Technology's ethics board as an exempt protocol (ID E-4618).

System and Interaction Design

Story Prompts and Retrieved Stories

To prompt vulnerable and meaningful personal stories from users, we used questions from the Life Story Interview, an approach from social science that gathers key moments from a person's life [29] (eg, "Look back over your life and tell us and emotional moment or experience you had in the past..."). In order to ensure that topics were constrained to stories present in our retrieval database, we used topic modeling to identify key clusters in the personal narratives from EmpathicStories. To identify these topics, we used Latent Dirichlet Allocation and KeyBERT on the clusters [30]. Users were instructed to reflect on their life in relation to one of the chosen topics (eg, family and mental health). Personal stories that participants wrote were restricted to be between 500 and 10,000 characters. See Multimedia Appendix 1 for full participant instructions. Stories retrieved by our model were either pulled from the EmpathicStories data set (1568 stories) or generated by ChatGPT (gpt-3.5-turbo). These stories covered a diverse set of personal experiences including mental health, life changes, loneliness, depression, substance abuse, and trauma.

Story Retrieval Model

Because our study aimed to assess differences in empathy toward human- versus AI-written stories, both the user's experiences and the stories returned by our system were important. Returning a story at random could undermine the user's experiences and hinder their empathy toward the retrieved story. Although many methods exist to retrieve semantically similar pieces of text [31], few focus on retrieving stories that users would emotionally resonate with, given their own story context. As such, we used a fine-tuned BART-base model from Shen et al [28], which was trained on the EmpathicStories data set, a corpus containing pairs of stories each annotated with an "empathic similarity" score from 1 to 4, where empathic similarity refers to how likely the narrators of both stories would empathize with one another. Using this model, we improved retrieval of stories that are empathetically relevant to a user's own personal story.

User Study Interface

We deployed a web interface similar to a guided journaling application where users write and read personal stories. The interface connects to a server run on a graphics processing unit machine (4x Nvidia A40s, 256 GB of RAM, and 64 cores; Nvidia), which retrieves story responses in real time. In addition, the server connects the front end to Firebase Realtime storage in order to track interaction data throughout the course of the study.

Participants and Recruitment

We recruited a pool of 985 participants from Prolific, a crowdsourcing platform that connects researchers with survey respondents for high-quality data collection. Participants across the studies were predominantly female and White. All participants on average had high trait empathy and neutral arousal and valence before starting the study. Full demographic distributions across the 4 studies are shown in Table 1. All participants were paid US $12 per hour for their time.

XSL•FO
RenderX

Table 1. Participant demographic distribution.

Demographic	H-CG	H-CR	H-CR+T	H-CG+T
Number of participants	300	299	197	189
Age (y), mean (SD), range	37.60 (12.54), 18-75	40.18 (14.31), 18-79	40.16 (13.76), 19-77	38.82 (13.52), 18-79
Gender, n (%)				
Women	173 (58)	161 (54)	100 (51)	111 (58)
Men	120 (40)	132 (44)	93 (47)	76 (40)
Nonbinary	5 (1)	3 (1)	2 (1)	1 (1)
N/A[a]	2 (1)	3 (1)	2 (1)	1 (1)
Ethnicity, n (%)				
White	228 (76)	242 (81)	160 (81)	145 (77)
Black	24 (8)	16 (5)	20 (10)	13 (7)
Asian	14 (5)	15 (5)	6 (3)	9 (5)
Other	13 (4)	7 (2)	4 (2)	5 (2)
Indian	10 (3)	7 (2)	4 (2)	13 (7)
NA	5 (2)	8 (3)	3 (2)	1 (0.05)
Hispanic	4 (1)	2 (1)	—[b]	1 (0.05)
Middle Eastern	1 (0.05)	1 (0.05)	—	2 (1)
Native	1 (0.05)	—	—	—
Islander	—	1 (0.05)	—	—
Empathy level, mean (SD), range	4.26 (0.83), 1-5	4.18 (0.79), 1-5	4.31 (0.79), 1-5	4.24 (0.69), 2-5
Arousal, mean (SD), range	4.42 (1.84), 1-9	4.80 (1.78), 1-9	4.81 (1.94), 1-9	4.48 (1.94), 1-9
Valence, mean (SD), range	5.75 (1.68), 1-9	5.76 (1.70), 1-9	5.75 (1.86), 1-9	5.76 (1.58), 1-9

[a]N/A: not applicable.

[b]—: not available.

Data Collection and Analysis

At the beginning of the study, we measured the user's valence and arousal using a self-report 9-point Likert scale, as current emotional state could influence empathy. Participants on average had neutral arousal and valence, so we did not exclude any participants from the study. For our empathy measurement, we used a shortened version of the State Empathy Scale [32], which contains 7 questions covering affective (sharing of others' feelings), cognitive (adopting another's point of view), and associative (identification with others) aspects of situational empathy. Users additionally provided free-text responses about their empathy toward the story as well as multiple-choice questions listing reasons why they did or did not empathize with the story (ie, the feelings the narrator had, the authenticity of the story, or providing free response reasons). At the end of the study, users self-reported how their empathy would change if the stories they read in the session were written by AI (which we term as perceived empathy with AI). Finally, at the end of the study, we collected demographic information including trait empathy, age, gender, and ethnicity. All survey questions were mandatory for participants (with "prefer not to answer" options for sensitive demographic information). Experiments were conducted throughout the month of June 2023. A full flow of the user study procedure is shown in Figure 1.

We used both quantitative and qualitative approaches to understand the effects of empathy toward a story from human versus AI narrators and offer insights around why empathy shifts under certain conditions. To analyze differences in empathy with the State Empathy Scale, we used a 2-tailed paired *t* test, as we identified through a Shapiro-Wilke test that the data are normally distributed. Note that we computed total empathy toward a story using the mean of the State Empathy Scale survey questions, and we present descriptive analyses of survey responses across conditions as mean (SD). To compare perceived empathy across studies, we used an independent, 2-tailed *t* test. All quantitative analyses were conducted via statistical libraries written in Python (SciPy Statistics). For qualitative analysis, open-ended explanations for the empathy rating were thematically coded using an inductive approach [33]. Two researchers (JS and DP) independently coded a subset of the data using Excel spreadsheets (Microsoft Corp) and reached substantial agreement with a Cohen κ value of 0.70.

Figure 1. Flow of the user study procedure. Participants recognize their own emotions, write their personal stories, reflect on the stories, and then fill out demographic and self-perception of state empathy information in a survey. Then, they read and perceive given stories, compare and rank the stories, and rate how willing they are to empathize with artificial intelligence (AI) in general.

Results

Effects of Human Versus AI Authorship on Empathy Toward Stories

Participants Generally Felt More Empathy for Human-Written Stories Than AI-Written Stories

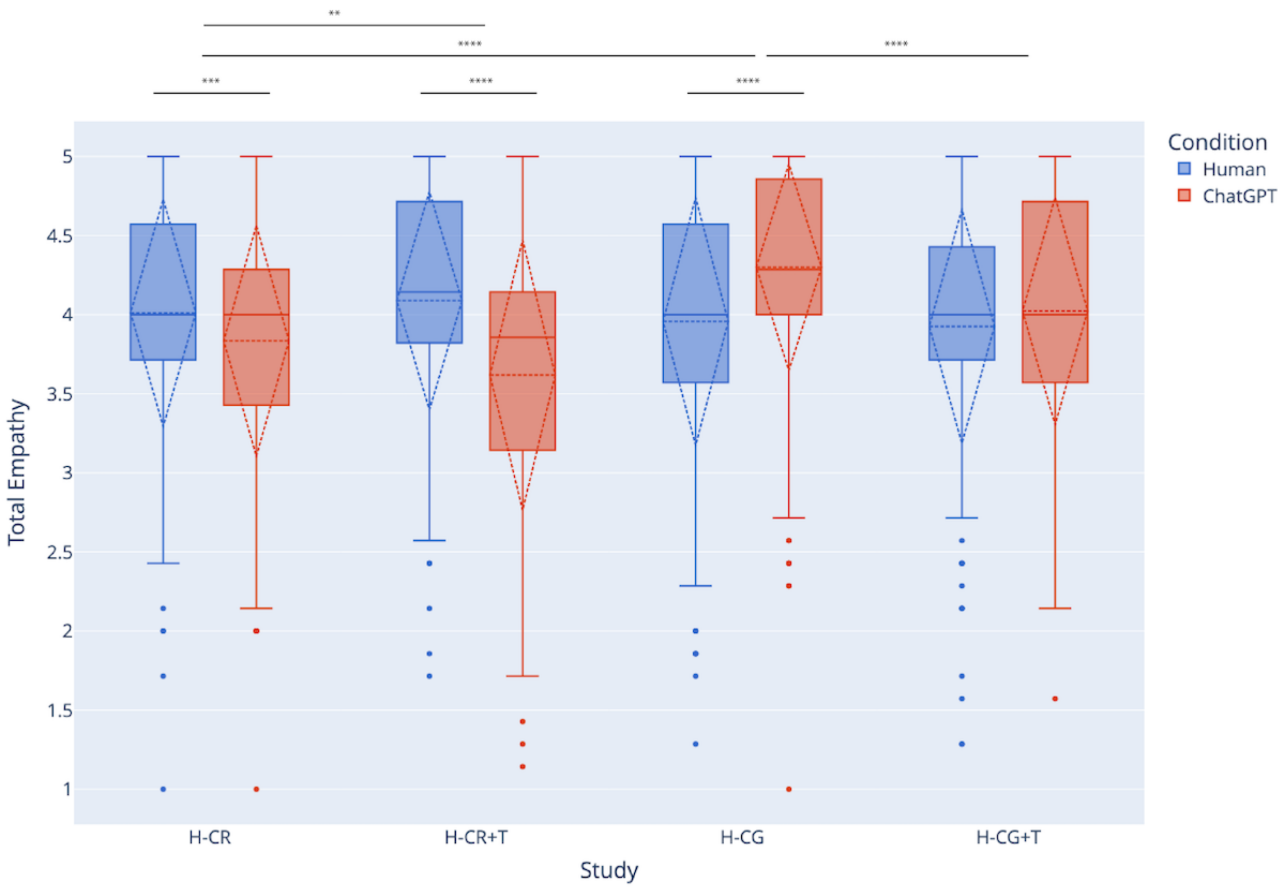
When we instead retrieved stories from a corpus of narratives generated by ChatGPT (H-CR), total empathy decreased across AI-written (mean 3.83, SD 0.73) versus human-written (mean 4.01, SD 0.72) stories (t_{298}=3.46, *P*<.001, Cohen *d*=0.24). This indicates a noticeable difference between human- versus. AI-written stories, which we explore further through qualitative analysis in later sections. When we made transparent to the user the author of the retrieved story (H-CR+T), we saw an even greater decrease in total empathy toward AI-written (mean 3.61, SD 0.85) stories relative to human-written (mean 4.1, SD 0.69) stories (t_{196}=7.07, *P*<.001, Cohen *d*=0.60).

When comparing *human-retrieved* (mean 3.96, SD 0.78) stories and *ChatGPT-generated* stories (mean 4.30, SD 0.65) based on the user's original story (H-CG), we found that participants empathized significantly more with ChatGPT-generated stories than human-retrieved stories (t_{299}=6.14, *P*<.001, Cohen *d*=0.47). Following this trend, we found that there was no statistically significant difference between empathy toward *human-retrieved* (mean 3.93, SD 0.73) and *ChatGPT-generated* (mean 4.02, SD 0.71) stories when the author was made transparent (H-CG+T).

Generated Stories Elicit More Empathy Than Retrieved Stories

Next, we cross-compared total empathy toward AI-written stories in H-CG (mean 4.3, SD 0.65) and H-CR (mean 3.83, SD 0.73), allowing us to explore differences in ChatGPT responding directly to a user's personal story context as compared with retrieving a relevant AI-generated story. From Figure 2, we see that empathy statistically significantly decreases in H-CR when stories are retrieved instead of generated directly from the user's written story (t_{597}=8.20, *P*<.001, Cohen *d*=0.67).

Figure 2. Changes in total empathy toward stories participants read across conditions (human-written vs AI-written story) and studies (author made transparent vs author not transparent, AI story was retrieved vs generated). **P*<.05, ***P*<.01, ****P*<.001, *****P*<.000.



Disclosure of Story Author Reduces Empathy in ChatGPT-Generated Stories

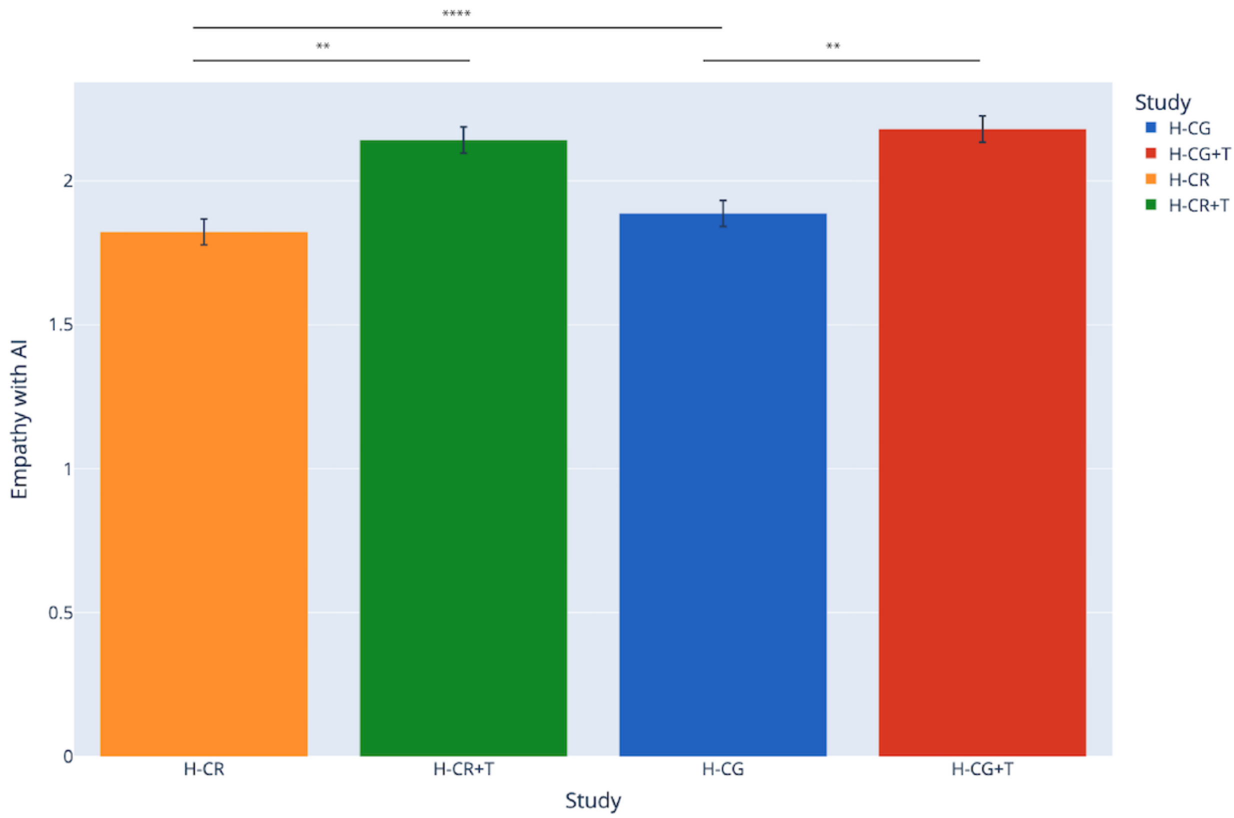
We cross-compared total empathy toward AI-written stories in H-CR and H-CR+T (mean 3.62, SD 0.86), allowing us to assess how transparency about a story being written by ChatGPT shifts empathy. We found that empathy toward the AI-written stories statistically significantly decreased when users were told before reading that the story was written by ChatGPT (t_{494}=3.02, *P*<.001, Cohen *d*=0.27), as shown in Figure 2.

Finally, we cross-compared total empathy toward AI-written stories in H-CG and H-CG+T (mean 4.02, SD 0.72) and saw that empathy in H-CG+T statistically significant decreased (t_{487}=4.37, *P*<.001, Cohen *d*=0.40). This confirmed the aforementioned result that telling participants a story was written by AI decreases empathy (Figure 2).

Effects of Author Disclosure on Willingness to Empathize With AI: People Are More Willing to Empathize With AI-Written Stories if Author Is Transparent

In addition to raw, self-reported empathy toward the narrator of each story, we also asked participants to rate how much they believe their empathy would shift if the stories they read were all written by AI, where scores are from Likert 1 (empathize a lot less) to 4 (empathize a lot more). As shown in Figure 3, we find that across all 4 studies, participants would, on average, empathize less (scores are generally at or below 2) with AI-written stories using our survey measurements (H-CG: mean 1.88, SD 0.91; H-CR: mean 1.82, SD 0.90; H-CR+T: mean 2.14, SD 0.89; H-CG+T: mean 2.18, SD 0.87). However, interestingly, we see that willingness to empathize with AI-written stories statistically significantly increases when we are transparent about the story being written by ChatGPT (ie, participants read a story knowing it was generated by ChatGPT). These results are shown in cross-comparing H-CR and H-CR+T for retrieved ChatGPT stories (t_{494}=–5.49, *P*<.001, Cohen *d*=0.36) as well as cross-comparing H-CG and H-CG+T for directly generated ChatGPT stories (t_{494}=–4.99, *P*<.001, Cohen *d*=0.33).

Figure 3. Self-reported willingness to empathize with AI-written stories across all 4 studies. AI: artificial intelligence. *P<.05, **P<.01, ***P<.001, ****P<.000.



Understanding Mechanisms Behind Empathy Toward Human Versus AI Stories

Through qualitative coding of participant free responses, conducted by 2 independent coders, we revealed 9 unique themes around why participants did or did not empathize with the stories (Table 2). Participants explained their reasoning by commenting on the narrator's perspective, including empathizing with the *situation* in the story or the *emotions* the narrator describes. Some participants expressed that the story was *not relatable* enough for them to empathize with. Two themes appeared around the way that the story was written: some expressed *story confusion* due to some of the logic of the story not being clear, and there was also a common theme around the *word choice* of the narrator, such as the writing style or phrasing. There were some participants who *mentioned AI* explicitly, and others who talked about the *authenticity* of the story, or whether it was real or fake. Some participants spoke about their *personality* being a factor in whether they were able to empathize. The *other* category was used if a response did not fit into an existing category.

We assigned a theme (or themes) to each response, and a percentage was calculated to account for the number of participants in each study. As a whole, the *emotional* (30.38%) and *situational* (24.74%) codes showed up most frequently across all conditions. One notable difference is that participants in H-CG (35.38%) and H-CR (31.99%) had a higher percentage of *emotional* codes than H-CR+*t* (23.40%) and H-CG+*t* (27.29%). H-CR+T and H-CG+T had a higher percentage of *word choice*, *authenticity*, and *mention AI* codes than H-CG and H-CR.

We broke themes down into individual studies and conditions. Conditions H-CR and H-CR+T were compared, as they compared the same types of stories (human-retrieved vs AI-retrieved), with H-CR+T explicitly telling participants when the stories were AI generated. Interestingly, codes for *emotional* were less common in the H-CR+T condition. H-CG and H-CG+T were compared (human-retrieved vs AI-generated) and showed a similar decrease in *emotional* codes.

Table 2. Themes resulting from a qualitative analysis across all four studies[a].

Code	Definition	Example quote	Total (%)	H-CG (%)	H-CR (%)	H-CG+T (%)	H-CR+T (%)
Emotional	Empathize with the emotions that the narrator describes in the story	"I can recognize the fact that someone would feel conflicted in this scenario. Simultaneously happy and envious."	30.38[b]	35.38[b,c]	31.99[b]	23.40[b]	27.29[b]
Situational	Empathize with the situation or context that the narrator is in	"The theme of sudden and unexpected loss was similar to my experience. I have lost a parent too which makes it easy to empathize. My parent had a long illness but the end was very sudden, similar to the narrators experience."	24.74	26.88[c]	25.94	23.19	21.18
Story confusion	Mention of specific details in the story that are not clear, including details or logic that does not add up	"I think her level of self doubt was a bit over the top for a position his finances did not depend on."	11.41	8.91	12.97[c]	11.49	12.88
Not relatable	Mention of not empathizing because the story was not relatable or they did not agree with the narrator	"I don't want to pursue the same goals as the narrator, so our feelings are different."	11.20	9.33	14.41[c]	10.43	10.04
Word choice	Mention of the writing style, phrasing, or grammar, typically to reduce feelings of empathy	"It was very poorly written and organized - just a bunch of ideas on the page, rather forced and disjointed. The style did not encourage me to believe it."	7.39	6.55	5.19	9.57	9.83[c]
Authenticity	Mention of the story being "real" or "fake," any mention of believability or originality	"It seems genuine, although I am aware that it may not be, as it is a common experience."	6.67	5.15	4.03	10.85[c]	8.73
Mentions AI	Explicitly mentions AI or automation	"It feels like an AI generated response based on keywords from my experience."	4.15	3.06	1.15	7.87[c]	6.55
Personality	Mention of personal ability to empathize	"I am a person who hardly empathizes and more so the story did not feel consistent to me."	0.47	0.56	0.29	0.43	0.66[c]
Other	Does not fit into any category, restates the question or generic	"I did empathize with the story."	3.59	4.18[c]	4.03	2.77	2.84

[a]Percentages are shown as the number of times a code was mentioned out of the total number of participants within each study.

[b]The top code in each column.

[c]The top percent in each row.

Discussion

Principal Results

From our work, we show that it is important to be intentional in how one presents outputs from generative AI systems. In summary, we find that empathy is higher for ChatGPT-generated stories than ChatGPT-retrieved stories; total empathy toward the story is generally higher for stories written by humans than AI, but that transparency creates greater willingness to empathize with AI.

Generated Versus Retrieved Stories

First, through cross-comparisons between ChatGPT-written retrieved stories (H-CR) and ChatGPT-generated stories (H-CG), we find that empathy is higher for ChatGPT-generated stories rather than ChatGPT-retrieved stories. Interestingly, we find that empathy is higher toward ChatGPT-generated stories than human-written retrieved stories. Thus, we did not validate that humans would empathize more with human-written stories in all conditions [H1]. These results on generated versus retrieved stories highlight the importance of context awareness. Generated stories directly respond to the user's story, and previous literature shows that a direct response to one's story increases empathy [34]. Output that is generated from conditioning on the stories can take much more from the input story, thus probably reaching a higher level of similarity, beyond what our retrieval algorithm is based on [35]. These results further suggest that trust in the author of the story, rather than the author's identity, may play a role in empathy toward personal stories.

Transparent Versus Opaque Story Author

In studies H-CR and H-CR+T, we find that people significantly empathize less with retrieved AI-written stories than human-written stories, which is in line with and supports previous research findings [26,27]. We find that empathy decreases most between human-written and AI-retrieved stories in H-CR+T when we are transparent about the author of the story. This indicates that knowing when a story is written by AI alters our empathy toward that story and ability to relate to the narrator, possibly because AI is conveying experiences that are not its "own."

Interestingly, participants' willingness to empathize with AI systems significantly increases across both retrieval and generation conditions when the author of the story is made transparent (validating [H2]). Prior works indicate that transparency about AI's lack of human qualities can reduce perceived similarity [27], but that transparency can increase trust toward AI systems [36]. Our results may indicate that disclosing a story's author could increase willingness to empathize through trust, or through demonstration that AI stories contain relatable qualities.

This finding also has implications for the design of empathetic AI systems outside of therapy. For instance, for social companions, or AI health care providers, where empathy may be an essential social interaction, designers must consider transparency in the AI system to increase users' willingness to empathize through trust. In practical tool design, this may take the form of disclosing the content creator in the user interactions, such as in the visual user interface or voices used [37]. There is much argument for using more anthropomorphic representations of AI technologies such as robots to increase trust [38], but in our empathetic storytelling interactions, where there is actual uncertainty about the story author, results may indicate otherwise.

In the H-CR+T condition, participants' reasoning for not empathizing with AI-written stories was more centered around themes relating to how the story was written, including "story confusion" and "word choice," similar to research that showed that "linguistic style" was a reported indicator for AI-generated text [39]. For example, one participant stated, "The story and feelings described feel really fake and over the top. It does not feel genuine and has clearly been written by a robot."

Others mention not being able to empathize with the story because the story did not actually happen, but they are still capable of engaging with it as a made-up story. For example, one participant shared, "Because I know it's written by AI then I can't think that it is genuine. However, as a work of fiction I can immerse myself in it and connect with the characters portrayed." This sentiment opens up the potential for AI-written stories to be contextualized for the user in a way that does not feel like they are being deceived by a fake story.

We see no difference in empathy between retrieved human stories and ChatGPT stories generated in direct response to the user (H-CG+T), indicating that responding directly to a user's story might overshadow the underlying empathic benefits of human-written stories. In this condition, more participants mentioned the "authenticity" of the story or mentioned AI explicitly as a factor against empathizing with the story they read. Participants tended to focus more on the author of the story instead of the content of the story in their open-ended responses. One participant shared, "The story felt similar to the content of my story, which made me feel like I could empathize with it. But knowing the story was written by an AI makes me feel less connected to the story because I know it's not real."

Ethical Considerations and Implications in Mental Health

From our studies, we show that retrieval of human-written stories can encourage human-human empathy rather than empathy toward AI systems, which has broader implications in the digital mental health domain. Large, pretrained generative models do not truly experience the situations present in stories. As such, mental health or social support chatbots powered by AI represent a population sourced from large quantities of human data, but still fall short of human-written stories in their empathic quality [9,13,25,40]. This appropriation of human experiences could be subverted by using AI to instead retrieve more empathically similar texts between human authors [28], such as in social support group settings via web, or to mediate human-human communications, such as between the patient and therapist [41].

To ensure ethical deployment of chatbots and LLMs more broadly in the mental wellness domain, the field of AI has historically advocated for transparency as an ethical design tenet [42]. The more transparent a system is, the more agency one has in the way they use it. However, we show that in framing interactions with stories, a one-sentence disclosure of the author significantly decreased empathy. This finding might be in tension with systems that rely on empathy for efficacy, such as in persuasive technologies that use bonds with AI to improve mental wellness outcomes [15,43]. The empathy and transparency trade-off might not be mutually exclusive, as transparency can breed trust, which also influences interaction. Our work paves directions research should be conducted to understand long-term effects of transparency on the outcomes of chatbots for mental health.

Limitations and Future Work

The primary limitation in our study design is that not all participants were exposed to all conditions. Given the number of conditions (varying generation or retrieval and transparent or not transparent author), we opted to mix within-subject comparisons and cross-study comparisons, resulting in a less clean study design. However, given the size of our online study, with around 200 participants per study, our results are still statistically sound. Future work can aim to replicate our findings with different study designs to confirm the psychological insights' soundness.

In addition, given the nature of crowdsourcing and the demographic pool of participants we surveyed, it is important to ensure that findings are replicated in other diverse populations. Although our studies were roughly balanced by gender, prolific respondents are predominantly White. Future work can assess the impact of identity on empathetic reaction to stories told by AI systems.

Another limitation of this work is that the quality of stories written by users may have affected the generated or retrieved stories from ChatGPT and the human-written stories database. This could have downstream effects on the user's empathy toward the story. Although we did not explore this confound in this paper, comparisons between human-written and AI-written stories were both conditioned on just the user's story. Our findings indicate that, at large, empathy patterns shift depending on transparency of the author but did not explore personal nuances in the quality of the user's story. Future works can aim to quantify the quality of written stories and how this might affect empathetic response.

This work focuses on human perceptions of AI story sharing, which can have implications in chatbot design. Such implications are extendable to mental health or social support chatbots that have their own identities or self-disclose their own personal experiences. However, these implications might not apply specifically to chatbots that serve the function of delivering therapy sessions without story sharing. As such, future work should explore the role of transparency regarding machine-like quality or human-like quality in mental health chatbot sessions that are not specific to story sharing.

Finally, there is still a key question to be asked about what role the agent should play in mental health domains, and where empathy fits into this context of human-AI interaction. In traditional patient-therapist relationships, therapeutic alliance, or the working relationship between the two, is a key component and leads to stronger patient outcomes [44]. AI chatbots have been designed to model this type of alliance through verbal empathy or expressing their understanding of the user. The stories presented to the study participants in this study are one way an agent can demonstrate empathy [45]. It is important to note that disclosing personal anecdotes as a form of empathy is different from traditional therapist-client relationships, where the therapists typically shared limited information about themselves [46]. However, there are other supportive relationships or interactions, including companion agents or coaches, that could be mediated by AI technologies. This work paves interesting future directions for how to think about the presentation of model outputs in the context of empathy and personal experiences across a multitude of domains.

Conclusions

A growing number of companies and research institutions propose using language models and AI chatbots to improve mental well-being or social companionship. Empathy is a core tenet at the center of these chatbot designs, making it crucial to consider the ethical question of how empathy unfolds toward human versus AI narrators, and the role of transparency in this effect. To this end, we conducted 4 crowdsourced studies to assess how empathy differs across human-written versus AI-written stories, varying how stories are selected (generation vs retrieval) and author disclosure (transparency that story was written by an AI author vs no transparency). Although we use current state-of-the-art empathetic retrieval and generation in this work, our findings provide more generalized future insights around human behavior when interacting with AI chatbots. We find that transparency of the author plays an important role in empathy toward an AI story as well as people's willingness to empathize toward machines. In particular, people generally empathize more with human-written stories but self-report more willingness to empathize with AI, indicating that transparency might play a role in fostering greater trust. This has implications in the design of AI systems intended to empathize *with* or evoke empathy *from* people. Designers of AI applications should consider explainable AI frameworks to make transparent how system content has been generated, as these can affect interaction outcomes. Our work motivates future directions regarding the social, psychological, and ethical implications of nuanced AI system design considerations that can drastically affect the ways in which humans extend empathy to artificial agents in the broader mental health and social support domains.

Acknowledgments

We would like to thank all of our participants, annotators, and teammates for their invaluable contributions to this project. This work was supported by an NSF GRFP (grant 2141064) and the IITP grant funded by the Korean Ministry of Science and ICT (2020-0-00842).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Full user study task instructions and surveys.
[DOCX File , 759 KB - mental_v11i1e62679_app1.docx]

References

1. Decety J, Jackson PL. The functional architecture of human empathy. Behav Cogn Neurosci Rev 2004;3(2):71-100. [doi: 10.1177/1534582304267187] [Medline: 15537986]
2. Cuff BMP, Brown SJ, Taylor L, Howat DJ. Empathy: a review of the concept. Emot Rev 2014;8(2):144-153. [doi: 10.1177/1754073914558466]
3. Cho E, Jeon S. The role of empathy and psychological need satisfaction in pharmacy students' burnout and well-being. BMC Med Educ 2019;19(1):43 [FREE Full text] [doi: 10.1186/s12909-019-1477-2] [Medline: 30717723]
4. Ta V, Griffith C, Boatfield C, Wang X, Civitello M, Bader H, et al. User experiences of social support from companion chatbots in everyday contexts: thematic analysis. J Med Internet Res 2020;22(3):e16235 [FREE Full text] [doi: 10.2196/16235] [Medline: 32141837]
5. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth 2018;6(11):e12106 [FREE Full text] [doi: 10.2196/12106] [Medline: 30470676]
6. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
7. Adam M, Wessel M, Benlian A. AI-based chatbots in customer service and their effects on user compliance. Electron Markets 2020;31(2):427-445. [doi: 10.1007/s12525-020-00414-7]
8. Yoon N, Lee H. AI recommendation service acceptance: assessing the effects of perceived empathy and need for cognition. J Theor Appl Electron Commer Res 2021;16(5):1912-1928. [doi: 10.3390/jtaer16050107]
9. Pelau C, Dabija D, Ene I. What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. Comput Human Behav 2021;122:106855. [doi: 10.1016/j.chb.2021.106855]
10. Paiva A, Dias J, Sobral D, Aylett R, Sobreperez P, Woods S, et al. Caring for agents and agents that care: building empathic relations with synthetic agents. 2005 Presented at: Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems, 2004 (AAMAS 2004); 2004; New York, NY.
11. Tsumura T, Yamada S. Influence of agent's self-disclosure on human empathy. PLoS One 2023;18(5):e0283955 [FREE Full text] [doi: 10.1371/journal.pone.0283955] [Medline: 37163467]
12. Spitale M, Okamoto S, Gupta M, Xi H, Matarić MJ. Socially assistive robots as storytellers that elicit empathy. ACM Trans Hum.-Robot Interact 2022;11(4):1-29. [doi: 10.1145/3538409]
13. Schaaff K, Reinig C, Schlippe T. Exploring ChatGPT's empathic abilities. ArXiv Preprint posted online on 2023 September 22, 2023. [doi: 10.48550/arXiv.2308.03527]
14. Lee Y, Ha M, Kwon S, Shim Y, Kim J. Egoistic and altruistic motivation: how to induce users' willingness to help for imperfect AI. Comput Hum Behav 2019;101:180-196. [doi: 10.1016/j.chb.2019.06.009]
15. Lv X, Yang Y, Qin D, Cao X, Xu H. Artificial intelligence service recovery: the role of empathic response in hospitality customers' continuous usage intention. Comput Hum Behav 2022;126:106993. [doi: 10.1016/j.chb.2021.106993]
16. Chung LL, Kang J. "I'm Hurt Too": the effect of a chatbot's reciprocal self-disclosures on users' painful experiences. Arch Des Res 2023;36(4):67-85.
17. Oatley K. Fiction: simulation of social worlds. Trends Cogn Sci 2016;20(8):618-628. [doi: 10.1016/j.tics.2016.06.002] [Medline: 27449184]
18. Djikic M, Oatley K, Moldoveanu MC. Reading other minds: effects of literature on empathy. Sci Study Lit 2013;3(1):28-47. [doi: 10.1075/ssol.3.1.06dji]
19. Abercrombie G, Curry AC, Dinkar T, Rieser V, Talat Z. Mirages: on anthropomorphism in dialogue systems. 2023 Presented at: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; December 1, 2023; Singapore p. 4776-4790.
20. Alonso M. Can robots have personal identity? Int J Soc Robotics 2023;15(2):211-220. [doi: 10.1007/s12369-022-00958-y]
21. Gabriel S, Puri I, Xu X, Malgaroli M, Ghassemi M. Can AI relate: testing large language model response for mental health support. ArXiv Preprint posted online on May 20, 2024. [doi: 10.48550/arXiv.2405.12021]
22. Brandtzaeg PB, Skjuve M, Følstad A. My AI friend: how users of a social chatbot understand their human–AI friendship. Hum Commun Res 2022;48(3):404-429.
23. Croes EAJ, Antheunis ML. Can we be friends with Mitsuku? a longitudinal study on the process of relationship formation between humans and a social chatbot. J Soc Pers Relatsh 2020;38(1):279-300. [doi: 10.1177/0265407520959463]
24. Zaki J. The War for Kindness: Building Empathy in a Fractured World. New York: Crown; 2019.
25. Giorgi S, Markowitz DM, Soni N, Varadarajan V, Mangalik S, Schwartz HA. "I slept like a baby": using human traits to characterize deceptive ChatGPT and human text. 2023 Presented at: 1st International Workshop on Implicit Author Characterization from Texts for Search and Retrieval (IACT'23); 2023; Taipei, Taiwan p. 23-37.
26. Ishowo-Oloko F, Bonnefon JF, Soroye Z, Crandall J, Rahwan I, Rahwan T. Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. Nat Mach Intell 2019;1(11):517-521. [doi: 10.1038/s42256-019-0113-5]

27. Straten CLV, Peter J, Kühne R, Barco A. Transparency about a robot's lack of human psychological capacities: effects on child-robot perception and relationship formation. ACM Trans Hum-Robot Interact 2020;9(2):1-22. [doi: 10.1145/3365668]
28. Shen J, Sap M, Colon-Hernandez P, Park H, Breazeal C. Modeling empathic similarity in personal narratives. 2023 Presented at: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; December 1, 2023; Singapore.
29. Atkinson R. The Life Story Interview. Thousand Oaks, CA: Sage Publications; 1998.
30. Grootendorst M. KeyBERT: Minimal keyword extraction with BERT. Zenodo. 2020. URL: https://doi.org/10.5281/zenodo.4461265 [accessed 2024-09-02]
31. Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. 2019 Presented at: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019 November; Hong Kong, China.
32. Shen L. On a scale of state empathy during message processing. West J Commun 2010;74(5):504-524. [doi: 10.1080/10570314.2010.512278]
33. Patton MQ. Qualitative research. In: Everitt BS, Howell DC, editors. Encyclopedia of Statistics in Behavioral Science. Hoboken: John Wiley & Sons, Ltd; 2005:bsa514.
34. Rashkin H, Smith EM, Li M, Boureau YL. Towards empathetic open-domain conversation models: a new benchmark and dataset. 2019 Presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; July 1, 2019; Florence, Italy.
35. Krishna K, Song Y, Karpinska M, Wieting J, Iyyer M. Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. ArXiv Preprint posted online on October 18, 2023. [doi: 10.48550/arXiv.2303.13408]
36. Liu B. In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human–AI interaction. J Comput Mediat Commun 2021;26(6):384-402.
37. Ehsan U, Liao QV, Muller M, Riedl MO, Weisz JD. Expanding explainability: towards social transparency in AI systems. 2021 Presented at: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; May 7, 2021; Yokohama, Japan p. 1-19.
38. Natarajan M, Gombolay M. Effects of anthropomorphism and accountability on trust in human robot interaction. 2020 Presented at: Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction; March 9, 2020; Cambridge, United Kingdom p. 33-42.
39. Jones CR, Bergen BK. Does GPT-4 pass the turing test? ArXiv Preprint posted online on April 20, 2024. [doi: 10.48550/arXiv.2310.20216]
40. Montemayor C, Halpern J, Fairweather A. In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. AI Soc 2022;37(4):1353-1359 [FREE Full text] [doi: 10.1007/s00146-021-01230-z] [Medline: 34054228]
41. Hohenstein J, Jung M. AI as a moral crumple zone: the effects of AI-mediated communication on attribution and trust. Comput Hum Behav 2020;106:106190. [doi: 10.1016/j.chb.2019.106190]
42. Yampolskiy RV. Taxonomy of pathways to dangerous artificial intelligence. 2016 Presented at: The Workshops of the 30th AAAI Conference on Artificial Intelligence; March 29, 2016; San Jose, CA p. 143-148.
43. Jeong S, Aymerich-Franch L, Alghowinem S, Picard R, Breazeal C, Park H. A robotic companion for psychological well-being: a long-term investigation of companionship and therapeutic alliance. 2023 Presented at: Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction; March 13, 2023; Stockholm, Sweden p. 485-494. [doi: 10.1145/3568162.3578625]
44. Martin DJ, Garske JP, Davis MK. Relation of the therapeutic alliance with outcome and other variables: a meta-analytic review. J Consult Clin Psychol 2000;68(3):438-450. [Medline: 10883561]
45. Bickmore T, Gruber A, Picard R. Establishing the computer-patient working alliance in automated health behavior change interventions. Patient Educ Couns 2005;59(1):21-30. [doi: 10.1016/j.pec.2004.09.008] [Medline: 16198215]
46. Irarrázaval L, Kalawski JP. Phenomenological considerations on empathy and emotions in psychotherapy. Front Psychol 2022;13:1000059 [FREE Full text] [doi: 10.3389/fpsyg.2022.1000059] [Medline: 36300059]

Abbreviations

AI: artificial intelligence
LLM: large language model

Edited by C Blease; submitted 28.05.24; peer-reviewed by J Hagstrom, RS Gomaa Mahmoud, A Jain, H Susanto; comments to author 17.07.24; revised version received 05.08.24; accepted 20.08.24; published 25.09.24.

Please cite as:
Shen J, DiPaola D, Ali S, Sap M, Park HW, Breazeal C
Empathy Toward Artificial Intelligence Versus Human Experiences and the Role of Transparency in Mental Health and Social Support Chatbot Design: Comparative Study
JMIR Ment Health 2024;11:e62679
URL: https://mental.jmir.org/2024/1/e62679
doi:10.2196/62679
PMID:

Coding of Childhood Psychiatric and Neurodevelopmental Disorders in Electronic Health Records of a Large Integrated Health Care System: Validation Study

Jiaxiao M Shi[1], PhD; Vicki Y Chiu[1], MS; Chantal C Avila[1], MA; Sierra Lewis[1], MPH; Daniella Park[1], MPH; Morgan R Peltier[2], PhD; Darios Getahun[1], MD, MPH, PhD

1

2

Corresponding Author:
Jiaxiao M Shi, PhD

Abstract

Background: Mental, emotional, and behavioral disorders are chronic pediatric conditions, and their prevalence has been on the rise over recent decades. Affected children have long-term health sequelae and a decline in health-related quality of life. Due to the lack of a validated database for pharmacoepidemiological research on selected mental, emotional, and behavioral disorders, there is uncertainty in their reported prevalence in the literature.

Objectives: We aimed to evaluate the accuracy of coding related to pediatric mental, emotional, and behavioral disorders in a large integrated health care system's electronic health records (EHRs) and compare the coding quality before and after the implementation of the *International Classification of Diseases, Tenth Revision, Clinical Modification* (*ICD-10-CM*) coding as well as before and after the COVID-19 pandemic.

Methods: Medical records of 1200 member children aged 2-17 years with at least 1 clinical visit before the COVID-19 pandemic (January 1, 2012, to December 31, 2014, the *ICD-9-CM* coding period; and January 1, 2017, to December 31, 2019, the *ICD-10-CM* coding period) and after the COVID-19 pandemic (January 1, 2021, to December 31, 2022) were selected with stratified random sampling from EHRs for chart review. Two trained research associates reviewed the EHRs for all potential cases of autism spectrum disorder (ASD), attention-deficit hyperactivity disorder (ADHD), major depression disorder (MDD), anxiety disorder (AD), and disruptive behavior disorders (DBD) in children during the study period. Children were considered cases only if there was a mention of any one of the conditions (yes for diagnosis) in the electronic chart during the corresponding time period. The validity of diagnosis codes was evaluated by directly comparing them with the gold standard of chart abstraction using sensitivity, specificity, positive predictive value, negative predictive value, the summary statistics of the *F*-score, and Youden *J* statistic. κ statistic for interrater reliability among the 2 abstractors was calculated.

Results: The overall agreement between the identification of mental, behavioral, and emotional conditions using diagnosis codes compared to medical record abstraction was strong and similar across the *ICD-9-CM* and *ICD-10-CM* coding periods as well as during the prepandemic and pandemic time periods. The performance of AD coding, while strong, was relatively lower compared to the other conditions. The weighted sensitivity, specificity, positive predictive value, and negative predictive value for each of the 5 conditions were as follows: 100%, 100%, 99.2%, and 100%, respectively, for ASD; 100%, 99.9%, 99.2%, and 100%, respectively, for ADHD; 100%, 100%, 100%, and 100%, respectively for DBD; 87.7%, 100%, 100%, and 99.2%, respectively, for AD; and 100%, 100%, 99.2%, and 100%, respectively, for MDD. The *F*-score and Youden *J* statistic ranged between 87.7% and 100%. The overall agreement between abstractors was almost perfect (κ=95%).

Conclusions: Diagnostic codes are quite reliable for identifying selected childhood mental, behavioral, and emotional conditions. The findings remained similar during the pandemic and after the implementation of the ICD-10-CM coding in the EHR system.

(JMIR Ment Health 2024;11:e56812) doi:10.2196/56812

KEYWORDS

autism; autism spectrum disorder; ASD; attention deficit hyperactivity disorder; ADHD; disruptive behavioral disorders; DBD; anxiety disorders; AD; major depression disorder; MDD; autistic; coding; neurodevelopmental; psychiatric; electronic health record; electronic health records; validation; accuracy; mental health; emotional; behavior; behaviors; behavioral; disorder; disorders; pediatric; pediatrics; paediatric; infant; paediatrics; infants; infancy; baby; babies; neonate; neotnates; neonatal; toddler; toddlers; child; children; hospital; hospitals

Introduction

Children and adolescents are particularly vulnerable to chronic mental and behavioral conditions because their brain continues to develop. Childhood mental and behavioral disorders, including autism spectrum disorder (ASD), attention-deficit hyperactivity disorder (ADHD), disruptive behavior disorders (DBD), anxiety disorder (AD), and major depressive disorder (MDD), are common neurological disorders and are on the rise in recent decades [1-4]. Affected children and adolescents are subjected to long-term negative health and social consequences [5,6], leading to significant health care costs and public health burden [7,8]. Therefore, accurately estimating their incidence and prevalence is important to guide policy-making, resource allocation, and implementation of different intervention programs.

Trends in mental and behavioral disorders are difficult to examine using routinely collected data and often are difficult to compare across studies because of differences in case ascertainment methods. Therefore, reported incidence and prevalence rates vary widely across studies [9], and the accuracy of case ascertainment has been challenged by researchers [10,11]. This problem is further complicated by the accuracy of coding after the mandatory introduction of the *International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM)* coding systems to classify diagnoses and procedures in the United States, which occurred on October 1, 2015 [12]. The *ICD-10-CM* provides increased specificity and detail for many health conditions [12], including childhood psychiatric and neurodevelopmental conditions. Many health care providers adopted the electronic health record (EHR) enacted by the Health Information Technology for Economic and Clinical Health Act (HITECH Act) in 2009 [13,14].

The Kaiser Permanente Southern California (KPSC)–integrated EHR data provide researchers with important health information to perform pharmacoepidemiological studies of children's behavioral, medical, and psychiatric conditions, including examining the public health impact of childhood psychiatric and neurodevelopmental conditions; however, there is uncertainty over the accuracy of the clinical diagnosis codes requiring validation before their use. Therefore, the objective of this study was to perform an EHR review on the validity of diagnosis codes during 3 time points (*ICD-9-CM* and *ICD-10-CM* coding and before and after the COVID-19 pandemic) for ascertaining psychiatric and neurodevelopmental conditions in a large socioeconomically diverse pediatric population aged 2-17 years [15]. The validation for the pre- and post–COVID-19 pandemic is important to assess how much the diagnosis coding has been impacted by increasing the use of virtual visits (telephone and video-assisted encounters).

Methods

Study Setting

This study was conducted using data on member children extracted from the KPSC EHR. The KPSC health care system provides services to over 4.8 million members in 15 hospitals and 234 medical offices throughout southern California. Mental health services are provided to member patients by qualified providers at in- and outpatient psychiatric care facilities in KPSC settings. Although most members receive their care at KPSC hospitals and <10% use contracting hospitals, all diagnostic, procedural, and pharmacy records are captured and maintained by the KPSC EHR since its full implementation in 2008. The sociodemographic characteristics of KPSC members closely reflect the California population [15].

Ethical Considerations

This study was conducted with approval by the KPSC Institutional Review Board (IRB# 13114). Informed consent was waived, as the study was low risk and strictly involved the use of internal EHR data, accessible only to authorized personnel when needed.

Study Design and Sample Selection

Data for this validation study were obtained retrospectively from children who were members of the KPSC health care system during three distinct time periods: (1) January 1, 2012, through December 31, 2014; (2) January 1, 2017, through December 31, 2019; and (3) January 1, 2021, through December 31, 2022. We carefully selected the 3 time periods to investigate the medical coding accuracy of selected psychiatric and neurodevelopmental conditions encompassing both the *ICD-9-CM* and *ICD-10-CM* periods as well as the pre- and post–COVID-19 pandemic eras. To be included in the validation study cohort for each time period, children must have been enrolled in the KPSC health care system for at least 1 year during the corresponding time period at specific age ranges varying by condition (5-17 years of age for ADHD, 2-17 years of age for ASD, and 3-17 years of age for the other 3 conditions—AD, MDD, and DBD). Children may present with signs and symptoms of the specified conditions very early in life; however, for this study, we used reported and reliable lower age groupings for ASD and ADHD diagnoses [16,17] and widely published age goupings for anxiety and depressive disorders [18,19]. Furthermore, at least 1 clinical visit, including virtual visits, in each corresponding time period was required.

In the KPSC system, diagnosis codes for clinical visits (eg, hospitalization, outpatient office visits, and emergency department visits) from all KPSC facilities are extracted from EHRs and entered into a structured database by professional medical coders from the clinical data management team. In this validation study, coding-based outcomes of interest were ascertained using these clinical diagnosis codes within the 3 specified time periods (Multimedia Appendix 1 presents the *ICD-9-CM* and *ICD-10-CM* codes).

For each of the investigated conditions, we randomly sampled 40 cases according to the following strata: (1) those without a diagnosis (No-Dx) and (2) those with a diagnosis (Dx). Thus, records for a total of 1200 individual children were selected. The accuracy for each of the 2 strata, with or without documented diagnostic records, was expected to be around 85%. A sample size of 120 per diagnostic condition would provide less than a 5% one-sided margin for a 90% CI of the accuracy for each stratum.

Diagnosing ASD, ADHD, DBD, and MDD

The KPSC system has an integrated framework for in- and outpatient as well as emergency department encounter services. During the child's visit to any of these facilities, the practitioner has access to the child's diagnoses, but often the diagnosis of ASD, ADHD, and DBD is made in an outpatient setting. The following criteria were used to diagnose and code ASD, ADHD, and DBD within the KPSC setting: (1) a Child Behavior Checklist must be filled out by parents and teachers to describe the child's behavioral and emotional problems and (2) a clinical interview must be performed by a qualified mental health professional. In a preliminary study conducted for this project, 96% of children with ASD, ADHD, and DBD were found to have had their conditions diagnosed by KPSC child and adolescent psychiatrists, developmental and behavioral pediatricians, child psychologists, and neurologists consistent with the diagnostic criteria from the *Diagnostic and Statistical Manual of Mental Disorders-Fifth Edition* (*DSM-V*). Diagnosis of the remaining 4% was confirmed upon membership, as these cases had been previously diagnosed outside the KPSC system [20,21].

The diagnosis of depression disorder, in the KPSC system, was based on the US Preventive Services Task Force on screening for depression in children and adolescents recommendation [22]. The KPSC system uses the Patient Health Questionnaire (PHQ-9) and the PHQ-9 modified for Adolescents (PHQ-A). If a patient's score on the PHQ-9 or PHQ-A does not seem to accurately reflect observed clinical symptoms, *DSM-V* criteria are recommended for diagnosis.

To ascertain the neurodevelopmental conditions investigated in this study, we relied on systemwide clinical diagnoses made by experts in the field as mentioned above.

Chart Abstraction Process

Trained research associates (abstractors) reviewed EHRs for documentation of a diagnosis (yes/no) of each condition under investigation for the selected sample of children during the study period. Children were considered to have the disorder in the presence of documented evidence of that condition noted in the chart during the corresponding time period of investigation. To ensure data quality and consistency of chart reviews between the 2 abstractors, a total of 180 cases, stratified by the 2 strata (Dx and No-Dx), were randomly selected for reabstraction (90 per abstractor and 36 per condition). There was a total of 4 possible responses for each chart reviewed case, as follows: diagnosis "Yes," diagnosis "No," "Unable to ascertain due to blocked notes," and "Unable to ascertain due to insufficient notes." The abstractors based their responses on the information they were able to ascertain from the clinical notes. For example, due to the sensitive nature of psychiatric diagnoses, some progress notes may have been blocked; therefore, if the abstractors could not ascertain the diagnosis due to blocked notes, it was coded as such. Similarly, if the abstractors could not ascertain a diagnosis due to a lack of documentation or notes in the KPSC system with underused care (eg, outside claims data), the record was coded as "Unable to ascertain due to insufficient notes."

The results of this assessment informed our full chart review process of the 1200 medical records as mentioned above. In other words, we excluded those claims data records from the random selection of the validation sample. Furthermore, during chart review, records in which the abstractors encountered blocked notes and were therefore unable to make a determination were flagged. These flagged records (n=38, 3%) were replaced with another randomly selected record for chart review from the same strata. Potential cases that were still unclear in the clinical use records were adjudicated by the study investigator with expertise in the field (DG). The child psychiatric and neurodevelopmental condition cases abstracted through this process served as the gold standard.

Child Characteristics

Characteristics for KPSC member children included age (2-5, 6-11, and 12-17 years), sex (male and female), race/ethnicity (categorized as non-Hispanic White, non-Hispanic Black, Hispanic, Asian/Pacific Islander, other/multiple, and unknown), median household income in US dollars (<30,000; 30,000-49,999; 50,000-69,999; 70,000-89,999; and ≥90,000), and insurance type (Medicaid, commercial through employment, private/individual, and other).

We obtained the characteristics of all children of the state of California residents during the same time periods using publicly available data posted on the Centers for Disease Control and Prevention Wonder website [3]. Both the KPSC EHR and the Centers for Disease Control and Prevention (Wonder) sources provided information on child characteristics, including age and race/ethnicity. Data on median household income were estimated based on census tracts for KPSC patients.

Statistical Analysis

We described the characteristics of our study population between those with and without the conditions of interest. Furthermore, we investigated how representative the KPSC pediatric population is compared to the state of California pediatric population during the entire study period using frequency distributions. For the purpose of comparison with the California children population, the age of the children for the KPSC population was evaluated based on the date of randomly selected clinical visits for each child during the entire study period.

The agreement between the 2 abstractors was evaluated with the interrater reliability assessment (κ statistic) by using the initial 180 chart reviews. We compared findings from the manual chart review of the 1200 cases, set as the gold standard, with corresponding diagnosis records for he *ICD-9-CM* and *ICD-10-CM* coding periods as well as before and after the COVID-19 pandemic through sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). These performance measurements were reported as weighted percentages with corresponding 95% CIs using normalized sampling weights (*W*), as defined below:

$$Wij=\#\text{ in stratum } j120\times(\#\text{total study population } i), i=1 \text{ to } 5, j=1,2$$

where i is for each of the 5 conditions and j is for the corresponding two strata (Dx and No-Dx) within each condition. To evaluate the overall performance, we also reported the summary statistics of *F*-score and Youden *J* statistic, which are composite measurements of sensitivity and PPV (*F*-score) or sensitivity and specificity (*J* statistic).

All analyses were conducted using SAS statistical software (version 9.4; SAS Institute, Inc).

Results

An overview of the patient characteristics for our sample, study population, and the state of California pediatric population are shown in Table 1. Overall, 2,283,931 children from all KPSC hospitals and medical offices were obtained during the 3 time periods. Compared with the state of California pediatric population, KPSC had similar distributions on the children’s sex but had slightly higher percentages of children for ages 2-5 years (25.6% for KPSC vs 24.7% for the state of California) and 6-11 (37.8% for KPSC vs 37.4% for the state of California). In addition, KPSC had slightly lower percentages of children identified as non-Hispanic White (24.5% vs 28.5%) or Asian/Pacific Islander (10.0% vs 13.1%). However, this could be partially attributed to the larger percentage of unknown, missing, or multiple race/ethnicity in the KPSC group (6.0% vs 0.5%, respectively). Overall, the KPSC pediatric population in this study was a good representation of the entire state of California’s pediatric population. Our sample of 1200 children across the 5 conditions was broadly representative of the KPSC study population with respect to age, sex, and race/ethnicity. However, children in the older age groups were slightly oversampled. The overall κ between our 2 abstractors was 95%.

XSL•FO
RenderX

Table . Characteristics of the study cohort and the state of California pediatric population aged 2-17 years (2012-2014, 2017-2019, and 2021-2022).

Characteristics		Chart reviewed		Pediatric population	
		Yes (n=1200)	No (n=2,282,731)	KPSC[a] (n=2,283,931)	California state[b] (n=48,566,041)
Age [c] (years), n (%)					
	2-5	198 (16.5)	584,657 (25.6)	584,855 (25.6)	11,975,915 (24.7)
	6-11	473 (39.4)	863,628 (37.8)	864,101 (37.8)	18,169,267 (37.4)
	12-17	529 (44.1)	834,446 (36.6)	834,975 (36.6)	18,420,859 (37.9)
Child's sex, n (%)					
	Female	536 (44.7)	1,117,357 (48.9)	1,117,893 (48.9)	23,755,606 (48.9)
	Male	664 (55.3)	1,165,374 (51.1)	1,166,038 (51.1)	24,810,435 (51.1)
Race/ethnicity, n (%)					
	Non-Hispanic White	322 (26.8)	558,254 (24.5)	558,576 (24.5)	13,818,677 (28.5)
	Non-Hispanic Black	90 (7.5)	180,348 (7.9)	180,438 (7.9)	3,056,231 (6.3)
	Hispanic	622 (51.8)	1,178,606 (51.6)	1,179,228 (51.6)	25,096,264 (51.7)
	Asian/Pacific Islander	98 (8.2)	228,687 (10.0)	228,785 (10.0)	6,344,897 (13.1)
	Unknown/other/multiple	68 (5.7)	136,836 (6.0)	136,904 (6.0)	249,972 (0.5)
Household income (US $), n (%)					
	<30,000	24 (2.0)	58,366 (2.6)	58,390 (2.6)	—[d]
	30,000-49,999	256 (21.3)	481,524 (21.1)	481,780 (21.1)	—
	50,000-69,999	336 (28.0)	609,418 (26.7)	609,754 (26.7)	—
	70,000-89,999	260 (21.7)	482,178 (21.1)	482,438 (21.1)	—
	≥90,000	323 (26.9)	649,109 (28.4)	649,432 (28.4)	—
	Missing	1 (0.1)	2136 (0.1)	2137 (0.1)	—
Insurance type, n (%)					
	Medicaid	302 (25.2)	534,606 (23.4)	534,908 (23.4)	—[d]
	Commercial	843 (70.3)	1,642,058 (71.9)	164,2901 (71.9)	—
	Private	49 (4.1)	93,984 (4.1)	94,033 (4.1)	—
	Other	6 (0.5)	12,083 (0.5)	12,089 (0.5)	—

[a]KPSC: Kaiser Permanente Southern California; the sample is based on the KPSC data from the electronic health records (2012-2014, 2017-2019, and 202-2022).

[b]Data are from the natality information of Centers for Disease Control and Prevention website [3].

[c]Starting age is different for each condition.

[d]Median household income and insurance type information are not available for the California state data.

Table 2 shows the distribution and their sample sizes for the two strata (Dx and No-Dx) among our entire study population and by the 3 time periods. Based on these comparisons against the chart review results, the performance measurements of our EHRs are shown in Table 3. The weighted sensitivity, specificity, PPV, and NPV for each of the 5 conditions were as follows:

- ADHD: sensitivity 100%, specificity 99.9%, PPV 99.2%, and NPV 100%
- ASD: sensitivity 100%, specificity 100%, PPV 99.2%, and NPV 100%
- MDD: sensitivity 100%, specificity 100%, PPV 99.2%, and NPV 100%
- AD: sensitivity 87.7%, specificity 100%, PPV 100%, and NPV 99.2%
- DBD: sensitivity 100%, specificity 100%, PPV 100%, and NPV 100%

The corresponding *F*-score and Youden *J* statistic were 99.6% and 99.9% for ADHD, 99.6% and 100% for ASD, 99.6% and 100% for MMD, 93.4% and 87.7% for AD, and 100% and 100% for DBD, respectively. Results were similar across the *ICD-9-CM* and *ICD-10-CM* coding time periods as well as the before and after the COVID-19 pandemic.

Table . Frequencies of study population and chart review results by diagnosis codes. Condition-specific row percentages may not sum to 100% due to rounding.

Condition and diagnosis status		Study population				Sample		
		Overall, n	Before COVID-19 pandemic, n (%)		After COVID-19 pandemic (2021-2022), n (%)	Charts abstracted, n		Normalized sampling weight
			2012-2014	2017-2019		Total	Confirmed cases	
DBD [a]								
	Dx[b]	41,351	11,787 (28.5)	14,907 (36.1)	14,657 (35.5)	120	120	0.0001601556
	No-Dx[c]	2,110,254	697956 (33.1)	755,726 (35.8)	656,572 (31.1)	120	0	0.0081731777
Anxiety								
	Dx	120,484	26039 (21.6)	49,720 (41.3)	44,725 (37.1)	120	120	0.0004666439
	No-Dx	2,031,121	683704 (33.7)	720,913 (35.5)	626,504 (30.9)	120	1	0.0078666894
ASD [d]								
	Dx	47,441	9953 (21.0)	16,355 (34.5)	21,133 (44.6)	120	119	0.000173097
	No-Dx	2,236,490	741,058 (33.1)	802,267 (35.9)	693,165 (31.0)	120	0	0.0081602363
MDD [e]								
	Dx	43,919	7846 (17.9)	19,036 (43.3)	17,037 (38.8)	120	119	0.0001701017
	No-Dx	2,107,686	701,897 (33.3)	751,597 (35.7)	654,192 (31.0)	120	0	0.0081632316
ADHD [f]								
	Dx	108,717	35,668 (32.8)	40,791 (37.5)	32,258 (29.7)	120	119	0.0004826384
	No-Dx	1,768,413	587,686 (33.2)	630,533 (35.7)	550,194 (31.1)	120	0	0.0078506949

[a]DBD: disruptive behavior disorders.

[b]Dx: with confirmed diagnosis codes in electronic data.

[c]No-Dx: without confirmed diagnosis codes in electronic data.

[d]ASD: autism spectrum disorder.

[e]MDD: major depressive disorder.

[f]ADHD: attention deficit hyperactivity disorder.

Table . Weighted performance measurements for childhood psychiatric and neurodevelopmental disorders by data sources before and after the implementation of the *International Classification of Diseases, Ninth/Tenth Revisions, Clinical Modification* codes in the Kaiser Permanente Southern California (KPSC) system in 2015 (ample size=1200) and before and after the COVID-19 pandemic.

Disease and periods		Weighted performance measurements					
		Sensitivity (95% CI)	Specificity (95% CI)	PPV[a] (95% CI)	NPV[b] (95% CI)	*F*-score	Youden *J* statistic
DBD [c]							
	Overall	100 (100-100)	100 (100-100)	100 (100-100)	100 (100-100)	100	100
	2012-2014	100 (100-100)	100 (100-100)	100 (100-100)	100 (100-100)	100	100
	2017-2019	100 (100-100)	100 (100-100)	100 (100-100)	100 (100-100)	100	100
	2021-2022	100 (100-100)	100 (100-100)	100 (100-100)	100 (100-100)	100	100
Anxiety							
	Overall	87.7 (63.4-100)	100 (100-100)	100 (100-100)	99.2 (97.5-100)	93.4	87.7
	2012-2014	100 (100-100)	100 (100-100)	100 (100-100)	100 (100-100)	100	100
	2017-2019	100 (100-100)	100 (100-100)	100 (100-100)	100 (100-100)	100	100
	2021-2022	70.4 (11.3-100)	100 (100-100)	100 (100-100)	97.5 (92.5-100)	82.6	70.4
MDD [d]							
	Overall	100 (100-100)	100 (99.9-100)	99.2 (97.5-100)	100 (100-100)	99.6	100
	2012-2014	100 (100-100)	100 (100-100)	100 (100-100)	100 (100-100)	100	100
	2017-2019	100 (100-100)	100 (100-100)	100 (100-100)	100 (100-100)	100	100
	2021-2022	100 (100-100)	99.9 (99.8-100)	97.5 (92.5-100)	100 (100-100)	98.7	99.9
ASD [e]							
	Overall	100 (100-100)	100 (99.9-100)	99.2 (97.5-100)	100 (100-100)	99.6	100
	2012-2014	100 (100-100)	99.9 (99.8-100)	97.5 (92.5-100)	100 (100-100)	98.7	99.9
	2017-2019	100 (100-100)	100 (100-100)	100 (100-100)	100 (100-100)	100	100
	2021-2022	100 (100-100)	100 (100-100)	100 (100-100)	100 (100-100)	100	100
ADHD [f]							
	Overall	100 (100-100)	99.9 (99.8-100)	99.2 (97.5-100)	100 (100-100)	99.6	99.9
	2012-2014	100 (100-100)	100 (100-100)	100 (100-100)	100 (100-100)	100	100
	2017-2019	100 (100-100)	100 (100-100)	100 (100-100)	100 (100-100)	100	100
	2021-2022	100 (100-100)	99.8 (99.5-100)	97.5 (92.5-100)	100 (100-100)	98.7	99.8

[a]PPV: positive predictive value.
[b]NPV: negative predictive value.
[c]DBD: disruptive behavior disorders.
[d]MDD: major depressive disorder.
[e]ASD: autism spectrum disorder.
[f]ADHD: attention deficit hyperactivity disorder.

Discussion

Principal Findings

This validation study was performed to determine the accuracy of clinical diagnosis codes in ascertaining childhood psychiatric and neurodevelopmental cases using data abstracted from the EHR of a large integrated health care system serving a demographically diverse patient population. To our knowledge, the accuracy of data on studied conditions using clinical diagnostic codes has not been validated using EHR data. Furthermore, the extent to which the transition of *ICD-9-CM* to the *ICD-10-CM* coding system as well as the pre- and post–COVID-19 pandemic eras have impacted the ascertainment of the studied behavioral and developmental conditions is unclear. Our study showed that within a large integrated health system, there is strong agreement between the diagnosis codes (*ICD-9-CM* or *ICD-10-CM*) and the patients' conditions, including ASD, ADHD, DBD, AD, and MDD, both before and after the COVID-19 pandemic eras.

In recent years, EHRs have become important data sources for various epidemiological study designs investigating potential associations between exposures and outcomes that have become standard among researchers and health care providers as part of the American Recovery and Reinvestment Act of 2009 (specifically the HITECH Act) [14]. Although the EHR has become an important information management and care delivery system tool ensuring quality of care by providing access to comprehensive treatment-related data, its validity and completeness for conducting pharmacoepidemiological studies have been challenged by many due to how information is coded [23]. In the KPSC health care system, the process of using data coding and coding rules of the medical diagnoses and procedures recorded in patients' health records is performed by trained medical coders from the clinical data management team. Individual diagnostic conditions in the EHR need to be evaluated critically for accuracy and consistency. Furthermore, whether the accuracy of case ascertainment has been impacted by the introduction of the *ICD-10-CM* coding system as well as the effect of the COVID-19 pandemic on the quality of data capture needs to be investigated. Therefore, we performed this validation study to evaluate (1) the accuracy of pediatric mental, emotional, and behavioral disorders identified in EHRs and (2) the coding quality before and after the implementation of the *ICD-10-CM* codings as well as before and after the COVID-19 pandemic.

The findings of this study suggest that the validity of EHR data for the identification of childhood mental, behavioral, and emotional conditions is quite strong and the transition from the *ICD-9-CM* coding system to the *ICD-10-CM* coding system as well as coding during the COVID-19 pandemic had minimal impact on the overall accuracy of case ascertainment.

The main strength of this study is the large chart abstraction conducted to assess the validity and reliability of childhood mental, behavioral, and emotional disorders case ascertainment using EHR data extracted from a large integrated health care system. The EHR system database provides an opportunity for neurodevelopmental outcome investigation with a high degree of validity of case ascertainment for epidemiological studies, in addition to current developments by others using natural language processing algorithms [24-27]. This validation study, comprised of a demographically diverse southern California population, is likely generalizable to health care settings with similar EHR database systems. Although the overall agreement between our abstractors was almost perfect (κ=95%), a potential limitation of this study was the use of medical record abstractors who were not blinded to the source of the data. However, a previous study that evaluated the agreement between masked and unmasked medical record abstraction reported no impact of bias in case ascertainment [28]. A further potential limitation is the fact that we had some records with blocked notes that needed to be replaced because of incompleteness in ascertaining the conditions of interest. However, in close examination, we found that the blocked notes were only 38 (3%) out of the 1200 cases, which we believe will have minimal impact on the overall analysis. In addition, we did not consider oversampling of the older age strata in the summarized performance metrics. Considering that the older age group could be given a more accurate diagnosis coding, the actual concordance of the electronic coding might be slightly lower than what we observed in this study.

Conclusions

Our findings suggest that childhood mental, behavioral, and emotional disorders are reliably coded in the EHRs and can be used for pharmacoepidemiological studies. Furthermore, the completeness of data remained similar during the pre- and postpandemic eras and after the implementation of the *ICD-10-CM* coding in the EHR system.

Acknowledgments

This research was funded by the Garfield Memorial Fund (project number RNG211534) to DG. The opinions expressed are solely the authors' responsibility and do not necessarily reflect the official views of the funding agency. The Garfield Memorial Fund study team would like to thank Kaiser Permanente members who contributed electronic health information to this study. The authors thank Ms Sole Cardoso for her technical support.

Conflicts of Interest

None declared.

Multimedia Appendix 1
International Classification of Diseases, Ninth/Tenth Revisions, Clinical Modification diagnostic codes to ascertain mental, emotional, and behavioral disorders.
[DOCX File, 14 KB - mental_v11i1e56812_app1.docx]

References

1. Khadka N, Peltier MR, Fassett MJ, et al. Rising trends of childhood attention-deficit/hyperactivity disorder in a large integrated healthcare delivery system in Southern California, 2010-2021. J Pediatr 2024 Mar 1;269:113997. [doi: 10.1016/j.jpeds.2024.113997] [Medline: 38432293]
2. Lebrun-Harris LA, Ghandour RM, Kogan MD, Warren MD. Five-year trends in US children's health and well-being, 2016-2020. JAMA Pediatr 2022 Jul 1;176(7):e220056. [doi: 10.1001/jamapediatrics.2022.0056] [Medline: 35285883]

3. Natality information. Centers for Disease Control and Prevention. URL: https://wonder.cdc.gov/natality.html [accessed 2023-12-25]
4. Bitsko RH, Claussen AH, Lichstein J, et al. Mental health surveillance among children. MMWR 2022 Feb;71:1-42. [doi: 10.15585/mmwr.su7102a1]
5. Wang PS, Berglund PA, Olfson M, Kessler RC. Delays in initial treatment contact after first onset of a mental disorder. Health Serv Res 2004 Apr;39(2):393-415. [doi: 10.1111/j.1475-6773.2004.00234.x] [Medline: 15032961]
6. Clement S, Schauman O, Graham T, et al. What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies. Psychol Med 2015 Jan;45(1):11-27. [doi: 10.1017/S0033291714000129] [Medline: 24569086]
7. Tkacz J, Brady BL. Increasing rate of diagnosed childhood mental illness in the United States: incidence, prevalence and costs. Public Health Pract (Oxf) 2021 Oct 15;2:100204. [doi: 10.1016/j.puhip.2021.100204] [Medline: 36101631]
8. AAP-AACAP-CHA declaration of a national emergency in child and adolescent mental health. American Academy of Pediatrics. URL: https://www.aap.org/en/advocacy/child-and-adolescent-healthy-mental-development/aap-aacap-cha-declaration-of-a-national-emergency- in-child-and-adolescent-mental-health/ [accessed 2023-12-25]
9. Merikangas KR, He JP, Burstein M, et al. Lifetime prevalence of mental disorders in U.S. adolescents: results from the National Comorbidity Survey Replication--Adolescent Supplement (NCS-A). J Am Acad Child Adolesc Psychiatry 2010 Oct;49(10):980-989. [doi: 10.1016/j.jaac.2010.05.017] [Medline: 20855043]
10. Bethell CD, Garner AS, Gombojav N, Blackwell C, Heller L, Mendelson T. Social and relational health risks and common mental health problems among US children: the mitigating role of family resilience and connection to promote positive socioemotional and school-related outcomes. Child Adolesc Psychiatr Clin N Am 2022 Jan;31(1):45-70. [doi: 10.1016/j.chc.2021.08.001] [Medline: 34801155]
11. Polanczyk GV, Salum GA, Sugaya LS, Caye A, Rohde LA. Annual research review: a meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. J Child Psychol Psychiatry 2015 Mar;56(3):345-365. [doi: 10.1111/jcpp.12381] [Medline: 25649325]
12. World Health Organization. ICD-10: International Statistical Classification of Diseases and Related Health Problems: Tenth Revision, 2nd edition: World Health Organization; 2004. URL: https://iris.who.int/handle/10665/42980 [accessed 2024-05-07]
13. HITECH Act. HIPAA Survival Guide. 2009. URL: https://www.hipaasurvivalguide.com/hitech-act-text.php [accessed 2023-12-25]
14. Adler-Milstein J, Jha AK. HITECH Act drove large gains in hospital electronic health record adoption. Health Aff (Millwood) 2017 Aug 1;36(8):1416-1422. [doi: 10.1377/hlthaff.2016.1651] [Medline: 28784734]
15. Koebnick C, Langer-Gould AM, Gould MK, et al. Sociodemographic characteristics of members of a large, integrated health care system: comparison with US Census Bureau data. Perm J 2012;16(3):37-41. [doi: 10.7812/TPP/12-031] [Medline: 23012597]
16. Baker BL, Neece CL, Fenning RM, Crnic KA, Blacher J. Mental disorders in five-year-old children with or without developmental delay: focus on ADHD. J Clin Child Adolesc Psychol 2010;39(4):492-505. [doi: 10.1080/15374416.2010.486321] [Medline: 20589561]
17. Lord C, Risi S, DiLavore PS, Shulman C, Thurm A, Pickles A. Autism from 2 to 9 years of age. Arch Gen Psychiatry 2006 Jun;63(6):694-701. [doi: 10.1001/archpsyc.63.6.694] [Medline: 16754843]
18. Luby JL, Belden AC, Pautsch J, Si X, Spitznagel E. The clinical significance of preschool depression: impairment in functioning and clinical markers of the disorder. J Affect Disord 2009 Jan;112(1-3):111-119. [doi: 10.1016/j.jad.2008.03.026] [Medline: 18486234]
19. Ingeborgrud CB, Oerbeck B, Friis S, et al. Anxiety and depression from age 3 to 8 years in children with and without ADHD symptoms. Sci Rep 2023 Sep 16;13(1):15376. [doi: 10.1038/s41598-023-42412-7] [Medline: 37717097]
20. Getahun D, Fassett MJ, Peltier MR, et al. Association of perinatal risk factors with autism spectrum disorder. Am J Perinatol 2017 Feb;34(3):295-304. [doi: 10.1055/s-0036-1597624] [Medline: 28099978]
21. Getahun D, Jacobsen SJ, Fassett MJ, Chen W, Demissie K, Rhoads GG. Recent trends in childhood attention-deficit/hyperactivity disorder. JAMA Pediatr 2013 Mar 1;167(3):282-288. [doi: 10.1001/2013.jamapediatrics.401] [Medline: 23338799]
22. Siu AL, U.S. Preventive Services Task Force. Screening for depression in children and adolescents: U.S. Preventive Services Task Force recommendation statement. Ann Intern Med 2016 Mar 1;164(5):360-366. [doi: 10.7326/M15-2957] [Medline: 26858097]
23. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 1;20(1):144-151. [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
24. Lingren T, Chen P, Bochenek J, et al. Electronic health record based algorithm to identify patients with autism spectrum disorder. PLoS One 2016;11(7):e0159621. [doi: 10.1371/journal.pone.0159621] [Medline: 27472449]
25. Leroy G, Gu Y, Pettygrove S, Galindo MK, Arora A, Kurzius-Spencer M. Automated extraction of diagnostic criteria from electronic health records for autism spectrum disorders: development, evaluation, and application. J Med Internet Res 2018 Nov 7;20(11):e10497. [doi: 10.2196/10497] [Medline: 30404767]

26. Washington P, Wall DP. A review of and roadmap for data science and machine learning for the neuropsychiatric phenotype of autism. Annu Rev Biomed Data Sci 2023 Aug 10;6:211-228. [doi: 10.1146/annurev-biodatasci-020722-125454] [Medline: 37137169]
27. Slaby I, Hain HS, Abrams D, et al. An electronic health record (EHR) phenotype algorithm to identify patients with attention deficit hyperactivity disorders (ADHD) and psychiatric comorbidities. J Neurodev Disord 2022 Jun 11;14(1):37. [doi: 10.1186/s11689-022-09447-9] [Medline: 35690720]
28. Reisch LM, Fosse JS, Beverly K, et al. Training, quality assurance, and assessment of medical record abstraction in a multisite study. Am J Epidemiol 2003 Mar 15;157(6):546-551. [doi: 10.1093/aje/kwg016] [Medline: 12631545]

Abbreviations

AD: anxiety disorder
ADHD: attention deficit hyperactivity disorder
ASD: autism spectrum disorder
DBD: disruptive behavioral disorders
DSM-V: Diagnostic and Statistical Manual of Mental Disorders-Fifth Edition
Dx: with confirmed diagnosis codes in electronic data
EHR: electronic health record
HITECH: Health Information Technology for Economic and Clinical Health
***ICD-10-CM*:** *International Classification of Diseases, Tenth Revision, Clinical Modification*
***ICD-9-CM*:** *International Classification of Diseases, Ninth Revision, Clinical Modification*
KPSC: Kaiser Permanente Southern California
MDD: major depression disorder
No-Dx: without confirmed diagnosis codes in electronic data
NPV: negative predictive value
PHQ: Patient Health Questionnaire
PHQ-A: Patient Health Questionnaire modified for Adolescents
PPV: positive predictive value

Edited by J Torous; submitted 31.01.24; peer-reviewed by A Hassan, P Washington, T Adebambo; revised version received 08.04.24; accepted 09.04.24; published 14.05.24.

Please cite as:
Shi JM, Chiu VY, Avila CC, Lewis S, Park D, Peltier MR, Getahun D
Coding of Childhood Psychiatric and Neurodevelopmental Disorders in Electronic Health Records of a Large Integrated Health Care System: Validation Study
JMIR Ment Health 2024;11:e56812
URL: https://mental.jmir.org/2024/1/e56812
doi:10.2196/56812

Original Paper

Data-Driven Exploration of National Health Service Talking Therapies Care Pathways Using Process Mining: Retrospective Cohort Study

Elizabeth Yardley[1], BSc, MSc; Alice Davis[2], PhD; Chris Eldridge[2], PGD; Christos Vasilakis[1], PhD

[1]School of Management, University of Bath, Bath, United Kingdom

[2]Mayden, Bath, United Kingdom

Corresponding Author:
Elizabeth Yardley, BSc, MSc
School of Management, University of Bath
Claverton Down
Bath, BA2 7AY
United Kingdom
Phone: 44 01249 701100
Email: egmy20@bath.ac.uk

Abstract

Background: The National Health Service (NHS) Talking Therapies program treats people with common mental health problems in England according to "stepped care," in which lower-intensity interventions are offered in the first instance, where clinically appropriate. Limited resources and pressure to achieve service standards mean that program providers are exploring all opportunities to evaluate and improve the flow of patients through their service. Existing research has found variation in clinical performance and stepped care implementation across sites and has identified associations between service delivery and patient outcomes. Process mining offers a data-driven approach to analyzing and evaluating health care processes and systems, enabling comparison of presumed models of service delivery and their actual implementation in practice. The value and utility of applying process mining to NHS Talking Therapies data for the analysis of care pathways have not been studied.

Objective: A better understanding of systems of service delivery will support improvements and planned program expansion. Therefore, this study aims to demonstrate the value and utility of applying process mining to NHS Talking Therapies care pathways using electronic health records.

Methods: Routine collection of a wide variety of data regarding activity and patient outcomes underpins the Talking Therapies program. In our study, anonymized individual patient referral records from two sites over a 2-year period were analyzed using process mining to visualize the care pathway process by mapping the care pathway and identifying common pathway routes.

Results: Process mining enabled the identification and visualization of patient flows directly from routinely collected data. These visualizations illustrated waiting periods and identified potential bottlenecks, such as the wait for higher-intensity cognitive behavioral therapy (CBT) at site 1. Furthermore, we observed that patients discharged from treatment waiting lists appeared to experience longer wait durations than those who started treatment. Process mining allowed analysis of treatment pathways, showing that patients commonly experienced treatment routes that involved either low- or high-intensity interventions alone. Of the most common routes, >5 times as many patients experienced direct access to high-intensity treatment rather than stepped care. Overall, 3.32% (site 1: 1507/45,401) and 4.19% (site 2: 527/12,590) of all patients experienced stepped care.

Conclusions: Our findings demonstrate how process mining can be applied to Talking Therapies care pathways to evaluate pathway performance, explore relationships among performance issues, and highlight systemic issues, such as stepped care being relatively uncommon within a stepped care system. Integration of process mining capability into routine monitoring will enable NHS Talking Therapies service stakeholders to explore such issues from a process perspective. These insights will provide value to services by identifying areas for service improvement, providing evidence for capacity planning decisions, and facilitating better quality analysis into how health systems can affect patient outcomes.

(JMIR Ment Health 2024;11:e53894) doi:10.2196/53894

KEYWORDS

electronic health record; EHR; electronic health records; EHRs; health record; data science; secondary data analysis; mental health services; mental health; health information system; HIS; information system; information systems; process mining; flow; flows; path; pathway; pathways; delivery; visualization

Introduction

The National Health Service Talking Therapies Program

Common mental health disorders such as anxiety and depression are among the top causes of disability in England [1], affecting 1 in 6 adults [2], and their prevalence is increasing [2,3]. This increasing health burden is reflected in the growing number of referrals received by the National Health Service (NHS) Talking Therapies program (formerly Improving Access to Psychological Therapies [IAPT]) since its inception in 2008 [4]. Each year, >1 million patients access Talking Therapies through the program [5], and like all parts of the NHS in England, the program is constrained by limited resources while facing additional pressure to achieve service standards for recovery rates and waiting times. On the basis of 2021 to 2022 access rates [5], the volume of people receiving treatment through the program will need to increase by 50% by 2024 to meet the targets set out in the NHS Long Term Plan [6]. Therefore, providers are looking at all opportunities to increase capacity and productivity to better understand and improve the flow of patients through care systems and to evaluate the performance of their services.

The Talking Therapies treatment model is based on the principle of stepped care, which is the approach recommended by the National Institute for Health and Care Excellence (NICE) [7]. The stepped care approach maintains that those with milder conditions should be treated initially with lower-intensity interventions, and these individuals might then be “stepped up” to higher-intensity interventions if clinically appropriate [7]. The stepped care approach has been shown to be associated with improved patient outcomes [8,9].

As detailed in Table 1, the first step in the stepped care model represents the presentation of a common mental health problem. This initial presentation is usually followed by a referral to an NHS Talking Therapies service, which generally offers step 2 and 3 interventions, while more specialist services offer higher-intensity interventions at steps beyond this.

While each therapy service will follow the overall model of care prescribed by the IAPT model [11], each service has a uniquely configured care pathway, which represents the plan for the implementation of clinical guidance. Studies have shown that there is variation in how the principle of stepped care is implemented [12,13], and clinical performance has also been shown to vary across Talking Therapies program sites [9,14]. Despite this variation in implementation and performance, methods of implementation and systems of treatment delivery are relatively understudied in comparison to the psychological therapies offered [12,13].

Studies have demonstrated the relevance of implementation methods to the program’s performance by identifying features of service delivery that are predictive of clinical outcomes and patient engagement. For instance, longer waiting times have been associated with patient disengagement with IAPT therapy services [15,16] and worse recovery outcomes [9]. More treatment sessions, a larger service size, and a greater proportion of therapy sessions delivered by experienced staff have been shown to be predictive of reliable recovery [8]. Furthermore, patients face additional waiting times between first and second treatment appointments [5]; however, only 1 known study has explored the relationship between these additional waiting times and patient engagement [16].

Furthermore, the NHS Talking Therapies program faces the challenge of patient attrition: only approximately 40% of those referred to the program attend ≥2 treatment sessions [5]. Many patients drop out of the program for unknown reasons; however, the relationship between waiting times and attrition rates cannot be explored using aggregated data. Therefore, tools that enable services to analyze wait durations will enable better research into the relationship between waiting times and patient attrition. Electronic health records (EHRs) capture a rich set of data documenting the details of patient journeys through the Talking Therapies program and patient outcomes. However, the data are complex, and their analysis is not trivial. Treating the care pathway as an operational process allows patient flow and service use to be analyzed using process-centered methods such as process mining.

Table 1. The stepped care model for common mental health disorders, revised from the National Institute for Health and Care Excellence guidance (CG123) and the Improving Access to Psychological Therapies manual [10,11].

Step	Examples of conditions	Examples of interventions
1	Presentations of all known and suspected common mental health disorders	Identification, assessment, psychoeducation, active monitoring, and referral for further assessment and interventions
2	Depression (either mild to moderate or persistent subthreshold symptoms), generalized anxiety disorder, panic disorder, and obsessive compulsive disorder	Self-help and guided self-help based on CBT[a], computerized CBT, psychoeducational groups, behavioral activation, and structured group physical activity program
3	Depression (either moderate or severe, prevention of relapse, or mild to moderate where individuals have not responded to low-intensity interventions), moderate to severe panic disorder, obsessive compulsive disorder with moderate or severe functional impairment, posttraumatic stress disorder, and generalized anxiety disorder, with marked functional impairment or in which individuals have not responded to low-intensity interventions	CBT (individual or group), applied relaxation, mindfulness-based cognitive therapy, interpersonal psychotherapy, behavioral activation, couple therapy, counseling for depression, brief psychodynamic therapy, trauma-focused CBT, EMDR[b], and graded exercise therapy

[a]CBT: cognitive behavioral therapy.

[b]EMDR: eye movement desensitization and reprocessing.

Process Mining in Health Care

While traditional business process modeling involves reaching a consensus about model design, process mining is based on the assumption that a process model can be extracted from the information available in systems [17]. Such information can be used to generate "event logs," the ordered records of events corresponding to the activities encountered by entities that have traveled through a process. An active area of research within process mining is regarding process discovery algorithms, which are often referred to as "process miners" [17]. These algorithms use the information contained within an event log to create a process model that is representative of the behavior captured within the log [18].

In health care, process mining techniques can be used to assess conformance to clinical guidelines and protocols and to assess performance [17]. Frequently asked questions within process mining projects are as follows: "What are the most followed paths and what exceptional paths are followed?" "Are there differences in care paths followed by different patient groups?" "Do we comply with internal and external guidelines?" and "Where are the bottlenecks in the process?" [19].

While process mining has been frequently applied in health care [17,20], in applications such as disease trajectory modeling and clinical pathway analysis [21], research has highlighted that there has been limited uptake within health care organizations, apart from specific research case studies [21,22]. Endeavors are being made toward systematic adoption of process mining in the health care domain through initiatives such as the Process Mining for Healthcare manifesto [21]; however, evidence shows an absence of efforts to integrate process mining tools into systems that record patient data [23].

Recent reviews of process mining in health care [20] and data-driven care pathway mapping from care records [24] have found applications in a number of medical fields, most commonly in oncology; however, neither found evidence of application in the domain of mental health. Outside the field of process mining, the study by Richards et al [12] in 2012 analyzed the delivery of stepped care within mental health services in the United Kingdom; however, there is no evidence of such analysis being developed into a reproducible methodology or tool that could be applied to routinely collected data. Furthermore, a review of the wider literature shows no evidence of the application of process mining to psychological therapy care pathways. The Scopus query "process mining" AND ("psychological therap*" OR "psychological intervention" OR "psychological treatment") returned only 2 results. Neither of the returned studies explored the care pathway for psychological therapies using process mining. The first study used process mining to explore health care pathways for patients presenting to emergency departments with functional neurological disorders, where onward referral to psychological therapy formed part of the pathway [25]. The second study used data from a psychological therapies service but did so to explore transitions between pre- and posttherapy clinical outcome bands to investigate the impact of appointment attendance on patient outcomes [26].

Objectives

The application of process mining techniques to NHS Talking Therapies care pathways using EHR data has not been studied; therefore, this study intends to demonstrate the value and utility of doing so. Our study applies process mining to the local care pathways of two NHS Talking Therapies sites and shows how the use, efficiency, and effectiveness of the Talking Therapies care pathway can be explored using a process mining approach. As part of a Knowledge Transfer Partnership project focused on embedding innovation into health care software, our study aims to address the absence of a systematic uptake of process mining by demonstrating how access to analytical tools that allow stakeholders to explore characteristics of their service implementation will enable improved monitoring of system use and better quality analysis into how health systems can affect patient-level outcomes.

Methods

Ethics Considerations

The Knowledge Transfer Partnership project titled "Developing innovative, advanced analytical tools to help improve IAPT demand and capacity planning" was approved by the HRA (Health Research Authority) and Health and Care Research Wales (HCRW; integrated research application system project ID 320525), and the University of Bath Psychology Research Ethics Committee (23-031).

Process Mining

Process mining offers a data-driven approach to analyzing health care processes using the data stored within health information systems in event logs. There are 3 main areas of process mining: process discovery, which involves discovering a process model; conformance, which involves checking an existing process model against an event log; and enhancement, which involves enhancing a model either by repairing incorrect aspects of the model or by extending a model to add a new perspective, including adding extra information such as frequencies, timings, and bottlenecks [18]. In our analysis of Talking Therapies EHR data, we treated the Talking Therapies care pathway as a process, where care pathway stages form the activities in the process, while referred patients (termed "referrals" within IAPT services) are the entities who travel through the process. We demonstrated how process mining can be used for process discovery, evaluation of conformance with system design principles, and performance analysis (enhancement through extension) of the care pathway from referral to discharge.

Throughout our study, the terms "care pathway" and "pathway" are used to describe the whole process of interest, whereas unique sequences of activities are termed "routes" through the care pathway. In the field of process mining, these routes are often referred to as traces or variants.

Research has highlighted the importance of domain expert involvement in process mining projects [21,27,28]. This study is part of a collaborative project with Mayden, the company that provides the "iaptus" digital care record software to NHS Talking Therapies service providers; therefore, expert involvement has been integrated throughout this study. System experts from Mayden work with relevant stakeholders within all Talking Therapies sites using the software to configure each care pathway within the software, ensuring the suitability of the data collected. The specific data for this project were extracted by ETL (extract, transform, and load) experts at Mayden, and data preparation was conducted in consultation with data analysts at Mayden, who are familiar with the data structure and any data quality issues that are universal across sites. The choice of sites for this work was guided by recommendations from service representatives. The approach and results have been reviewed by a user representative from Mayden; furthermore, feedback on this approach was also gathered from Talking Therapies service representatives through user engagement meetings and workshops.

We transformed EHR data into an event log format, manipulated event logs using filtering and stage aggregation, and produced a process map directly from event logs in the form of a directly-follows graph: a descriptive mapping of data to a directed graph of nodes (care pathway stages) and edges (patient flows between stages) in R software (version 4.1.2; R Foundation for Statistical Computing) using the *bupaR* package (version 0.5.2) [29]. Additional R code was developed to extend the process maps with summary statistics and additional formatting, including a bottleneck indicator that uses colors to identify where the edge between the two stages was both highly traveled and had a lengthy median duration. Bright colors are assigned to the edges that fall into the highest percentile when this information is collected into a single metric (the product of the number of patients who moved between the two stages and their median wait duration). A visualization of common routes through the care pathway was built using the *ggplot2* R package (version 3.3.5) [30].

Data

Data Source and Inclusion Criteria

iaptus is the digital care record used by approximately two-thirds of NHS Talking Therapies service providers in England [31]. In this study, we analyzed anonymized EHRs from iaptus relating to the patients referred to two sites between June 1, 2019, and June 1, 2021. These two sites were selected due to their differing size and clinical performance so that the generalizability of the approach could be demonstrated. From the patients referred to the sites during this period, those who consented to their data being processed as part of the IAPT data set, those who were aged ≥18 years at the time of referral, and those who had been discharged by February 8, 2023, were included in study data.

Types of Data Used

The NHS target for recovery is that 50% of those who have completed treatment should recover [11]. Recovery is calculated using the notion of clinical "caseness," which is based on threshold levels of patient reported outcome measures. Patients are considered recovered if they were above the caseness threshold at the beginning of treatment and below the threshold at the end of treatment [11]. Patient reported outcome measures are routinely collected at each session; therefore, patient outcomes for the program have a high degree of data completeness for those who attend ≥2 treatment sessions. Recovery outcome data were available for 94.98% (17,151/18,058) and 96.44% (5034/5220) of the patients who had attended ≥2 sessions at site 1 and site 2, respectively.

Patient geographical data, such as lower super output area (LSOA), was joined with 2019 Office for National Statistics (ONS) data [32] to calculate the proportion of referrals to each site that were assigned to each of the index of multiple deprivation (IMD) deciles. IMD data were available for 99.54% (45,193/45,401) and 99.78% (12,561/12,590) of the patients with referrals to site 1 and site 2, respectively.

The therapist role was available for 71% (375/528) and 82.7% (105/127) of the therapists in the data set for site 1 and site 2, respectively. Role data were mostly missing for the remaining therapists because their role data had been deleted and was no longer available in the system. These therapists had most likely left the service.

Patient appointment records contain information about appointment attendance, which is used to calculate the proportion of patients who completed treatment by attending ≥2 treatment sessions and the proportion of sessions that were not attended. These data were 99.62% (206,564/207,357) and 99.89% (55,415/55,478) complete across site 1 and site 2, respectively. Missing data were removed before calculating the recovery rates, the IMD decile breakdown, the therapist role proportions and missed appointment rates.

Patient movements through the stages of the Talking Therapies care pathway are recorded in the system by service staff as time-stamped events. Each row of data in the event log identifies the referral that the movement relates to, the stage in the pathway that the patient has moved into, and the date and time of the movement. The pathway stages include all aspects of the service's implementation of the stepped care model, from receipt of referral to assessment, low- (step 2) and high-intensity (step 3) interventions, and the eventual discharge of the patient. Some Talking Therapies services may provide other specialist interventions in addition to step 2 and step 3 treatment; therefore, data were filtered to only include patients with referrals for NHS Talking Therapies treatment.

Data Preparation

The event log data needed to undergo data quality assessment and be prepared for analysis. In the analysis, the time stamp of events was used to determine their sequence. However, as the time stamp is recorded manually by staff, there were some issues with the event data, which meant that manually inputted time stamps did not always reflect the actual sequence of events that occurred. Nevertheless, this was imputed from other time stamp data within the movement record, using the time stamps of surrounding events in the sequence, for example, by taking the last observation carried forward. More information on this can be found in Multimedia Appendix 1.

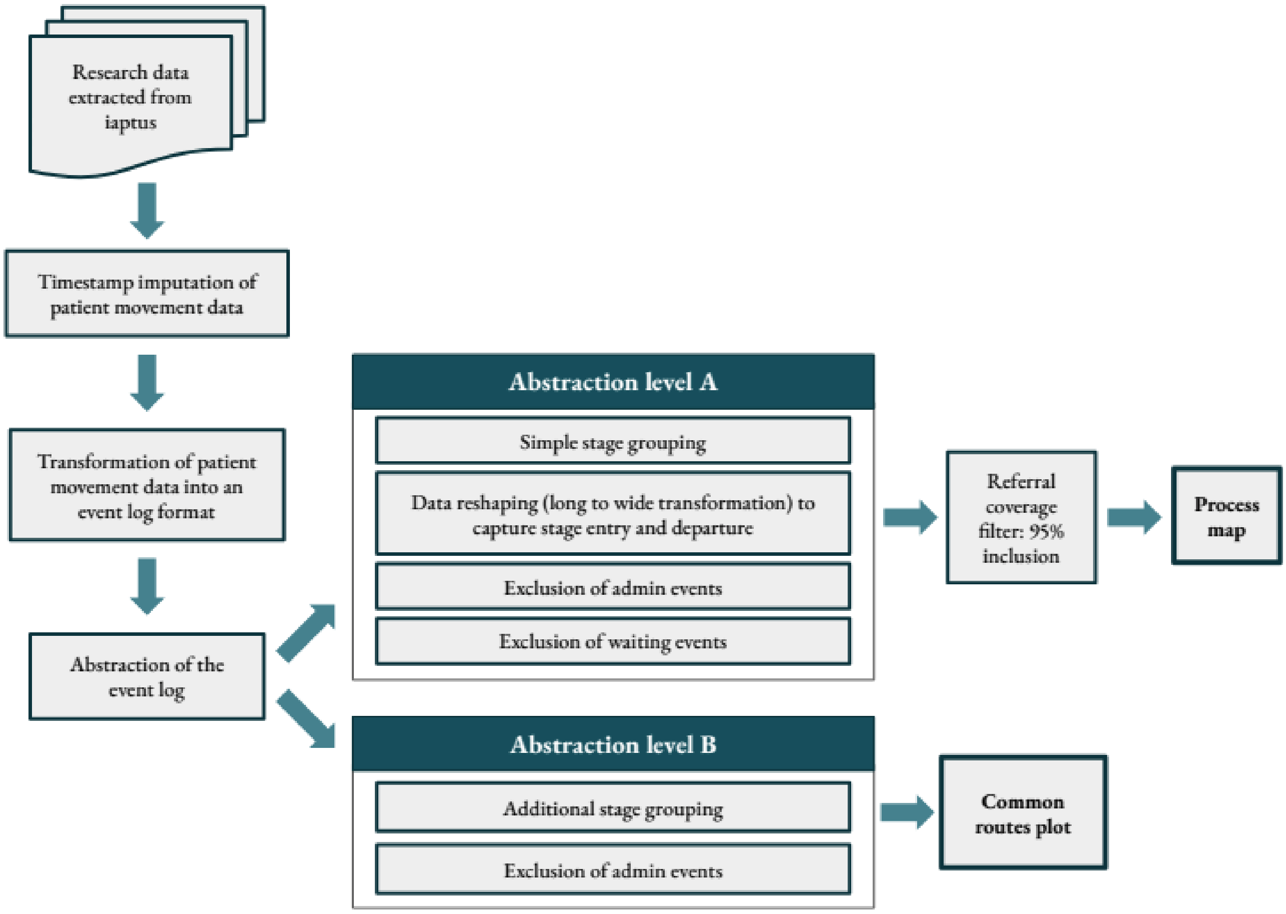
Care pathway configurations can be complex as the pathway is often used within services for patient management and data reporting purposes. In addition, the configuration and the resulting data collected can change over time. For these reasons, a high degree of variation was found in the basic event log; therefore, activities were aggregated and filtered to produce two abstracted event logs for each site that would provide two simplified views of the data set (as shown in Figure 1). These two levels of abstraction were designed in collaboration with data analysts at Mayden.

The abstraction level A involved grouping care pathway stages and excluding events. Grouping care pathway stages included remapping the names of duplicated stages with minor distinctions to appropriate descriptive names; for example, outdated stage names were replaced with newer versions of the name. Other suitable stages were collapsed into overarching subprocesses, whereby consecutive instances of events within the same subprocess are transformed into a single instance of the subprocess. For example, a discharge planning stage was always followed by the final stage representing the discharge of the patient; therefore, these two stages were collapsed into a single "discharged" stage to reduce granularity in the data.

To further reduce the granularity of the data presented in the process map visualizations, events relating to administrative stages, such as moving to a different step intensity or joining a waiting list, were excluded from the event log, as this information was contained implicitly within the data. To produce the process maps, the event log was filtered using a percentage coverage level to include data relating to patients who experienced the most common pathway routes.

For abstraction level B, stage names were remapped in the same way as abstraction level A, and subprocesses were collapsed to a further degree to create larger groups that encompassed more pathway stages. For example, the subprocess "step 3 treatment" included the individual treatment stages at step 3, such as cognitive behavioral therapy (CBT) and counseling. Waiting stage events were not excluded from this version of the event log, as these additional event data proved useful for analyzing distinct pathway routes; however, events relating to other administrative stages were excluded.

Figure 1. Data preparation steps.



Results

Data Description

The two sites in our sample differed in size, performance, and referral patient characteristics (Table 2). Site 1 had >3.5 times referrals as site 2. Both sites had similar proportions of referred patients in the lowest IMD deciles; however, site 2 had considerably more referred patients in decile 3 and marginally more patients in deciles 9 and 10, while site 1 had more patients in deciles 5, 7, and 8. On the whole, this suggests that patients referred to site 2 had more variation in deprivation status, whereas site 1 saw a more even distribution across the IMD deciles. Site 2 had a slightly greater proportion of higher-intensity therapists than site 1. While site 1 fell short of the recovery rate target by more than 2%, site 2 exceeded the target by >2%. Site 2 had a higher treatment completion rate (the proportion of referred patients who had ≥2 treatment sessions) and shorter waiting times than site 1 and a higher missed appointment rate. Although site 1 had a lower treatment completion rate, patients with referrals to the site had a marginally higher mean number of treatment sessions than at site 2, suggesting that those who did have treatment had more treatment sessions on average at site 1. The patients with referrals to site 1 were younger, more likely to be female patients, more likely to have an unspecified presenting problem (as opposed to an initial diagnosis), more likely to be self-referred, and less likely to have had a previous referral, in comparison to those with referrals to site 2.

Imputation summary statistics in Table 3 describe the time stamp adjustments for events that exhibited illogical time stamp sequencing. In total, 1.9% (6491/336,637) of the events were adjusted at site 1, with a median adjustment size (the absolute difference between the original value and the imputed value) of 2.3 weeks. In total, 7.2% (8921/123,523) of the events were adjusted at site 2, with a smaller median adjustment size of 0.5 weeks. Across all events (including those with no adjustment), the adjustments averaged very close to 0 at both sites (site 1: mean 0.2, SD 2.3; median 0, IQR 0-0 and site 2: mean 0.1, SD 1.7; median 0, IQR 0-0), suggesting that the impact on the overall results was likely to be minimal. Further information is provided in Multimedia Appendix 1. Furthermore, Table 3 shows the reduction in event log variation from applying abstraction to the log.

Table 2. Data summary for the site characteristics, performance measures, and patient referral characteristics present in the data set for discharged patients with referrals received by sites 1 and 2 between June 1, 2019, and June 1, 2021.

	Site 1	Site 2
Site characteristics		
Patient referrals received, N	45,401	12,590
Index of multiple deprivation deciles of referrals[a], n/N (%)		
1	4783/45,193 (10.58)	1263/12,561 (10.05)
2	5041/45,193 (11.15)	1335/12,561 (10.63)
3	4207/45,193 (9.31)	2083/12,561 (16.58)
4	4937/45,193 (10.92)	1361/12,561 (10.84)
5	4077/45,193 (9.02)	787/12,561 (6.27)
6	4091/45,193 (9.05)	1235/12,561 (9.83)
7	5502/45,193 (12.17)	1084/12,561 (8.63)
8	3980/45,193 (8.81)	682/12,561 (5.43)
9	3707/45,193 (8.20)	1194/12,561 (9.51)
10	4868/45,193 (10.77)	1537/12,561 (12.24)
Therapist role, n/N (%)		
High-intensity therapists	201/375 (53.6)	59/105 (56.2)
Low-intensity therapists	174/375 (46.4)	46/105 (43.8)
Site performance indicators, n/N (%)		
Treatment completion rate (of all patient referrals)	18,058/45,401 (39.77)	5220/12,590 (41.46)
Recovery rate (of patient referrals who completed treatment)	8212/17,151 (47.88)	2632/5034 (52.28)
Missed appointment rate (of all scheduled appointments for all patient referrals)	21,757/206,564 (10.53)	6524/55,415 (11.77)
Summary statistics for patient referrals, mean (SD); median (IQR)		
Age (years)	35 (14); 31 (25-42)	38 (15); 35 (26-48)
Referral duration (weeks)	23 (23.6);13.9 (4.1-36.9)	18.3 (16.8); 14.7 (2.7-30)
Total waiting time duration (weeks)	17.1 (19.2); 8.6 (3.1-23.6)	12.4 (12.8); 7.7 (1-22)
Number of treatment sessions (all patient referrals)	3.5 (4.9); 1 (0-6)	3.1 (4.1); 1 (1-5)
Number of treatment sessions (patient referrals who completed treatment)	8 (5.1); 7 (5-10)	6.6 (4.4); 6 (3-9)
Categorical data for all patient referrals, n/N (%)		
Gender identity		
Female patients (including trans women)	30,257/45,401 (66.64)	8179/12,590 (64.96)
Male patients (including trans men)	14,934/45,401 (32.89)	4399/12,590 (34.94)
Nonbinary patients	192/45,401 (0.42)	6/12,590 (0.05)
Unspecified	18/45,401 (0.04)	6/12,590 (0.05)
Presenting problem		
Anxiety and stress-related disorders	14,477/45,401 (31.89)	4473/12,590 (35.53)
Depression	12,100/45,401 (26.65)	4342/12,590 (34.49)
Other mental health problems	1256/45,401 (2.77)	303/12,590 (2.41)
Other recorded problems	142/45,401 (0.31)	83/12,590 (0.66)
Unspecified	17,426/45,401 (38.38)	3389/12,590 (26.92)
Number of previous National Health Service Talking Therapies program referrals		
No previous referrals	38,456/45,401 (84.70)	10,097/12,590 (80.20)
1 previous referral	5099/45,401 (11.23)	1874/12,590 (14.88)

	Site 1	Site 2
≥2 previous referrals	1846/45,401 (4.07)	619/12,590 (4.92)
Referral source		
Self	42,901/45,401 (94.49)	9853/12,590 (78.26)
General practitioner	1432/45,401 (3.15)	867/12,590 (6.89)
Other	1068/45,401 (2.35)	1870/12,590 (14.85)

[a]Where 1 represents the most deprived 10% of lower super output areas.

Table 3. Data preparation summary: time stamp imputation and event log abstraction.

Data preparation summary	Site 1			Site 2		
	Values, n	Values, mean (SD)	Values, median (IQR)	Values, n	Values, mean (SD)	Values, median (IQR)
Imputation summary statistics						
Size of time stamp adjustment (weeks)						
Total per referral for all referrals	45,401	1.2 (7.3)	0 (0-0)	12,590	1.5 (7.1)	0 (0-0.1)
Total per referral for adjusted referrals only	3218	17.4 (21.9)	11 (0.4-25.9)	3489	5.3 (12.8)	1.4 (0.5-3.5)
Per adjustment for all events	336,637	0.2 (2.3)	0 (0-0)	123,523	0.1 (1.7)	0 (0-0)
Per adjustment for adjusted events only	6491	8.6 (14.3)	2.3 (0.6-10.2)	8921	2.1 (5.8)	0.5 (0.2-1)
Number of time stamp adjustments						
Per referral—for all referrals	45,401	0.1 (0.6)	0 (0-0)	12,590	0.7 (1.5)	0 (0-1)
Per referral— for adjusted referrals only	3218	2 (1.2)	2 (1-3)	3489	2.6 (1.4)	2 (2-3)
Event log abstraction summary						
Raw event log						
Number of unique pathway stages	189	—[a]	—	117	—	—
Number of unique pathway routes	2454	—	—	1388	—	—
Processed event log—abstraction level A						
Number of unique pathway stages	26	—	—	23	—	—
Number of unique pathway routes	388	—	—	268	—	—
Processed event log—abstraction level B						
Number of unique pathway stages	8	—	—	9	—	—
Number of unique pathway routes	239	—	—	67	—	—

[a]Not applicable.

Process Mining

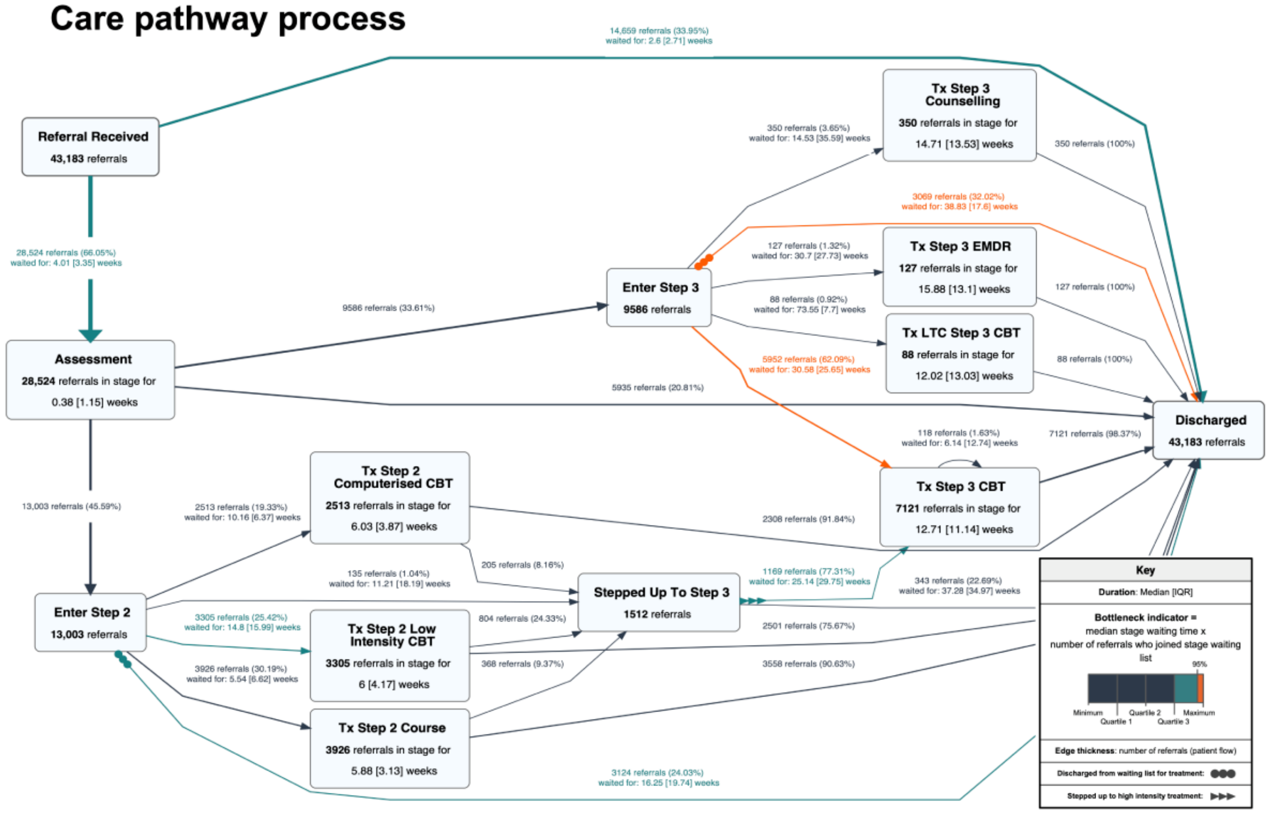
Process Mapping

Figure 2 presents a process map with 95% coverage for site 1. The map summarizes the pathway structure, patient flow, and service performance for referred patients who followed common pathway routes. Patients with referrals to the service receive an assessment before being triaged to step 2 (low-intensity treatment) or step 3 (high-intensity treatment). Patients can be "stepped up" from low- to high-intensity treatment (indicated by a triple arrowhead on the edge tail, as shown in the process map key in Figure 2) and can be discharged from any point in the pathway. A process map for site 2 is presented in Figure S1 in Multimedia Appendix 1.

Edge thickness, referral volumes, and branching probabilities demonstrate patient flow. In the process map, one-third (14,659/43,183, 33.95%) of the referred patients were discharged immediately, and a further 20.81% (5935/28,524) of those assessed were discharged after assessment, equating to 13.74% (5935/43,183) of all referrals represented in the map. In total, 27.42% (6193/22,589) of those triaged into steps 2 or 3 were discharged before receiving the intended treatment (3124/13,003, 24.03% of the referrals at step 2 and 3069/9586, 32.02% of the referrals at step 3), equating to 14.34% (6193/43,183) of all the referrals represented in the process map. Flows that involved patients being discharged before receiving the intended treatment are indicated on the map with a set of triple dots on the edge tail (as shown in the process map key in Figure 2). Site 1 had higher attrition rates from waiting lists for treatment than site 2, and at both sites, the attrition rate was higher at step 3. Relatively few patients followed the “step-up” route from low- to high-intensity treatment.

Edge color and duration summary statistics quantify pathway performance for 95.11% (43,183/45,401) of referrals. For example, the initial waiting time for assessment had a median value of 4 weeks, followed by a secondary waiting time for treatment with a median value ranging from 5.5 to 73.6 weeks across treatment types. The duration on each node summarizes the length of treatment (number of weeks) and shows that higher-intensity treatment generally had a longer duration than lower-intensity treatment. The bottleneck indicator highlights the waiting time for step 3 CBT and the waiting time experienced by those who waited for step 3 treatment but were discharged.

Figure 2. Process map of the care pathway at site 1 using event log A (n=43,183 patient referrals). Referral coverage level=95%. CBT: cognitive behavioral therapy; EMDR: eye movement desensitization and reprocessing; Tx: treatment.



Common Route Analysis

Figure 3 presents the 10 most common care pathway routes at site 1 and the referral outcomes for each route, using event log B. The remaining uncommon routes are summarized under the common routes using the same metrics. The x-axis represents the median duration of activities throughout the course of the referral. Furthermore, a common route analysis is presented for site 2 in Figure 4. The findings from Figure 3 correspond with those from Figure 2 for site 1: three of the 5 most common routes involved no treatment, and routes that involved discharge from waiting lists are also observed. The relative frequencies of common treatment routes show that step 2 treatment alone was more common than step 3 treatment alone, and the comparative infrequency of “stepped care” is evident. Over the common routes at both sites, >5 times as many patients experienced direct access to high-intensity treatment rather than stepped care. In addition, many common routes involved no treatment at site 2. A smaller proportion of patients were discharged from the waiting list for assessment (1442/12,590, 11.45%) in comparison to site 1 (8437/45,401, 18.58%); however, a greater proportion of patients were discharged after being assessed at site 2 (3035/12,590, 24.11%) compared to site 1 (5970/45,401, 13.15%). A total of 1507 of the patients with referrals to site 1 traveled a stepped care pathway route (1117 patients who experienced a common stepped care route + 390 patients who experienced step-up through less common routes), equating to 3.32% (1507/45,401) of all patients referred to the service. At site 2, the equivalent figure was 527 patients (365 patients who experienced a common stepped care route + 162 patients who experienced step-up through less common routes), equating to 4.19% (527/12,590) of all patients referred to the service.

The duration of common pathway routes differed between sites. Longer median waiting times for common assessment and treatment pathway routes were observed for site 1, including the stepped care route. At both sites, the common high-intensity route had a median wait duration of 2.6 times longer than that of the low-intensity route. There was a distinction at both sites between the median wait durations of routes involving discharge from a waiting list, compared to routes where patients completed waiting and commenced treatment. For example, at site 1, the median value of the total wait duration was 13.3 weeks for the step 2 treatment route compared to 20.6 weeks for the attrition alternative. The step 3 treatment route had a total wait duration with a median value of 34 weeks compared to 41.9 weeks for the attrition alternative. At site 2, the equivalent comparisons were 8.1 versus 13.1 weeks at step 2 and 21 versus 26.8 weeks at step 3. This distinction was most severe for the stepped care route at site 1, where those who were stepped up from step 2 but were discharged from the high-intensity waiting list experienced a total wait duration with a median value of 11.4 weeks longer than those who successfully waited for treatment.

Figure 3. Common route analysis of the care pathway at site 1 using event log B (n=45,401 patient referrals). Coverage level=100%. Top 10 routes only plotted. Tx: treatment; WL: waiting list.

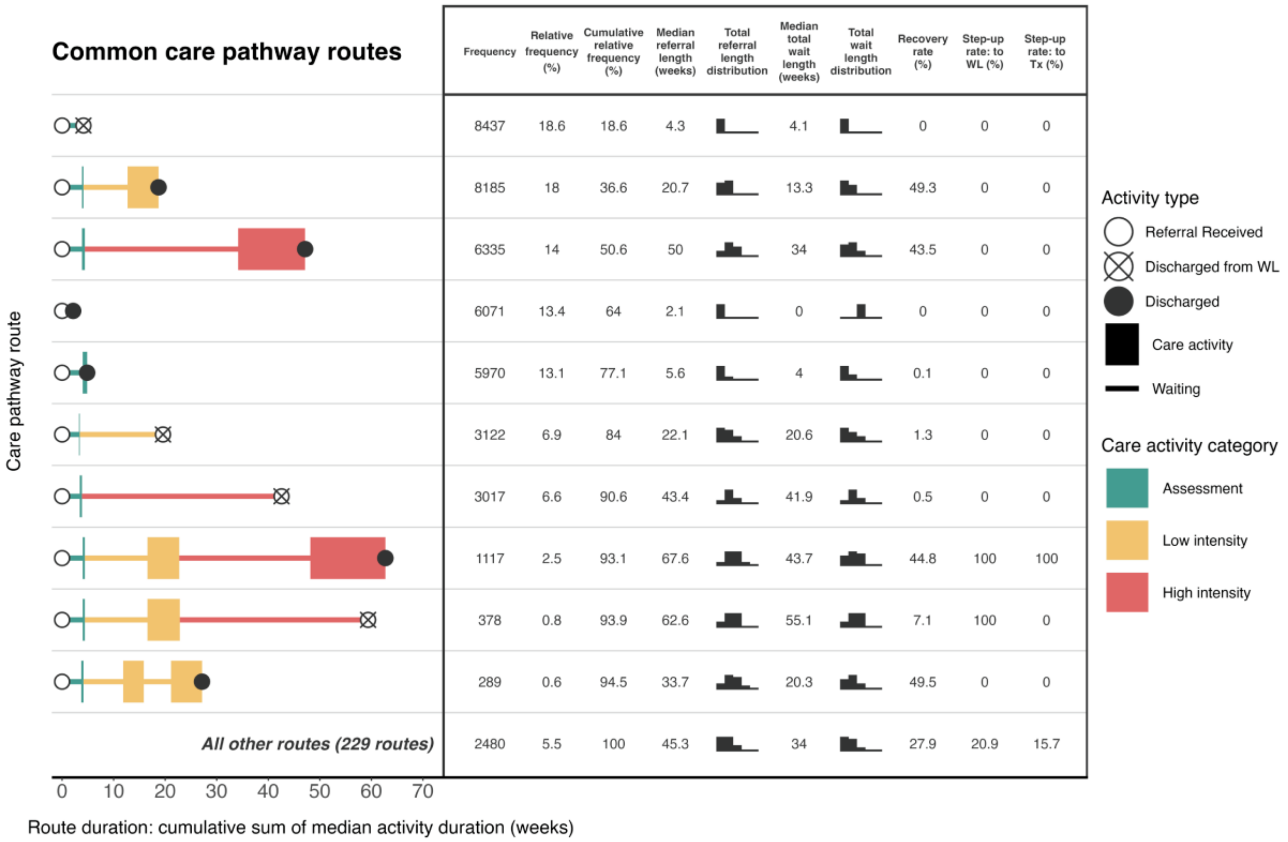
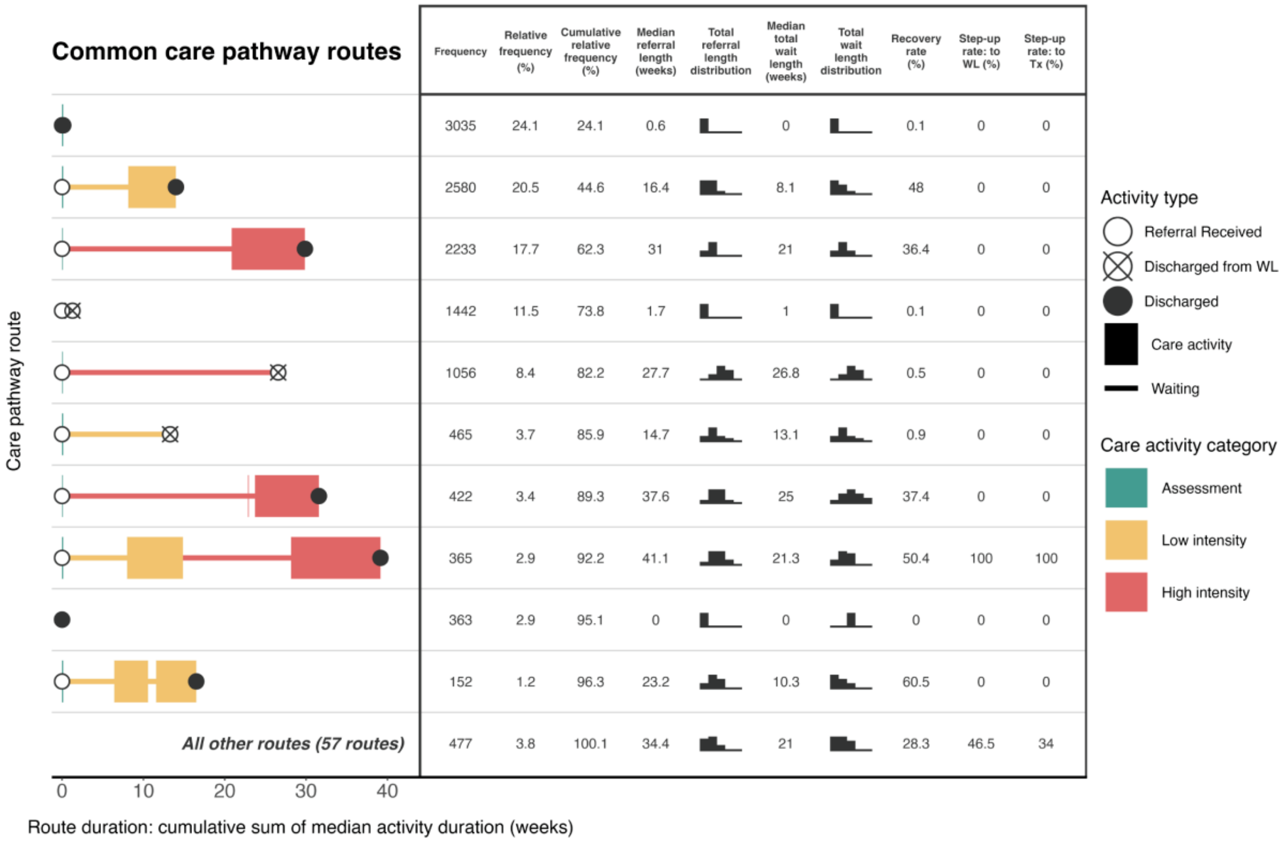


Figure 4. Common route analysis of the care pathway at site 2 using event log B (n=12,590 patient referrals). Coverage level=100%. Top 10 routes only plotted. Tx: treatment; WL: waiting list.



Discussion

In this paper, we demonstrated how process mining techniques, such as process discovery using directly-follows graphs (data-driven process maps), and process enhancement (through extension) using performance and common route analysis can be used to explore "as-is" mental health care pathway use and patient outcomes from a process perspective using routinely collected anonymized EHRs.

Clinical Guidelines and Outcomes

Our results show how process mining can be used to investigate care system implementation and explore adherence to clinical guidelines and the principles that have informed the design of the care pathway of a service. We presented a data-driven process map of the referrals to a Talking Therapies service over 2 years. The map indicated lower rates of patients being stepped up from low- to high-intensity treatments than might be expected from a "stepped care" system. The common route analysis confirmed that approximately 3% to 4% (site 1: 1507/45,401 and site 2: 527/12,590, respectively) of patients referred to the sites received stepped care.

The stepped care approach is associated with better patient outcomes [8,9], but research has found differences in the implementation of "stepped care" within psychological therapy services [12], for example, a stratified model, where a specific treatment intensity is selected after an initial assessment, versus a progressive model, where low-intensity treatment takes place in the first instance [13]. Therefore, data-driven pathway mapping is a fitting tool to investigate not only the design and implementation decisions of the care system but also the actual use of the system; for example, the National Institute for Health and Care Excellence guidelines instruct that pathways should "allow services to be built around the pathway and not the pathway around the services" [10].

The low stepped care rates presented in this study could be attributed to a number of factors, which could be further explored using process-centered methods. For example, Talking Therapies services are rewarded based on assessment volumes and outcome-based performance [33]; therefore, services can be motivated to process stepped up referrals as separate referrals entirely. Patient-level instances of care, as opposed to referral-level ones, could be used to investigate rereferral pathway routes to explore whether this type of behavior is present in the data. In addition, step-up rates will be impacted by the clinical composition of the referrals received by the service; therefore, by filtering the analysis by individual patient groups, adherence to clinical guidelines and operational principles could be explored further. The relative infrequency of stepped care in comparison to direct access to high-intensity treatment supports the findings of previous research into stepped care implementation [12].

Furthermore, we identified common routes through the Talking Therapies care pathway and presented the clinical performance of these routes in terms of patient recovery. This type of analysis could be used to compare clinical outcomes between pathway routes and could provide a valuable basis for future evaluation of the efficacy of treatment pathways.

Patient Attrition

Our results established common routes that did not involve treatment. Service managers may wish to monitor routes involving discharge from waiting lists for care. Waiting list attrition is likely to be an undesirable outcome that indicates system inefficiency as well as potentially leading to negative patient outcomes due to untreated symptoms. Analytical tools that incorporate pathway routes can be used to locate patient attrition and monitor patient flows through undesirable routes. For example, although the two sites had similar treatment completion rates of approximately 40% (in line with national rates), route analysis revealed differences in the dominant nontreatment routes between the sites. The distinction between the location of early attrition can help differentiate between that which is patient initiated and that which is attributable to triaging or signposting onward.

Furthermore, we identified attrition later in the course of care, provoking questions regarding the association between system performance and patient attrition. At both sites, the pathway routes involving discharge from a treatment waiting list featured in the most common routes. Moreover, patients who were discharged before receipt of treatment appeared to wait for a longer duration than those who received treatment when comparing the median of the total wait duration of each route. Those discharged before receipt of treatment at site 1 waited 7.3 weeks longer at step 2 and 7.9 weeks longer at step 3 compared to those who received treatment. At site 2, those discharged before receipt of treatment waited 5 weeks longer at step 2 and 5.8 weeks longer at step 3.

In addition, areas of the pathway with longer median waiting times (ie, high-intensity treatment) had higher rates of waiting list attrition. Furthermore, site 2 had shorter median waiting times for treatment and less attrition from the treatment waiting lists than site 1. These initial findings could indicate a relationship between treatment waiting times and patient discharge before treatment commencement. Owing to the limitations of publicly available data, the research literature only seemed to include engagement studies that explored the association between waiting times and appointment nonattendance [15,16] rather than any quantitative study that has explored the relationship between secondary waiting times and patient attrition from the Talking Therapies program. However, our initial results about the relationship between treatment waiting times and patient disengagement are consistent with these studies. This relationship will be explored further in future work, as identifying the factors associated with waiting list attrition could have significant implications for policy.

Waiting Times and Pathway Bottlenecks

The Talking Therapies program's waiting times are subject to NHS service standards; however, the initial waiting time for a first appointment is often followed by a secondary waiting period, which, more often than not, is 3 times as long as the initial waiting time [14]. Secondary waiting times are now incorporated into national reporting requirements for Talking Therapies services; however, they are not held to NHS service standards, despite the IAPT manual declaring that they should not be "excessive" [11]. Furthermore, the manual states that the waiting time for high-intensity treatment should not be "substantially longer" than the waiting time for low-intensity treatment and that for those who are stepped up, the waiting time between the low-intensity treatment and high-intensity treatment should "certainly not exceed the waiting time standard for the first intervention" [11].

Accordingly, our analysis provides an overview of system performance in terms of the total wait duration of common pathway routes and further demonstrates patient waiting times throughout the course of care. In addition, the process map bottleneck indicator has been used to highlight areas of the pathway that involve both large patient flows and lengthy median waiting times, indicating pathway stages that might need more urgent attention, and could be used by decision makers to inform capacity allocation or pathway configuration decisions.

Existing studies into the clinical impact of waiting times appear to focus on first waiting times within the pathway. For example, Clark et al [9] used the time between the first referral date and the first treatment appointment as a feature in their predictive models of clinical outcomes. However, as the vast majority of first IAPT assessment appointments are categorized as involving some aspects of treatment [15], it is likely that the first recorded treatment appointment will be the assessment, which will be conducted before the commencement of a course of treatment, leading to potential underestimation of the time taken to enter treatment.

While secondary waiting times are now part of national reporting requirements for Talking Therapies services, these figures are only reported for those who eventually go on to receive a second treatment session; therefore, by definition, they did not drop out before the second treatment session. For this reason, exploration of the impact of these waiting times on patient dropout requires more advanced tools. Analytical tools that allow Talking Therapies services to monitor both initial and secondary waiting times will, therefore, enable better monitoring of excessive waiting times for patients and offer the potential for future research into the relationship between all waiting times and patient outcomes.

Limitations

Our results are subject to some limitations in relation to the data and preprocessing methods. To create the study data set, referrals in the available data were filtered using both the referral date and the discharge date. The patients with referrals received by the sites within the inclusion window (from June 1, 2019, to June 1, 2021) were excluded from the study sample if they were discharged after the discharge cut-off date (February 8, 2023). This filter could introduce bias toward the end of the referral inclusion window by excluding the patients with referrals with a longer duration; however, the number of patients with referrals received within the referral inclusion window who were not discharged on or before February 8, 2023, was minimal at both sites (292/45,401, 0.64% and 1/12,590, 0.008% of all patients referred to sites 1 and 2, respectively). Therefore, the impact on the overall results is considered to be negligible.

Second, the 2-year period of referral inclusion included the onset of the COVID-19 pandemic, which has been shown to have influenced the rates of access and methods of treatment delivery [5,34]. Some additional analysis has been included in Multimedia Appendix 1 (Figures S2-S8 and Table S1) to evaluate the impact of the pandemic on the results of this study. The volume of referrals to both sites dropped during the initial months of the COVID-19 pandemic and recovered in the months that followed, similar to the early findings presented in the study by Bauer-Staeb et al [34]. There were some differences in performance measures following the onset of the COVID-19 pandemic (Table S1 in Multimedia Appendix 1), such as an increase in recovery rates, a reduction in referral duration and wait duration, a reduction in the missed appointment rate, and changes in the proportion of patients who received ≥2 treatment sessions. These differences can be further explored using the approach presented in this study (Figures S5-S8 in Multimedia Appendix 1) by comparing the frequency and duration of pathway routes across the two time periods. Routes that involved discharge from a waiting list for treatment were less common following the onset of the COVID-19 pandemic. The overall rates of stepped care had some differences following the onset of the COVID-19 pandemic; however, the rates remained low overall, with <6% of all referred patients receiving stepped care across both sites, both before and following the onset of the COVID-19 pandemic.

Third, the imputation techniques applied to illogical event time stamps were based on the assumption that other events' time stamps could be used as a proxy. Such assumptions were based on our understanding of how iaptus users use the software. In addition, as part of the ongoing development and implementation of this study, elements of data processing that were constructed independently for the two sites, such as the design of the abstraction levels, are being developed into automated criteria that can be applied across all sites. More structured feedback will be gathered from multiple services about the suitability of these assumptions as part of this implementation process during planned future user engagement sessions.

Clinical Implications

Our study provides a contextual yet structured view of patient flow rather than using isolated metrics to describe pathway use. This study has been driven by the ongoing demand received by Mayden from NHS Talking Therapies services for ways to monitor access to services, explore changes in clinical outcomes, understand patient engagement, and manage service capacity. Ongoing elements of this project include developing the analysis presented in this study into tools that can be used for these purposes across all services that use iaptus, enabling purely data-driven exploration of care pathways from the data that are routinely collected by the services.

Feedback from Talking Therapies service representatives such as service managers, clinical leads, and data analysts through meetings; a webinar with representatives from 33 Talking Therapies services; and user workshops has highlighted that implementing such analysis into routine practice will enable key stakeholders within mental health services to analyze their implementation of the Talking Therapies treatment model, by monitoring actual system use and performance and by exploring how these might impact patient-level outcomes. Further implications proposed by service representatives at these sessions were a better understanding of patient outcomes and engagement, the potential to use this approach to analyze patient movements through the care pathway by their demographic or clinical backgrounds, and the ability to analyze repeated referrals. Service feedback has also suggested that understanding pathway use will support data-driven capacity allocation decisions; therefore, our future research also endeavors to integrate process insights with staff resource information to support such decision-making.

Analysis of patient journeys through pathway routes demonstrates how service users experience the care pathway and is, therefore, naturally patient centered. Integrating this information into an EHR would provide clinicians with an immediate overview of their patients' previous routes and associated wait times, providing patient-level and service-level insights.

Furthermore, identifying secondary waiting times within the care pathway is a key first step to exploring the relationship between waiting times and engagement outcomes. In addition, our future research aims to model this relationship, enabling services to target improvements to relevant elements of their performance.

Acknowledgments

This study forms part of a Knowledge Transfer Partnership (KTP) project between the University of Bath and Mayden, a software development company that provides "iaptus," the digital care record for psychological therapies used by the National Health Service and other providers of the National Health Service Talking Therapies program. The authors would like to thank the Mayden data services team for their role in curating the data for this work and their expertise in understanding the data and accompanying quality issues. The authors would also like to thank Barney Bristow and David Betts for their help in reviewing this manuscript. The authors are grateful to the Talking Therapies services included in the analysis for their support and use of their data. This research was funded by Innovate UK as part of a KTP between the University of Bath and Mayden (funding reference number 10023951 and project reference number KTP013141).

Data Availability

The data sets analyzed during this study are not publicly available due to the data sharing agreements in place between Mayden (the software provider responsible for iaptus) and the psychological therapies services. This study used anonymized electronic health record data from the digital care record software iaptus.

Authors' Contributions

EY designed and performed the analysis and drafted the manuscript with support from AD and CV. AD, CV, and CE devised the study. All authors discussed the results and reviewed the final manuscript.

Conflicts of Interest

AD and CE are employed by Mayden. EY is employed by the University of Bath as a Knowledge Transfer Partnership (KTP) associate but works at Mayden in line with their working practices and is supervised by AD (industrial supervisor) and CV (academic supervisor). Mayden contributes to the KTP associate's salary by paying the University of Bath, but most of the salary is paid by the research grant, which has been funded by Innovate UK. CV has no conflicts of interest to declare.

Multimedia Appendix 1

Supplemental materials including data preparation details and additional results.

[PDF File (Adobe PDF File), 1404 KB - mental_v11i1e53894_app1.pdf]

References

1. The burden of disease in England compared with 22 peer countries: a report for NHS England. Public Health England. 2020. URL: https://www.gov.uk/government/publications/global-burden-of-disease-for-england-international-comparisons [accessed 2023-10-23]
2. McManus S, Bebbington P, Jenkins R, Brugha T. Mental health and wellbeing in England: adult psychiatric morbidity survey 2014. NHS Digital. 2016. URL: https://digital.nhs.uk/data-and-information/publications/statistical/adult-psychiatric-morbidity-survey/adult-psychiatric-morbidity-survey-survey-of-mental-health-and-wellbeing-england-2014 [accessed 2023-10-23]
3. Coronavirus and depression in adults, Great Britain: July to August 2021. Office for National Statistics. 2021. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/articles/coronavirusanddepressioninadultsgreatbritain/julytoaugust2021 [accessed 2023-10-23]
4. Clark DM, Layard R, Smithies R, Richards DA, Suckling R, Wright B. Improving access to psychological therapy: initial evaluation of two UK demonstration sites. Behav Res Ther 2009 Nov;47(11):910-920 [FREE Full text] [doi: 10.1016/j.brat.2009.07.010] [Medline: 19647230]
5. Psychological therapies, reports on the use of IAPT services, 2021-2022. NHS Digital. 2022. URL: https://digital.nhs.uk/data-and-information/publications/statistical/psychological-therapies-annual-reports-on-the-use-of-iapt-services/annual-report-2021-22 [accessed 2023-10-23]
6. NHS mental health implementation plan 2019/20-2023/24. National Health Service. 2019. URL: https://www.longtermplan.nhs.uk/publication/nhs-mental-health-implementation-plan-2019-20-2023-24/ [accessed 2023-10-23]
7. Clark DM. Implementing NICE guidelines for the psychological treatment of depression and anxiety disorders: the IAPT experience. Int Rev Psychiatry 2011 Aug;23(4):318-327 [FREE Full text] [doi: 10.3109/09540261.2011.606803] [Medline: 22026487]
8. Gyani A, Shafran R, Layard R, Clark DM. Enhancing recovery rates: lessons from year one of IAPT. Behav Res Ther 2013 Sep;51(9):597-606 [FREE Full text] [doi: 10.1016/j.brat.2013.06.004] [Medline: 23872702]
9. Clark DM, Canvin L, Green J, Layard R, Pilling S, Janecka M. Transparency about the outcomes of mental health services (IAPT approach): an analysis of public data. Lancet 2018 Feb 17;391(10121):679-686 [FREE Full text] [doi: 10.1016/S0140-6736(17)32133-5] [Medline: 29224931]
10. Common mental health problems: identification and pathways to care: clinical guideline [CG123]. National Institute for Health and Care Excellence. 2011. URL: https://www.nice.org.uk/guidance/cg123/chapter/Recommendations [accessed 2024-02-27]
11. The improving access to psychological therapies manual. National Health Service. URL: https://www.rcpsych.ac.uk/docs/default-source/improving-care/nccmh/iapt/the-iapt-manual-v6.pdf?sfvrsn=79ece7e5_2 [accessed 2024-04-26]
12. Richards DA, Bower P, Pagel C, Weaver A, Utley M, Cape J, et al. Delivering stepped care: an analysis of implementation in routine practice. Implement Sci 2012 Jan 16;7:3 [FREE Full text] [doi: 10.1186/1748-5908-7-3] [Medline: 22248385]
13. Boyd L, Baker E, Reilly J. Impact of a progressive stepped care approach in an improving access to psychological therapies service: an observational study. PLoS One 2019;14(4):e0214715 [FREE Full text] [doi: 10.1371/journal.pone.0214715] [Medline: 30964883]

14. Mental health statistics: prevalence, services and funding in England. UK Parliament: House of Commons Library. 2024. URL: https://researchbriefings.files.parliament.uk/documents/SN06988/SN06988.pdf [accessed 2024-04-26]
15. Davis A, Smith T, Talbot J, Eldridge C, Betts D. Predicting patient engagement in IAPT services: a statistical analysis of electronic health records. Evid Based Ment Health 2020 Feb;23(1):8-14 [FREE Full text] [doi: 10.1136/ebmental-2019-300133] [Medline: 32046987]
16. Saxon D, Barkham M, Bee P, Gellatly J, Faija C, Bower P. Telephone treatments in improving access to psychological therapies services: an analysis of use and impact on treatment uptake. BMC Psychiatry 2023 Feb 07;23(1):95 [FREE Full text] [doi: 10.1186/s12888-022-04404-1] [Medline: 36750788]
17. Garcia CD, Meincheim A, Faria Junior ER, Dallagassa MR, Sato DM, Carvalho DR, et al. Process mining techniques and applications – a systematic mapping study. Expert Syst Appl 2019 Nov;133:260-295. [doi: 10.1016/j.eswa.2019.05.003]
18. van der Aalst W. Process Mining: Data Science in Action. 2nd edition. Berlin, Germany: Springer; 2016.
19. Mans RS, van der Aalst WM, Vanwersch RJ, Moleman AJ. Process mining in healthcare: data challenges when answering frequently posed questions. In: Proceedings of the 2012 Joint Workshop on Process Support and Knowledge Representation in Health Care. 2012 Presented at: KR4HC '12; September 3, 2012; Tallinn, Estonia p. 140-153 URL: https://doi.org/10.1007/978-3-642-36438-9_10 [doi: 10.1007/978-3-642-36438-9_10]
20. Rojas E, Munoz-Gama J, Sepúlveda M, Capurro D. Process mining in healthcare: a literature review. J Biomed Inform 2016 Jun;61:224-236 [FREE Full text] [doi: 10.1016/j.jbi.2016.04.007] [Medline: 27109932]
21. Munoz-Gama J, Martin N, Fernandez-Llatas C, Johnson OA, Sepúlveda M, Helm E, et al. Process mining for healthcare: characteristics and challenges. J Biomed Inform 2022 Mar;127:103994 [FREE Full text] [doi: 10.1016/j.jbi.2022.103994] [Medline: 35104641]
22. Martin N, De Weerdt J, Fernández-Llatas C, Gal A, Gatta R, Ibáñez G, et al. Recommendations for enhancing the usability and understandability of process mining in healthcare. Artif Intell Med 2020 Sep;109:101962 [FREE Full text] [doi: 10.1016/j.artmed.2020.101962] [Medline: 34756220]
23. Dallagassa MR, dos Santos Garcia C, Scalabrin EE, Ioshii SO, Carvalho DR. Opportunities and challenges for applying process mining in healthcare: a systematic mapping study. J Ambient Intell Human Comput 2021 Feb 08;13(1):165-182 [FREE Full text] [doi: 10.1007/s12652-021-02894-7]
24. Manktelow M, Iftikhar A, Bucholc M, McCann M, O'Kane M. Clinical and operational insights from data-driven care pathway mapping: a systematic review. BMC Med Inform Decis Mak 2022 Feb 17;22(1):43 [FREE Full text] [doi: 10.1186/s12911-022-01756-2] [Medline: 35177058]
25. Williams S, Southall C, Haley S, Ba Dhafari T, Kemp S, Relton SD, et al. To the emergency room and back again: circular healthcare pathways for acute functional neurological disorders. J Neurol Sci 2022 Jun 15;437:120251 [FREE Full text] [doi: 10.1016/j.jns.2022.120251] [Medline: 35429701]
26. Potts C, Bond RR, Jordan JA, Mulvenna MD, Dyer K, Moorhead A, et al. Process mining to discover patterns in patient outcomes in a psychological therapies service. Health Care Manag Sci 2023 Sep;26(3):461-476 [FREE Full text] [doi: 10.1007/s10729-023-09641-8] [Medline: 37191758]
27. van Eck ML, Lu X, Leemans SJ, van der Aalst WP. PM2: a process mining project methodology. In: Proceedings of the 27th International Conference on Advanced Information Systems Engineering. 2015 Jun 8 Presented at: CAiSE '15; June 8-12, 2015; Stockholm, Sweden p. 297-313 URL: https://link.springer.com/chapter/10.1007/978-3-319-19069-3_19 [doi: 10.1007/978-3-319-19069-3_19]
28. Johnson OA, Dhafari TB, Kurniati A, Fox F, Rojas E. The ClearPath method for care pathway process mining and simulation. In: Proceedings of the 2018 International Workshops on Business Process Management Workshops. 2018 Presented at: BPM '18; September 9-14, 2018; Sydney, Australia p. 239-250 URL: https://link.springer.com/chapter/10.1007/978-3-030-11641-5_19 [doi: 10.1007/978-3-030-11641-5_19]
29. Janssenswillen G, Depaire B, Swennen M, Jans M, Vanhoof K. bupaR: enabling reproducible business process analysis. Knowl Based Syst 2019 Jan;163:927-930. [doi: 10.1016/j.knosys.2018.10.018]
30. Wickham H. Elegant graphics for data analysis. ggplot2 - Tidyverse. 2016. URL: https://ggplot2.tidyverse.org [accessed 2023-10-23]
31. iaptus: clients and reviews. Mayden. URL: https://iaptus.co.uk/who-uses-iaptus/best-clinical-system-clients-reviews/ [accessed 2023-10-23]
32. Dataset: mapping income deprivation at a local authority level: 2019. Office for National Statistics. 2021. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/datasets/mappingincomedeprivationatalocalauthoritylevel [accessed 2024-02-27]
33. Guidance on blended payment for mental health services, national tariff 2019/20. National Health Service. 2019. URL: https://www.england.nhs.uk/publication/past-national-tariffs-documents-and-policies/ [accessed 2024-02-27]
34. Bauer-Staeb C, Davis A, Smith T, Wilsher W, Betts D, Eldridge C, et al. The early impact of COVID-19 on primary care psychological therapy services: a descriptive time series of electronic healthcare records. EClinicalMedicine 2021 Jul;37:100939 [FREE Full text] [doi: 10.1016/j.eclinm.2021.100939] [Medline: 34386738]

Abbreviations

CBT: cognitive behavioral therapy
EHR: electronic health record
ETL: extract, transform, and load
HCRW: Health and Care Research Wales
HRA: Health Research Authority
IAPT: Improving Access to Psychological Therapies
IMD: index of multiple deprivation
LSOA: lower super output area
NHS: National Health Service
NICE: National Institute for Health and Care Excellence
ONS: Office for National Statistics

Edited by J Torous; submitted 25.10.23; peer-reviewed by B Boeckmann, O Johnson, Z Valero-Ramón, C Potts; comments to author 10.01.24; revised version received 01.03.24; accepted 19.03.24; published 21.05.24.

Please cite as:
Yardley E, Davis A, Eldridge C, Vasilakis C
Data-Driven Exploration of National Health Service Talking Therapies Care Pathways Using Process Mining: Retrospective Cohort Study
JMIR Ment Health 2024;11:e53894
URL: https://mental.jmir.org/2024/1/e53894
doi:10.2196/53894
PMID:38771630

Research Letter

The Frequency of Design Studies Targeting People With Psychotic Symptoms and Features in Mental Health Care Innovation: Secondary Analysis of a Systematic Review

Lars Veldmeijer[1,2,3], MA; Gijs Terlouw[2], PhD; Jim Van Os[1], PhD; Job Van 't Veer[2], PhD; Nynke Boonstra[1,2,3], PhD

[1]Department of Psychiatry, Utrecht University Medical Center, Utrecht, Netherlands

[2]Department of Healthcare and Welfare, NHL Stenden University of Applied Sciences, Leeuwarden, Netherlands

[3]KieN VIP Mental Health Care Services, Leeuwarden, Netherlands

Corresponding Author:
Lars Veldmeijer, MA
Department of Psychiatry
Utrecht University Medical Center
Heidelberglaan 100
Utrecht, 3584 CX
Netherlands
Phone: 31 68 287 4768
Email: lars.veldmeijer@nhlstenden.com

Abstract

This study examined and reflected on the frequency of people with psychotic symptoms and features as the target population in design studies for mental health care innovation.

(JMIR Ment Health 2024;11:e54202) doi:10.2196/54202

KEYWORDS

design approaches; design; innovation; innovative; innovate; innovations; psychiatry; mental health care; mental health; mental illness; mental disease; involvement; service users; people with lived experience; people with lived experiences; lived experience; lived experiences; co-creation; cocreation; psychosis; psychotic; schizophrenia; schizoid; schizotypal; paranoia; neurosis; hallucinosis; hallucination; hallucinations

Introduction

There is growing evidence highlighting the importance of involving people with lived experience in design processes in mental health care [1,2]. Particular attention should be directed toward the engagement of people with psychotic symptoms and features [3], as they often feel misunderstood due to their altered perceptions and subjective experiences [4,5]. A bottom-up review of the lived experience of psychosis emphasizes the complexity of psychotic symptoms and features and recommends including lived experience in designing mental health services to address these experiences and needs [6]. Design approaches can promote the involvement of people with firsthand experiences in the development of treatment, therapy, and recovery interventions for mental health care innovation [2]. Currently, it is unknown how frequent design studies specifically target people with psychotic symptoms and features. There is a scoping review of coproducing research on psychosis [7], but coproduction and design approaches are distinct methodologies. Design approaches facilitate designing initiatives that prioritize participants' needs, expertise, and knowledge whereas coproduction facilitates collaborative delivery and knowledge production. In this research letter, we present findings on the frequency of design studies targeting people with psychotic symptoms by analyzing a prior systematic review data set that focused on involving people with firsthand experiences in designing mental health care innovations. The primary objective of this secondary data analysis was to elucidate how often design studies in mental health care target people with psychotic symptoms and features.

Methods

Primary Data Set and Secondary Data Analysis

We conducted a secondary data analysis using a data set from a prior systematic review that assessed the involvement of service users and people with lived experience in the design processes of mental health care innovation. In the screening process and study selection of the prior systematic review, 33 papers met the inclusion criteria [2]. All included papers were original reports or papers that (1) involved service users, people with lived experience, or both; (2) mentioned design approaches; (3) involved an empirical study; and (4) conducted the study in settings including mental health care services or psychiatry programs. In this secondary analysis, we examined the primary data set to provide an overview of the frequency of design studies in mental health care focusing on people with psychotic symptoms and features. This data set is suitable for this analysis since the search strategy of the systematic review did not target specific mental health conditions.

Data Extraction and Categorization

Studies were categorized based on their primary target population as reported in the studies (Multimedia Appendix 1). We categorized broad terms like psychosis, which encompasses various symptoms and features like altered perceptions, as well as mental health conditions in which psychotic symptoms and features are prevalent, such as schizophrenia. In the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision* (DSM-5-TR), these symptoms, features, and conditions fall under the category "schizophrenia spectrum and other psychotic disorders," covering a spectrum of related mental health conditions [8], also referred to as the psychosis spectrum [9]. Studies addressing psychotic symptoms and features or related conditions alongside unrelated mental health conditions were labeled "various mental health conditions" due to their comorbid nature. We did not count these studies as primarily focusing on people with psychotic symptoms and features. Studies not centered on psychotic symptoms or related conditions were also categorized to determine the frequency of studies that target people with psychotic symptoms and features in the data set.

Research Question

We aimed to answer the following research question: how frequently are people with psychotic symptoms and features the target group in design studies that involve service users or people with lived experience in mental health care innovation?

Results

The Frequency of Design Studies Targeting People With Psychotic Symptoms and Features

Of the studies in the data set that focused on specific target populations, 18% (6/33) centered on psychosis and 9% (3/33) concentrated on schizophrenia. Since psychosis and schizophrenia are considered part of a broader defined spectrum of psychotic-related mental health conditions in which psychotic symptoms and features are prevalent, the total proportion of studies that primarily focused on people with psychotic symptoms and features was 27% (9/33). The largest group of studies, accounting for 34% (11/33) of the data set, did not focus on specific target populations and were classified as "various mental health conditions." In this category, 12% (4/33) included psychotic symptoms and features as a target in addition to other mental health conditions (Table 1).

Table 1. Target populations in mental health design studies.

Target population	Count, n (%)
Psychosis	6 (18)
Depression	4 (12)
Schizophrenia	3 (9)
Self-harm	2 (6)
Eating disorders	2 (6)
Substance use disorders	1 (3)
Borderline	1 (3)
Attention-deficit/hyperactivity disorder	1 (3)
Autism spectrum disorder	1 (3)
Bipolar disorder	1 (3)
Various mental health conditions	11 (34)

Discussion

This secondary data analysis revealed a notable emphasis on studies primarily targeting people with psychotic symptoms and features in mental health care design studies. This is noteworthy given the extensive range of mental health conditions in psychiatry, encompassing 21 categories according to the DSM-5-TR [8]. Although "schizophrenia spectrum and other psychotic disorders" constitutes only 4.67% (1/21) of these categories, 27% (9/33) of studies in our data set focused primarily on psychotic symptoms and features. This percentage is high, considering the lifetime prevalence of psychotic disorders is approximately 1% [10]. Another 12% (4/33) of the studies mention psychotic symptoms and features alongside or as a result of other mental health conditions. Although these studies did not focus primarily on people with psychotic symptoms and features, they have shown that much attention has been given to psychotic experiences in design studies. The substantial research focus on people with psychotic symptoms and features in design studies may be attributed to the limited progress in prognosis for severe cases despite extensive research and treatment efforts [11]. This may prompt designers and researchers to look for less conventional strategies to enforce novel promising solutions. Additionally, there is a growing call for attention to the subjective experience of psychotic symptoms and features in clinical care, as these vary from individual to individual (eg, [4-6]). Both factors underscore the urgency of involving people with firsthand experiences to capture the vividness of psychotic experiences in the design of innovative services and interventions, ultimately aiming to improve outcomes for service users.

Comparing the 9 studies that primarily focused on people with psychotic symptoms and features in this secondary data analysis to the results of the prior systematic review, we observed that 44% (4/9) demonstrated a high level of participant involvement in their design processes [2]. This is crucial for the development of new innovations because research shows psychotic symptoms and features can seem very different from a lived experience perspective compared to conventional psychiatric conceptualizations [6]. At the same time, the results stress the ongoing need to engage people with lived experience of psychotic symptoms and features in design studies, as more than half of the studies did not show the substantial involvement that would be expected of design processes that aim to tailor innovations to the needs of the target group. Consequently, we recommend future design studies targeting people with psychotic symptoms and features to adopt the co-design methodology, as co-design shows the highest participant involvement levels in mental health care design studies [2]. Furthermore, researchers are encouraged to use the participation matrix [12] alongside co-design to make intentional methodological decisions regarding the phases and roles in which people with lived experience are involved. To prevent tokenism and cooptation in design processes, researchers and designers are recommended to systematically coreflect with people with lived experience, exploring the roles played and distilling benefits and challenges from both perspectives.

Acknowledgments

We appreciate the financial support of the FAITH Research Consortium, GGZ-VS University of Applied Science, as well as from the NHL Stenden University of Applied Sciences PhD program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Categorization of studies based on their target population.
[DOCX File , 19 KB - mental_v11i1e54202_app1.docx]

References

1. Veldmeijer L, Terlouw G, van 't Veer J, van Os J, Boonstra N. Design for mental health: can design promote human-centred diagnostics? Design Health 2023 Feb 14;7(1):5-23. [doi: 10.1080/24735132.2023.2171223]
2. Veldmeijer L, Terlouw G, Van Os J, Van Dijk O, Van 't Veer J, Boonstra N. The involvement of service users and people with lived experience in mental health care innovation through design: systematic review. JMIR Ment Health 2023 Jul 25;10:e46590 [FREE Full text] [doi: 10.2196/46590] [Medline: 37490326]
3. Lal S, Malla A. Service engagement in first-episode psychosis: current issues and future directions. Can J Psychiatry 2015 Aug;60(8):341-345 [FREE Full text] [doi: 10.1177/070674371506000802] [Medline: 26454555]
4. Postmes L, Sno HN, Goedhart S, van der Stel J, Heering HD, de Haan L. Schizophrenia as a self-disorder due to perceptual incoherence. Schizophr Res 2014 Jan;152(1):41-50 [FREE Full text] [doi: 10.1016/j.schres.2013.07.027] [Medline: 23973319]
5. Ritunnano R, Kleinman J, Whyte Oshodi D, Michail M, Nelson B, Humpston CS, et al. Subjective experience and meaning of delusions in psychosis: a systematic review and qualitative evidence synthesis. Lancet Psychiatry 2022 Jun;9(6):458-476 [FREE Full text] [doi: 10.1016/S2215-0366(22)00104-3] [Medline: 35523211]
6. Fusar-Poli P, Estradé A, Stanghellini G, Venables J, Onwumere J, Messas G, et al. The lived experience of psychosis: a bottom-up review co-written by experts by experience and academics. World Psychiatry 2022 Jun 07;21(2):168-188 [FREE Full text] [doi: 10.1002/wps.20959] [Medline: 35524616]
7. Jakobsson CE, Genovesi E, Afolayan A, Bella-Awusah T, Omobowale O, Buyanga M, et al. Co-producing research on psychosis: a scoping review on barriers, facilitators and outcomes. Int J Ment Health Syst 2023 Aug 30;17(1):25 [FREE Full text] [doi: 10.1186/s13033-023-00594-7] [Medline: 37644476]
8. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR®). Washington DC: American Psychiatric Association; 2022.
9. Guloksuz S, van Os J. The slow death of the concept of schizophrenia and the painful birth of the psychosis spectrum. Psychol Med 2018 Jan;48(2):229-244. [doi: 10.1017/S0033291717001775] [Medline: 28689498]

10. Moreno-Küstner B, Martín C, Pastor L. Prevalence of psychotic disorders and its association with methodological issues. A systematic review and meta-analyses. PLoS One 2018;13(4):e0195687 [FREE Full text] [doi: 10.1371/journal.pone.0195687] [Medline: 29649252]
11. Molstrom I, Nordgaard J, Urfer-Parnas A, Handest R, Berge J, Henriksen MG. The prognosis of schizophrenia: a systematic review and meta-analysis with meta-regression of 20-year follow-up studies. Schizophr Res 2022 Dec;250:152-163 [FREE Full text] [doi: 10.1016/j.schres.2022.11.010] [Medline: 36417817]
12. Smits D, van Meeteren K, Klem M, Alsem M, Ketelaar M. Designing a tool to support patient and public involvement in research projects: the Involvement Matrix. Res Involv Engagem 2020;6:30 [FREE Full text] [doi: 10.1186/s40900-020-00188-4] [Medline: 32550002]

Abbreviations

DSM-5-TR: *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision*

Edited by J Torous; submitted 01.11.23; peer-reviewed by W Yassine, C Lim, A Hudon; comments to author 24.11.23; revised version received 11.12.23; accepted 13.12.23; published 09.01.24.

Please cite as:
Veldmeijer L, Terlouw G, Van Os J, Van 't Veer J, Boonstra N
The Frequency of Design Studies Targeting People With Psychotic Symptoms and Features in Mental Health Care Innovation: Secondary Analysis of a Systematic Review
JMIR Ment Health 2024;11:e54202
URL: https://mental.jmir.org/2024/1/e54202
doi:10.2196/54202
PMID:38194249

An Ethical Perspective on the Democratization of Mental Health With Generative AI

Zohar Elyoseph[1,2], BA, MA, PhD; Tamar Gur[3], BA, MA, PhD; Yuval Haber[4], BA, MA; Tomer Simon[5], BA, MA, PhD; Tal Angert[6], BA, MA; Yuval Navon[6], BA; Amir Tal[7], BA, MA, PhD; Oren Asman[7,8], LLB, LLM, LLD

1
2
3
4
5
6
7
8

Corresponding Author:
Zohar Elyoseph, BA, MA, PhD

Abstract

Knowledge has become more open and accessible to a large audience with the "democratization of information" facilitated by technology. This paper provides a sociohistorical perspective for the theme issue "Responsible Design, Integration, and Use of Generative AI in Mental Health." It evaluates ethical considerations in using generative artificial intelligence (GenAI) for the democratization of mental health knowledge and practice. It explores the historical context of democratizing information, transitioning from restricted access to widespread availability due to the internet, open-source movements, and most recently, GenAI technologies such as large language models. The paper highlights why GenAI technologies represent a new phase in the democratization movement, offering unparalleled access to highly advanced technology as well as information. In the realm of mental health, this requires delicate and nuanced ethical deliberation. Including GenAI in mental health may allow, among other things, improved accessibility to mental health care, personalized responses, and conceptual flexibility, and could facilitate a flattening of traditional hierarchies between health care providers and patients. At the same time, it also entails significant risks and challenges that must be carefully addressed. To navigate these complexities, the paper proposes a strategic questionnaire for assessing artificial intelligence–based mental health applications. This tool evaluates both the benefits and the risks, emphasizing the need for a balanced and ethical approach to GenAI integration in mental health. The paper calls for a cautious yet positive approach to GenAI in mental health, advocating for the active engagement of mental health professionals in guiding GenAI development. It emphasizes the importance of ensuring that GenAI advancements are not only technologically sound but also ethically grounded and patient-centered.

(JMIR Ment Health 2024;11:e58011) doi:10.2196/58011

KEYWORDS

ethics; generative artificial intelligence; generative AI; mental health; ChatGPT; large language model; LLM; digital mental health; machine learning; AI; technology; accessibility; knowledge; GenAI

Introduction

The Democratization of Information: From Print to Artificial Intelligence–Generated Content

The democratization of information has been described as the process of making knowledge more accessible, inclusive, and transparent to a broad audience, often facilitated by technological advancements [1]. Over the past few centuries, a transformation has occurred in how knowledge is accessed, disseminated, and used. Historically, access to information and technology was often restricted to a privileged few—aristocrats, the church, academics, researchers, and professionals who had the means to gather and interpret data. The printing press served as an important milestone in the democratization of information. With the development of the steam locomotive (trains) in the 1800s, printed newspapers and journals that included news and ideas could be disseminated quickly and relatively cheaply across large distances. More recently, in the 1990s, when the internet became widely accessible, search engines enabled widespread and decentralized access to knowledge. Web 2.0, a participatory web with wiki platforms and other people-centric websites, later leveraged the web and engaged its users and elicited their collective intelligence [2]. This was followed by open-source movements that promoted sharing code and software frameworks freely, allowing developers globally to build upon and improve existing technologies. All these advancements led to an unprecedented amount of information freely accessible to billions. As technology continues to be developed and advanced, we argue that a new era in information democratization has begun in 2022 when various generative artificial intelligence (GenAI) platforms opened their platform for anyone with an internet connection to be used. The current phase of technology democratization marks a shift away from its exclusive use by computer scientists, researchers, and artificial intelligence (AI) professionals and toward reaching a broader audience with less expertise. Users now have more opportunities to actively participate in improving current technologies and may play a larger role in their advancement. GenAI technologies, such as large language models (LLMs) with visual and auditory elements, provide billions of people with direct access to cutting-edge technology, transcending the concept of "end users" and allowing them to perform tasks previously reserved for those with extensive computer science knowledge. Today, laypeople can use such technologies to create code, software, and GenAI models by expressing their desires in a natural language. These technologies drive the democratization of knowledge and technology by providing tailored, personalized, and on-demand information on a massive scale.

While the growing popularity of GenAI has undoubtedly aided the democratization of information, it also raises serious concerns about surveillance and control. Considering insights from Foucauldian theories, the widespread integration of GenAI into social discourse raises concerns about the potential abuse of authority and narrative manipulation. Furthermore, relying on GenAI-driven decision-making processes risks reinforcing existing power dynamics and marginalizing specific voices in society. As GenAI affects more aspects of our lives, it is crucial to critically evaluate its implications on privacy, autonomy, and the integrity of information dissemination. This paper provides a sociohistorical perspective for the theme issue on "Responsible Design, Integration, and Use of Generative AI in Mental Health." It considers the ethics of using GenAI for the democratization of mental health knowledge and practice.

Democratization of Knowledge in Mental Health: A Permanent Shift

Since the launch of ChatGPT (OpenAI) in November 2022, multiple studies have shown the transformative potential of GenAI in mental health [3-12]. This is crucial to recognize as we delve into the advantages and risks of GenAI in democratizing mental health knowledge. GenAI can address the global shortage of mental health professionals, reshape mental health care, advance diagnostic accuracy, improve treatment personalization, and enhance the overall accessibility of mental health services. It can facilitate mental health education and awareness, provide various self-help or self-paced mental health support tools, and so forth. However, it also poses risks, especially in the context of therapy and personalized mental health interventions.

Advantages of GenAI in Mental Health Democratization

Accessibility

A fundamental challenge in the mental health field is the limited access to mental health care both in developed and developing countries, as well as the disparities in access to mental health care [13-18]. Factors such as socioeconomic status [14], geographical location [19], linguistic barriers [20], and cultural disparities [21] present significant hurdles to the accessibility of mental health services. GenAI may be leveraged in mitigating these barriers through the development of linguistically and culturally attuned resources and potentially offering solutions adaptable to various economic backgrounds [19].

Personalized Responses

AI provides an opportunity for a new era of mental health services that are sharply attuned to the individual needs and preferences of each patient [22]. Within the framework of treatment with a mental health professional, AI and GenAI technologies can facilitate a deeper understanding of a person's unique psychological landscape by considering a myriad of factors such as their biological predispositions, societal and cultural influences, and personal preferences [23]. These technologies have the potential to analyze complex patterns and variances in individual experiences and histories, which can be instrumental in crafting responses and interventions that are tailored to suit individual needs as well as preferences. In contrast, non-AI–based search engines still used by many over the internet fall short of this goal, as searches often lack context and fail to grasp the nuanced complexities of individual experiences and histories. Moreover, personalized responses enabled by GenAI can envision a departure from generic diagnostic protocols, creating room for treatment paradigms that are designed with a more intricate understanding of individual idiosyncrasies. This method assumes that mental health is influenced by biological, social, and personal factors, aiming to provide solutions tailored to the distinct characteristics of each person's experiences. Services that are tailored for each user specifically may lead the mental health field toward a more inclusive and equitable future.

Theoretical or Conceptual Pluralism

In the pursuit of democratizing mental health, the goal is to enhance the field's adaptability and inclusivity by embracing a variety of perspectives. GenAI mental health tools may be leveraged in three main ways and they are (1) providing a specific theory-based or perspective-based analysis of various scenarios with the ability to shift between theories or perspectives; (2) generating a dialogue or a synthesis between various theories or perspectives; and (3) potentially create a new theory or hypothesis (see Textbox 1).

Textbox 1. Using generative artificial intelligence for multiple theory or perspective analysis, integration, and formation of novel ideas (generated by Claude AI 2.1 pro and ChatGPT 4; December 2023).

Provide theory-based analysis

- Cognitive theory (Beck): The patient exhibits the key hallmarks of severe major depressive disorder per Beck's cognitive model—pervasive negative cognitive schemas, automatized depressive cognitions, diminished mood, and inability to work.
- Object relations theory (Winnicott): The emergence of this severe depressive episode despite no clear environmental trigger suggests potential psychological vulnerabilities rooted in early childhood. Lack of good-enough mothering could have prevented the development of a cohesive self, resulting in a false self-presentation vulnerable to fragmentation and depressive affect under stress.

Synthesize perspectives

- This patient's sudden-onset depression despite minimal stressors points to vulnerabilities in self-structure from poor early nurturing. His fragmented functioning aligns with Klein's paranoid-schizoid position, while loss of values reflects ACT's model. Synthesis suggests deficient parenting prevented integrating difficult experiences into a cohesive self. When his false self faltered, overwhelming feelings emerged that he could neither accept mindfully nor confront to rebuild his personality. Addressing both positive introjections in Kleinian play therapy and value-guided committed action using ACT may enable overcoming this acutely disabled state.

Create new theory

- The Wilted Plant Theory of Depression presents a unique botanical metaphor to understand and address depression. It likens the human psyche to a plant that requires "emotional sunlight" (positive interactions), "psychological nutrients" (intellectual engagement and meaningful activities), and "emotional hydration" (free expression of emotions) to thrive. Just as a plant wilts without proper care, the theory posits that the human mind endures in the absence of these essential elements, leading to depression. This metaphorical approach highlights the importance of a nurturing environment for maintaining and restoring mental well-being.

In other words, this pluralism may facilitate the synchronous operation of a variety of therapeutic approaches, philosophies, and cultural viewpoints. For instance, we can observe opportunities where GenAI might enable the integration and dialogue between traditionally distinct therapeutic methodologies such as cognitive behavioral therapy and psychodynamic approaches [24]. Here, the structural and goal-oriented strategies of cognitive behavioral therapy could be married with the depth of insight derived from psychodynamic explorations, engendering a more rounded approach to mental health care [25]. Moreover, the perspective of psychiatry, with its medically grounded insights, could be brought into conversation with psychological approaches, nurturing a space where medical, psychological, and holistic strategies can come together to form a more comprehensive view of mental health care.

With that being said, the way current LLMs work, the mere ability to be creative in connecting various methods in a convincing manner may be wonderful for brainstorming new eclectic concepts and therapeutic approaches but is in no way, in itself, evidence of its feasibility and reliability in real life.

Equality and Reduction of Social Gaps

GenAI-powered LLMs hold the potential to foster greater equality and reduce prevailing social gaps [26]. By harnessing vast arrays of data and insights, such models may facilitate interventions crafted to meet the varied needs of different populations, including those historically underserved or marginalized [27,28]. For instance, developing and distributing mental health programs in languages and dialects that have historically lacked sufficient resources. It could further enable community-centric initiatives, enhancing the representation of diverse groups in the mental health discourse, thereby paving pathways for more localized and culturally sensitive interventions. Moreover, GenAI may be able to identify and relate to social aspects that are at times highly relevant in mental health scenarios. GenAI-based LLMs with access to information about symptoms, illnesses, and treatments, may allow laymen to ask questions and receive clarifications usually only available by contacting an expert. This may also facilitate in later contacting the relevant health care professionals, thus saving time and resources. This could become more relevant and useful when data sets used for training foundational models or fine-tuning general-purpose models have more representation of various languages and cultures.

Therapist-Patient Engagement

One of the notable strengths of GenAI is its ability to reduce bureaucratic and administrative burdens in mental health care settings. It can provide transformative solutions by automating tasks like transcription, summarization [29], and form filling. Using these technologies, therapists may simplify administrative processes, freeing up more time and attention to provide direct patient care. With AI handling routine paperwork and data entry tasks, clinicians are freed from screens and forms, allowing them to focus on building a connection, conducting assessments, and providing personalized interventions to their patients. This not only increases the efficiency of mental health services, but also improves the overall quality of patient care by encouraging more meaningful interactions between therapists and their clients.

Flattening of Hierarchies

The advent of GenAI bears the promising potential to flatten traditional hierarchical structures prevalent in the mental health sector, fundamentally altering the dynamics between health care providers and recipients [30]. Historically, psychiatrists and psychologists held a pronounced degree of authority, largely stemming from their exclusive access to specialized knowledge. If knowledge is no longer confined to a select few but is accessible to a wider population, this allows for a more balanced dynamic between mental health professionals and individuals seeking help [12,30]. It could empower individuals with insights and understanding of their own mental health conditions, fostering more collaborative therapeutic relationships and potentially leading to more fruitful and synergistic therapy sessions grounded on mutual understanding and shared knowledge. As we have previously defined, the introduction of GenAI into the field of mental health can be seen as an "artificial third" that changes the dynamic between mental health professionals and patients, so that in fact a new relationship triangle is created characterized by the flattening of the existing power hierarchy between experts and patients [9,12,31]. In this vision of a democratized mental health landscape, GenAI acts as an equalizer, breaking down barriers to knowledge accessibility and cultivating a health care landscape built on collaboration, understanding, and shared expertise. This is further evident with the context window of LLMs increasing dramatically over a short period of time (for instance, OpenAI ChatGPT increased from 4000 tokens to 128,000 on November 2023 and Google's Gemini increased on February 2024 to 1 million tokens). This allows end users to upload a large amount of information (such as hundreds of text pages, images, and videos with clinically relevant information) and discuss it with a chatbot during one prompt or conversation.

Risks to Mental Health Democratization Through AI

Corporate Centralization

Corporate centralization in mental health services, facilitated by GenAI, carries a significant risk of prioritizing profit over individual-centered care, widening disparities in access and quality of mental health care, and influencing public health narratives toward economic gains rather than genuine support and care [32]. GenAI can assume a therapeutic role and be designed to foster trust and build rapport with users [33], making it a potent instrument in the hands of entities that may have their own agendas. This includes but is not limited to the promotion of specific political narratives, ideological indoctrination, aggressive marketing or unnoticeable marketing strategies, also known as dark pattern AI [34], taking advantage of its persuasive power for psychological manipulation and control. The centralization of power-knowledge, without emphasizing checks and balances in a small number of economic corporations could potentially create a facade of democratization of mental health, but not reflect a true democratization in the field.

Information Transparency

Information transparency could be divided into 2 major aspects of the "one-way mirror," as only 1 party is exposed to the other party's information. On the provider side of the "mirror" there are real concerns about the management of user data. These include transactional misuses, such as unauthorized sales to third parties or its exploitation for targeted marketing [35]. However, more sinister breaches of personal privacy could also be achieved since GenAI systems have the potential to intrusively analyze personal conversations, behaviors, and emotions [5] without explicit consent. Moreover, the data harnessed might even be used in training AI systems, a process that remains largely concealed from the end users. Indeed, the algorithms driving this AI application function are much like a "black box" shrouded in mystery with no clarity as to how determinations and analyses are reached [36]. Alas, democratization is thus a double-edged sword; while GenAI may indeed democratize access to mental health resources, the current level of transparency and explainability to users of its operational mechanics may limit a truly informed user engagement, limiting the realization of a democratized system with empowered users [37]. When core aspects of an alignment process, including the embedded objectives, values, and ethics, are not made clear and transparent to users [7], it can result in power becoming concentrated in an entity whose true incentives remain obscured.

People's Misperceptions of AI

Overview

The level of expertise that people attribute to GenAI tools may be affected by their perceptions of technology. Numerous studies have shown that people tend to imbue AI systems with significant epistemic authority stemming largely from the veneer of impartiality and objectivity these technologies present. This attribution of high epistemic authority to GenAI systems may also pose a significant risk. Epistemic authority essentially refers to the weight and trust in a source as a repository of knowledge and information [38]. While GenAI systems can rely on a vast amount of data, the elevation of their epistemic authority could also carry detrimental effects for both health care providers and patients.

Risk of Misinformation

GenAI systems are not infallible; they can make mistakes, be based on incorrect data, or present biased viewpoints, thus generating incorrect advice or guidance. In the context of GenAI's mistakes, consider mentioning the term "hallucinations" [39] or "confabulations" (which in our eyes is problematic terminology because it can be perceived as offensive and because it has an anthropomorphic assumption). Attributing high epistemic authority to GenAI may lead to unconditional acceptance of its output, without a critical evaluation of the veracity of the information provided.

GenAI Overreliance With Reduced Patient Self-Engagement

While incorporating GenAI into mental health care has numerous advantages, it also highlights the serious risk of epistemic bias among both therapists and patients. Attributing high epistemic authority to AI may overshadow not only the expertise and nuanced understanding of health care providers but also the personal experiences and insights of the patients themselves. Overreliance on GenAI in healthcare may reduce patient self-engagement by prioritizing AI-generated insights over the comprehensive understanding provided by health care providers and patients, potentially undermining individuals' active participation in their mental health journey and resulting in less effective treatments. Relying on the AI's ability to articulate and construct our own thoughts and feelings, thus "Letting the Machine speak for us" could also mean relinquishing effort in our interpersonal engagements, including in therapy, reducing one's possibilities for self-understanding and growth [40]. Furthermore, therapists are vulnerable to epistemic bias by relying too heavily on AI-generated insights, potentially missing important nuances in patient narratives and clinical assessments. This overreliance on GenAI may unintentionally limit the therapist's ability to engage deeply with patients, as algorithmic recommendations may not fully capture the complexities of individual experiences. As a result, it is critical for therapists to be vigilant against the influence of epistemic bias in their practice, striking a balance between using GenAI tools and retaining the essential human elements of empathy [41], intuition, and clinical judgment.

Increased Power Imbalance

The elevation of GenAI as a central epistemic figure may lead to a power imbalance, where knowledge is centralized in the hands of a GenAI entity that is under the control of economic corporations. This undermines the democratization ethos which seeks to foster a collaborative and pluralistic approach to mental health and where knowledge is the result of collective insight, involving a harmonic convergence of professional guidance and personal experiences. Thus, while GenAI offers a promise of democratized access to information, it also threatens to replace a current knowledge monopoly (currently in the hands of mental health experts) with a monopoly of a small number of LLM companies, which is counterintuitive to the principles of democratization that advocate for a decentralized, collective approach to knowledge dissemination and use. It should be noted that open-source models that are available to the public enable decentralized technological development and constitute a decentralized force, and as these models continue to develop and improve, the risks of a few companies monopolizing a field (including mental health) will be diminished.

People in emotional need may become dependent on or attached to GenAIs in potentially nonadaptive ways. For instance, many of Replika's AI chatbot's 7 million users see it as their best friend or even a family member [33]. While examining the relationship with this chatbot, researchers found that the patterns of dependency were different from other technological dependencies as it involved people feeling that Replika had needs and emotions that they needed to cater to [42]. Accordingly, there is an additional layer of risk relating to the authenticity of this "relationship" with a machine [5-8] whereby the humanization of GenAI [27] may imitate human agency in a manner that could alter our perception of good and healthy lives [43].

Regulation Issues

As GenAI technology, driven mainly by for-profit private corporations, starts to enter the sphere of mental health services, there's a growing concern regarding its adherence to the established protocols that have historically governed mental health services. While the democratization endeavor seeks to foster inclusivity and accessibility, the introduction of GenAI poses a conundrum; it opens avenues for unprecedented access to mental health resources but at the potential cost of diluting the standard of care and ethical considerations traditionally upheld by mental health professionals. One of the major bioethical and legal challenges in this regard is how care ethics concepts could be relevant within the developing field of "responsible AI," to more fully consider AI's impact on human relationships [44].

Objective Perspective Versus Gender, Socioeconomic, and Ethnic Biases

Integrating GenAI in mental health services is challenged by how one balances between clinically informed judgments and reducing bias. AI systems rely on preexisting data that were produced, collected, and potentially also labeled by humans, thereby holding an intrinsic propensity to reflect societal biases, including those grounded in gender, socioeconomic factors, and ethnicity [45]. At the same time, demographic factors play a critical role in assessing individual health risks and conditions. Consequently, the AI alignment should continuously navigate the narrow pathway between eliminating biases and retaining critical data essential for accurate clinical judgments. From the democratization perspective, AI may perpetuate biases, and at the same time, if overly aligned, may fail to provide users' expectations of a personalized and efficient mental health service. Thus, the path forward calls for a nuanced and vigilant development process for AI systems, one that meticulously harmonizes statistical evidence with fundamental democratic values.

The claims raised above suggest that AI represents a real opportunity to advance the field of mental health, as it will likely be increasingly present in our lives and its adoption seems inevitable. We propose addressing the risks outlined in this study as thought tools in the development of applied AI tools for responsible use in mental health, rather than viewing them as warnings against using this technology.

Guiding Ethical Development: A Strategic Questionnaire for AI Mental Health Applications

Considering the potential risks and opportunities identified in the discourse on GenAI applications in mental health, we propose a set of carefully formulated questions designed to assess GenAI's capability to enhance mental health care (Textbox 2). These questions are intended for use in the development processes of mental health applications, ensuring a comprehensive evaluation of both the benefits and the risks involved. We have deliberately distinguished between risks and opportunities, recognizing that they do not always exist on the same scale. Namely, a significant risk does not necessarily negate the potential benefits of an AI application, and vice versa. Hence, it is imperative to conduct a differential assessment for each application, weighing its specific risks against its potential opportunities. This approach is grounded in a nuanced understanding that while GenAI offers remarkable prospects for democratizing mental health care, its implementation must be navigated with caution to avoid unintended consequences. The proposed questionnaire is thus an extension of the discourse presented so far in the study, bridging the theoretical considerations with practical evaluation tools.

Textbox 2. A strategic questionnaire for artificial intelligence mental health applications.

Promoting democratization

- Accessibility: Does it improve access to mental health services for diverse individuals, including marginalized communities?
- User empowerment: Does it provide tools for self-care and informed decision-making?
- Facilitating collaboration and shared decision-making: Does it facilitate a collaborative approach between patients and health care providers, allowing for an AI augmented shared decision-making process?
- Inclusivity: Can it adapt to diverse cultural, socioeconomic, and personal needs promoting inclusive care?
- Transparency: Does it provide clear information about its functionalities, limitations, and data usage?

Identifying potential risks

- Data privacy and security: How are privacy and security risks mitigated?
- Bias and inequality: Does it reinforce societal biases or exacerbate inequalities in mental health care?
- Overdependence or addiction: How likely is it for users to develop over reliance or dependence on this tool?
    - Misinformation: How likely is the system to provide false or misinformation or lead to neglecting of human-based professional advice?
    - Corporate involvement: Are intentional or nonintentional considertations steering the clinical information or advice provided? Or to compromising ethical standards in patient care?
- Overshadowing human expertise: Does it diminish the role or undermine the expertise of mental health professionals?

Discussion and Conclusion

The integration of GenAI in mental health care, as outlined in this paper, is a potent and inevitable aspect of the broader democratization movement. The ethical implications of not leveraging GenAI in this field are profound, given its potential to revolutionize care and treatment. GenAI introduces a paradigm shift, challenging existing dynamics within mental health care and presenting opportunities to resolve longstanding issues in the field. However, this shift is not without its challenges; it disrupts established power structures, provokes questions about truth and the nature of expertise, and raises concerns about the potential displacement of human roles by technology.

The transition to GenAI-driven mental health care is an inescapable reality, accompanied by considerable promises. It is imperative that the mental health field not only adapts to this new landscape but actively shapes it. This task should not be left solely to engineers and computer scientists; mental health professionals must play a pivotal role. Their involvement is critical to ensure that GenAI development aligns with the ethical standards and therapeutic goals of mental health care. In response to these challenges, our study proposes a structured questionnaire designed to guide responsible AI development in mental health. This questionnaire serves as a road map, delineating crucial considerations for balancing the opportunities and risks associated with GenAI integration. It emphasizes the need for a cautious yet optimistic approach to AI development and regulation, ensuring that advancements in mental health care are not only technologically sound but also ethically grounded and patient-centered. As we conclude, we call upon mental health associations and professionals to engage with these guidelines actively. By adopting a stance that is both critically vigilant and constructively engaged, the mental health field can navigate the complexities of GenAI integration. This approach is vital for harnessing AI's potential while safeguarding the foundational values and ethical principles of mental health care. Our contribution, through this discussion and the questionnaire, aims to ensure that the AI revolution in mental health is not only technologically advanced but also democratically enriched and ethically sound.

Conflicts of Interest

AT and OA are guest editors for the special issue "Responsible Design, Integration, and Use of Generative AI" in *JMIR Mental Health*. The other authors declare no conflict of interest.

References

1. Wallace DP, Van Fleet C. From the editors: the democratization of information? Wikipedia as a reference resource. Ref User Serv Quart 2005;45(2):100-103 [FREE Full text]
2. Murugesan S. Understanding web 2.0. IT Prof 2007;9(4):34-41. [doi: 10.1109/MITP.2007.78]
3. Elyoseph Z, Hadar-Shoval D, Asraf K, Lvovsky M. ChatGPT outperforms humans in emotional awareness evaluations. Front Psychol 2023 May;14:1199058. [doi: 10.3389/fpsyg.2023.1199058] [Medline: 37303897]
4. Elyoseph Z, Levkovich I. Beyond human expertise: the promise and limitations of ChatGPT in suicide risk assessment. Front Psychiatry 2023 Aug;14:1213141. [doi: 10.3389/fpsyt.2023.1213141] [Medline: 37593450]

5. Elyoseph Z, Refoua E, Asraf K, Lvovsky M, Shimoni Y, Hadar-Shoval D. Capacity of generative AI to interpret human emotions from visual and textual data: pilot evaluation study. JMIR Ment Health 2024 Feb 6;11:e54369. [doi: 10.2196/54369] [Medline: 38319707]
6. Elyoseph Z, Levkovich I, Shinan-Altman S. Assessing prognosis in depression: comparing perspectives of AI models, mental health professionals and the general public. Fam Med Community Health 2024 Jan 9;12(Suppl 1):e002583. [doi: 10.1136/fmch-2023-002583] [Medline: 38199604]
7. Elyoseph Z, Hadar Shoval D, Levkovich I. Beyond personhood: ethical paradigms in the generative artificial intelligence era. Am J Bioeth 2024 Jan;24(1):57-59. [doi: 10.1080/15265161.2023.2278546] [Medline: 38236857]
8. Elyoseph Z, Levkovich I. Comparing the perspectives of generative AI, mental health experts, and the general public on schizophrenia recovery: case vignette study. JMIR Ment Health 2024 Mar 18;11:e53043. [doi: 10.2196/53043] [Medline: 38533615]
9. Hadar-Shoval D, Asraf K, Mizrachi Y, Haber Y, Elyoseph Z. Assessing the alignment of large language models with human values for mental health integration: cross-sectional study using Schwartz's theory of basic values. JMIR Ment Health 2024 Apr 9;11:e55988. [doi: 10.2196/55988] [Medline: 38593424]
10. Levkovich I, Elyoseph Z. Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. Fam Med Community Health 2023 Sep;11(4):e002391. [doi: 10.1136/fmch-2023-002391] [Medline: 37844967]
11. Levkovich I, Elyoseph Z. Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: vignette study. JMIR Ment Health 2023 Sep 20;10:e51232. [doi: 10.2196/51232] [Medline: 37728984]
12. Tal A, Elyoseph Z, Haber Y, et al. The artificial third: utilizing ChatGPT in mental health. Am J Bioeth 2023 Oct;23(10):74-77. [doi: 10.1080/15265161.2023.2250297] [Medline: 37812102]
13. Araya R, Zitko P, Markkula N, Rai D, Jones K. Determinants of access to health care for depression in 49 countries: a multilevel analysis. J Affect Disord 2018 Jul;234:80-88. [doi: 10.1016/j.jad.2018.02.092] [Medline: 29524750]
14. Hodgkinson S, Godoy L, Beers LS, Lewin A. Improving mental health access for low-income children and families in the primary care setting. Pediatrics 2017 Jan;139(1):e20151175. [doi: 10.1542/peds.2015-1175] [Medline: 27965378]
15. The burden of mental disorders in the region of the Americas. Pan American Health Organization. 2018. URL: http://iris.paho.org/xmlui/handle/123456789/49578 [accessed 2024-09-04]
16. Vigo DV, Kestel D, Pendakur K, Thornicroft G, Atun R. Disease burden and government spending on mental, neurological, and substance use disorders, and self-harm: cross-sectional, ecological study of health system response in the Americas. Lancet Public Health 2019 Feb;4(2):e89-e96. [doi: 10.1016/S2468-2667(18)30203-2] [Medline: 30446416]
17. Whiteford HA, Degenhardt L, Rehm J, et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. Lancet 2013 Nov 9;382(9904):1575-1586. [doi: 10.1016/S0140-6736(13)61611-6] [Medline: 23993280]
18. Mental health action plan 2013-2020. World Health Organization. 2013. URL: https://www.who.int/mental_health/publications/action_plan/en/ [accessed 2024-09-04]
19. Cummings JR, Allen L, Clennon J, Ji X, Druss BG. Geographic access to specialty mental health care across high- and low-income US communities. JAMA Psychiatry 2017 May 1;74(5):476-484. [doi: 10.1001/jamapsychiatry.2017.0303] [Medline: 28384733]
20. Ohtani A, Suzuki T, Takeuchi H, Uchida H. Language barriers and access to psychiatric care: a systematic review. Psychiatr Serv 2015 Aug 1;66(8):798-805. [doi: 10.1176/appi.ps.201400351] [Medline: 25930043]
21. Byrow Y, Pajak R, Specker P, Nickerson A. Perceptions of mental health and perceived barriers to mental health help-seeking amongst refugees: a systematic review. Clin Psychol Rev 2020 Feb;75:101812. [doi: 10.1016/j.cpr.2019.101812] [Medline: 31901882]
22. Graham S, Depp C, Lee EE, et al. Artificial intelligence for mental health and mental illnesses: an overview. Curr Psychiatry Rep 2019 Nov 7;21(11):116. [doi: 10.1007/s11920-019-1094-0] [Medline: 31701320]
23. Hadar-Shoval D, Elyoseph Z, Lvovsky M. The plasticity of ChatGPT's mentalizing abilities: personalization for personality structures. Front Psychiatry 2023;14:1234397. [doi: 10.3389/fpsyt.2023.1234397] [Medline: 37720897]
24. Grodniewicz JP, Hohol M. Waiting for a digital therapist: three challenges on the path to psychotherapy delivered by artificial intelligence. Front Psychiatry 2023 Jun;14:1190084. [doi: 10.3389/fpsyt.2023.1190084] [Medline: 37324824]
25. Pilecki B, Thoma N, McKay D. Cognitive behavioral and psychodynamic therapies: points of intersection and divergence. Psychodyn Psychiatry 2015 Sep;43(3):463-490. [doi: 10.1521/pdps.2015.43.3.463] [Medline: 26301762]
26. van Heerden AC, Pozuelo JR, Kohrt BA. Global mental health services and the impact of artificial intelligence–powered large language models. JAMA Psychiatry 2023 Jul 1;80(7):662. [doi: 10.1001/jamapsychiatry.2023.1253]
27. Ferrario A, Sedlakova J, Trachsel M. The role of humanization and robustness of large language models in conversational artificial intelligence for individuals with depression: a critical analysis. JMIR Ment Health 2024 Jul 2;11:e56569. [doi: 10.2196/56569] [Medline: 38958218]
28. Fiske A, Henningsen P, Buyx A. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J Med Internet Res 2019 May 9;21(5):e13216. [doi: 10.2196/13216] [Medline: 31094356]

29. Adhikary PK, Srivastava A, Kumar S. Exploring the efficacy of large language models in summarizing mental health counseling sessions: a benchmark study. JMIR Ment Health 2024;11:e57306.
30. Cohen IG. What should ChatGPT mean for bioethics? Am J Bioeth 2023 Oct;23(10):8-16. [doi: 10.1080/15265161.2023.2233357] [Medline: 37440696]
31. Haber Y, Levkovich I, Hadar-Shoval D, Elyoseph Z. The artificial third: a broad view of the effects of introducing generative artificial intelligence on psychotherapy. JMIR Ment Health 2024 May 23;11:e54781. [doi: 10.2196/54781] [Medline: 38787297]
32. Zajko M. Artificial intelligence, algorithms, and social inequality: sociological contributions to contemporary debates. Sociol Compass 2022 Mar;16(3):e12962. [doi: 10.1111/soc4.12962]
33. Munn N, Weijers D. Corporate responsibility for the termination of digital friends. AI Soc 2023 Aug;38(4):1501-1502. [doi: 10.1007/s00146-021-01276-z]
34. Freeman R. Generative artificial intelligence, automated user interfaces, and the new laws of dark patterns. The National Law Review. 2023. URL: https://www.natlawreview.com/article/generative-artificial-intelligence-automated-user-interfaces-and-new-laws-dark [accessed 2024-10-04]
35. Coghlan S, Leins K, Sheldrick S, Cheong M, Gooding P, D'Alfonso S. To chat or bot to chat: ethical issues with using chatbots in mental health. Dig Health 2023 Jun;9:20552076231183542. [doi: 10.1177/20552076231183542] [Medline: 37377565]
36. Castelvecchi D. Can we open the black box of AI? Nat New Biol 2016 Oct;538(7623):20-23. [doi: 10.1038/538020a] [Medline: 27708329]
37. von Eschenbach WJ. Transparency and the black box problem: why we do not trust AI. Philos Technol 2021;34:1607-1622. [doi: 10.1007/s13347-021-00477-0]
38. Reed B. Epistemic authority: a theory of trust, authority, and autonomy in belief. Philos Rev 2015 Jan 1;124(1):159-162. [doi: 10.1215/00318108-2812701]
39. Hatem R, Simmons B, Thornton JE. A call to address AI "hallucinations" and how healthcare professionals can mitigate their risks. Cureus 2023 Sep;15(9):e44720. [doi: 10.7759/cureus.44720] [Medline: 37809168]
40. Hartford A, Stein DJ. The machine speaks: conversational AI and the importance of effort to relationships of meaning. JMIR Ment Health 2024 Jun 18;11:e53203. [doi: 10.2196/53203] [Medline: 38889401]
41. Rubin M, Arnon H, Huppert JD, Perry A. Considering the role of human empathy in AI-driven therapy. JMIR Ment Health 2024 Jun 11;11:e56529. [doi: 10.2196/56529] [Medline: 38861302]
42. Laestadius L, Bishop A, Gonzalez M, Illenčík D, Campos-Castillo C. Too human and not human enough: a grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. New Media Soc 2022:146144482211420. [doi: 10.1177/14614448221142007]
43. Asman O, Tal A, Barilan YM. Conversational artificial intelligence-patient alliance turing test and the search for authenticity. Am J Bioeth 2023 May;23(5):62-64. [doi: 10.1080/15265161.2023.2191046] [Medline: 37130413]
44. Tavory T. Regulating AI in mental health - the ethics of care perspective. JMIR Ment Health 2024 Jul 20 Accepted. [doi: 10.2196/58493]
45. Timmons AC, Duong JB, Simo Fiallo N, et al. A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. Perspect Psychol Sci 2023 Sep;18(5):1062-1096. [doi: 10.1177/17456916221134490] [Medline: 36490369]

Abbreviations

AI: artificial intelligence
GenAI: generative artificial intelligence
LLM: large language model

Edited by J Torous; submitted 02.03.24; this is a non–peer-reviewed article;accepted 24.07.24; published 17.10.24.

Please cite as:
Elyoseph Z, Gur T, Haber Y, Simon T, Angert T, Navon Y, Tal A, Asman O
An Ethical Perspective on the Democratization of Mental Health With Generative AI
JMIR Ment Health 2024;11:e58011
URL: https://mental.jmir.org/2024/1/e58011
doi:10.2196/58011

The Role of Humanization and Robustness of Large Language Models in Conversational Artificial Intelligence for Individuals With Depression: A Critical Analysis

Andrea Ferrario[1,2,*], PhD; Jana Sedlakova[1,3,4,*], MA; Manuel Trachsel[5,6,7], MD, PhD

1
2
3
4
5
6
7

[*]these authors contributed equally

Corresponding Author:
Andrea Ferrario, PhD

Abstract

Large language model (LLM)–powered services are gaining popularity in various applications due to their exceptional performance in many tasks, such as sentiment analysis and answering questions. Recently, research has been exploring their potential use in digital health contexts, particularly in the mental health domain. However, implementing LLM-enhanced conversational artificial intelligence (CAI) presents significant ethical, technical, and clinical challenges. In this viewpoint paper, we discuss 2 challenges that affect the use of LLM-enhanced CAI for individuals with mental health issues, focusing on the use case of patients with depression: the tendency to humanize LLM-enhanced CAI and their lack of contextualized robustness. Our approach is interdisciplinary, relying on considerations from philosophy, psychology, and computer science. We argue that the humanization of LLM-enhanced CAI hinges on the reflection of what it means to simulate "human-like" features with LLMs and what role these systems should play in interactions with humans. Further, ensuring the contextualization of the robustness of LLMs requires considering the specificities of language production in individuals with depression, as well as its evolution over time. Finally, we provide a series of recommendations to foster the responsible design and deployment of LLM-enhanced CAI for the therapeutic support of individuals with depression.

(JMIR Ment Health 2024;11:e56569) doi:10.2196/56569

KEYWORDS

generative AI; large language models; large language model; LLM; LLMs; machine learning; ML; natural language processing; NLP; deep learning; depression; mental health; mental illness; mental disease; mental diseases; mental illnesses; artificial intelligence; AI; digital health; digital technology; digital intervention; digital interventions; ethics

Introduction

What Are Large Language Models?

Large language models (LLMs) are a type of generative artificial intelligence (AI) that displays unprecedented performance in different downstream tasks, such as question answering and, in general, context-aware text generation [1-4]. They produce language using deep neural networks. These models consist of billions of parameters and are trained on huge amounts of data at the expense of notable computational power. LLMs have been recently popularized by services—such as OpenAI's ChatGPT-4, Google's BARD (now called "Gemini"), and Meta's Llama—that are currently used by millions of people every day, experts, and laypeople alike. These services are essentially conversational AI (CAI) enhanced with LLMs. They offer a more human-like, natural, and context-relevant interaction than other technological applications such as rule-based conversational agents (ie, traditional "chatbots"). They hold the potential to transform how we engage in conversations and manage the information therein. Consequently, they are expected to become much more widely adopted in different professional fields, research, and society alike.

LLMs in Health Care and Mental Health Applications

In health care, applications of LLMs are manifold, spanning from clinical research and processes to physician-patient relations [5-10]. For instance, LLMs can improve clinical processes by automating the generation of administrative text

[1,5]. Physician-patient relations could benefit from the use of LLM-enhanced patient decision aids and interventions that could support therapy and improve shared decision-making [1]. Context-relevant and personalized conversations with an LLM-enhanced CAI show the potential to promote patients' empowerment and individuals' reflection around their personal values and preferences for different health care scenarios in a way that is not possible with current methods, for example, filling out legal documents such as advance directives [11-13].

In the mental health domain, the use of CAI is no novelty. The very first chatbot ELIZA, was developed in 1966, and played the role of a digital psychotherapist [14]. Six decades later, it is possible to develop and test LLM-enhanced CAI leveraging an ample body of knowledge and use cases. In the mental health domain, CAIs are currently used as patient therapeutic support, for example, a simple psychotherapy, such as cognitive behavioral exercises [15]. Given their noteworthy ability to process and produce language, the use of LLMs holds the potential to provide more context-aware and effective psychotherapeutic support to their users than traditional CAI. In fact, once embedded in CAI, designers can instruct LLMs to provide a nonjudgmental, readily available platform for vulnerable individuals to discuss their feelings and mental health struggles as well as practice skills that they learned in a therapeutic session.

CAI is also used to collect data of patients with mental health disorders, carry out initial triage processes, and provide treatment recommendations [6]. Here, LLM-enhanced CAI could process written or spoken responses of patients with mental health disorders to support therapists in their diagnostics or track mental health changes in patients over time. They could also generate personalized treatment recommendations by taking an individual's mental health history, their symptoms, values, and care preferences as input. By collecting data on the web, such as social media posts or chat logs, LLMs could help detect signs of emotional distress and detect mental health issues promptly. To this end, preliminary results show that ChatGPT-3.5 achieves good performance at detecting stress and depression in written statements on web-based forums [16]. These results suggest that ChatGPT-3.5–enhanced CAI could be in the future used in psychotherapeutic scenarios, for example, among patients with depression [16].

Ethical, Technical, and Clinical Risks Posed by LLMs

Emerging technology, including LLMs, are not immune to risks. They stem from conceptual, ethical, technical, and clinical considerations and become especially important in domains such as mental health care. They comprise ethical issues linked to bias, global digital divide, trustworthiness, black-box nature and validation, and the generalizability challenges in training and deploying LLMs [1,4,17].

In addition, authors are addressing different ethical problems associated with the use of LLMs in mental health applications. For instance, Cabrera et al [18] relate these challenges to the 4 principles of biomedical ethics with emphasis on data privacy, confidentiality, avoiding manipulation, and safety. Similarly, Yang et al [19] emphasize the necessity of carefully evaluating LLM-enhanced CAI in mental health applications by advocating the use of explainability methods to make the outcomes of LLMs more transparent. They also suggest complementing the use of LLMs with additional sources of information, such as emotional cues and cause-effect reasoning to enhance the quality of mental health support [19]. Further, Thirunavukarasu et al [1] emphasized the importance of using domain-specific data to fine-tune LLMs to validate LLM-enhanced applications with real clinical use cases. Finally, research is also investigating these challenges, with a focus on the perspective of end users of these systems, for example, individuals engaged in digitally assisted therapies. To this end, Weidinger et al [4] identify 6 areas of risk and potential harm to users of LLMs, including issues such as discrimination, privacy risks, misinformation, and human-computer interaction challenges [4]. In particular, they emphasize the risks for the users of LLM-enhanced services, which stem from their "human-like" design.

Mental Health Use Case: LLM-Enhanced CAI to Support Individuals With Depression

In summary, regarding the use of LLMs, research needs to address a mix of familiar and novel conceptual, ethical, technical, and clinical issues. To improve our understanding of these challenges, we need to examine LLMs within specific domains. This approach becomes particularly pertinent in the mental health domain, where the high sensitivity of the use cases underscores the imperative for a responsible and effective implementation of LLMs in CAI systems that provide therapeutic support to vulnerable individuals.

In this work, we focus on the scenario where LLM-enhanced CAI systems are used in the mental health domain to promote therapeutic support focusing on individuals with depression. The rationale behind selecting this use case is as follows. First, depression affects over 300 million people and the World Health Organization identifies it as the largest single contributor to global disability [20]. From an economic standpoint, for instance, studies have estimated the economic impact of depression to be 1.6% of the US gross domestic product [21]. Further, burnout is a major issue among psychiatrists [22]. Then, it is imperative to integrate technology-mediated interventions alongside traditional therapy methods to enhance accessibility and effectiveness of mental health services. Here, the use of CAI for patients with depression is widespread and supported by an ample body of scientific evidence [15]. More recently, research has also started exploring the use of LLMs for addressing depression [3,23].

This said, it is still an open avenue of research to delineate the perimeter for the responsible use of LLMs in scenarios involving individuals with depression. Therefore, in this work, we contribute to research on the responsible design, integration, and use of generative AI in mental health by focusing on 2 challenges that affect all scenarios where individuals with depression interact with LLM-enhanced CAI. In particular, we address challenges that pertain the (1) humanization of LLM-enhanced CAI (philosophy and psychology) and (2) contextualization of the robustness desideratum (computer science).

Our approach is interdisciplinary and relies on theories and methods from philosophy, ethics, psychology, and computer science. Our aim is to conceptually analyze 2 topics that are underexplored in the literature on LLMs and their applications in mental health care despite their importance while highlighting their risks. With our analysis, we provide a critical perspective on the CAI-specific trend of humanizing CAI and the problem of treating the robustness of LLM-based CAI systems as a context-independent challenge. We cross the boundaries of the disciplines to show how our conceptual analysis can be informative for issues in computer science and health research [24]. In fact, when properly translated, our conceptual analysis can generate valuable insights in empirical disciplines [25]. Finally, we discuss recommendations to promote the responsible use of LLM-enhanced CAI in the mental health domain.

Use of LLM-Enhanced CAIs by Individuals With Depression: 2 Challenges

Humanizing LLM-Enhanced CAI: a Philosophical and Psychological Perspective

Humanization is intended to develop CAI with the goal of simulating human abilities and traits, such as humor, empathy, and politeness. It is different from anthropomorphism, which refers to users' tendency to attribute CAI with human abilities and traits [26,27], although the CAI does need to be intentionally designed to mimic humans [26]. In practice, humanization of CAI is achieved by developing verbal, nonverbal, visual, and relational cues to make the system more human-like [28,29]. For example, CAI can have a persona (relational cues) that simulates certain human personalities, such as being a friend or therapist. This persona is often implemented in the avatar (visual cues), the informal language (linguistic cues), and emojis (nonverbal cues) used by the CAI. Empirical studies suggest that the humanized abilities and characteristics of CAI, such as reciprocity or giving empathetic responses, has positive outcomes on digital health interactions, such as improved user experience and the formation of relationships with CAI, trust, or better engagement [27,30-33]. These outcomes are particularly relevant in mental health care applications, where therapeutic relationships with the CAI promote therapeutic effectiveness, and high levels of user engagement might limit drop-outs [34,35]. However, research studies lack consistency in conceptualizing humanization and manipulating different cues [26]. In fact, there is no systematic inquiry to understand the extent to which specific cues lead to specific outcomes. More investigations are needed to understand the underlying mechanisms of measured effects (eg, linguistic vs nonlinguistic cues) and assess how these differ on the basis of design choices of humanized CAI.

LLMs can simulate context-aware conversations with their users and demonstrate an ability to adopt conversational personas. This allows LLM-enhanced CAI displaying features that strongly resemble human abilities and characteristics to an unprecedented level [36]. As noted by Shanahan et al [36], LLMs are fundamentally dialogue agents that role-play an ample variety of human-like characters [36]. While there is a generally positive view of humanizing CAI in various domains, we argue for a critical view of humanization efforts. Particularly, the effort of humanizing LLM-enhanced CAI in mental health applications presents serious challenges that must be tackled. Research highlights concerns about the safety of vulnerable users interacting with "human-like" systems [4]. This perspective on humanization emphasizes the potential risks and challenges stemming from the interactions between LLM-enhanced CAI and users, such as individuals with depression. However, there appears to be a lack of theoretical perspectives and clarification on humanization although humanizing concepts are fundamentally rooted in describing and developing AI systems [27,37]. This theoretical clarification could inform the responsible development of these systems, particularly in mental health applications. In what follows, we address this gap by relying on philosophical, ethical, and psychological considerations.

Conceptual Considerations

First, a conceptual problem underlies the development of LLM-enhanced and "humanized" CAI. We argue that it is important to maintain a distinction between the characteristics and traits simulated by these systems and the human qualities that are referred to using the same concepts and terminology. Simulated abilities and characteristics of LLM-based CAI are not the same as the original human abilities and characteristics. There are fundamental differences between humans and AI that further problematize an uncritical adoption of human concepts in the context of CAI. These problems have been addressed in different research domains. For instance, Bender et al [38] focus on the difference between synthetic language produced by LLM and human natural language by arguing that LLMs are "stochastic parrots" producing language, but not understanding it. Felin and Holweg [39] similarly argue by reporting differences in human cognition and computation processes of AI. Such arguments are often based on linguistic, philosophical, and psychological knowledge about human cognition, understanding and belief systems that are based on meaning, intentions, theory-based logic, and experience and are embedded in social and normative space [39-43]. In philosophy, the argumentation can stem from the analysis of such concepts as rational and moral agency that are not present in CAI, but are inherent in humans and their activities such as conversations [43]. Another strategy could be to analyze CAI as a different system from humans and by showing the limits of their models that cannot reach the complexity of human intelligence as reported by Landgrebe and Smith [44]. All these considerations have in common the fact that they provide a diversity of arguments for the position that CAI's simulated abilities and characteristics differ from humans [45,46]. In line with this literature, we argue for careful descriptions of CAI when human concepts are used. Such human concepts and terms such as being genuinely "empathetic," "compassionate," "inclusive," "polite," or "authoritative" mean something different when applied to CAI. If possible, CAI should be described more appropriately to avoid misconceptions and conceptual confusion. In the next subsection, we will outline problems and risks that might stem from such misconceptions and conceptual confusion.

In mental health literature, we found specific examples criticizing the adoption of human concepts for LLM-enhanced

CAI. A good and common example is "empathy," which is a key component of psychotherapy [47,48]. Recently, researchers investigated the simulation of an LLM-based "empathetic therapist" with individuals with depression [16]. The fact that an LLM-enhanced CAI can generate a seemingly empathetic response is substantially different from a human actually expressing empathy [49]. This ability is linked with someone's personality and emotional profile, shared social space, and lived experiences [47]. To be empathetic means to achieve genuine *understanding* of what another person is experiencing or attempting to express. Empathy includes active listening, asking targeted questions, and expressing genuine concern effectively addressing emotional needs [47]. These activities lie beyond the capabilities of LLMs, which are disembodied statistical processes. Most importantly, LLMs do not understand users' inputs and, in particular, do not understand their semantics [50,51], despite representing a vast body of information in a neural network.

Here, understanding (eg, a statement) is a crucial epistemic accomplishment arising from a myriad of complex cognitive activities that result in grasping meaning (eg, of statements and their components) and causal relationships, testing alternative knowledge pathways, on top of providing well-grounded reasons for each of those. Furthermore, understanding emerges as the culmination of intricate processes that are socially and normatively embedded [52]. This attainment is fostered by virtues, such as perseverance, precision, and epistemic humility among others. These characterize, in particular, how human experts in a research domain structure knowledge and seek understanding. In contrast, LLMs compute answers through statistical processes that simply do not take into account the meaning of user's prompts [39]. As a result, understanding escapes the statistical manipulations that characterize the logic of LLMs [53,54]. In a nutshell, displaying—sometimes successful, as LLMs do hallucinate and generate "fake" references and justifications—ability to manipulate structured information does not guarantee understanding.

Vulnerable patients with depression may potentially misinterpret CAI as empathetic and caring, potentially leading to unrealistic expectations such as warmth and acceptance [55]. Due to CAI's limitations, such misconceptions could reinforce negative beliefs and worsen emotional states. Since LLMs lack understanding of user inputs, they may respond inappropriately, misunderstanding the nuances of individual situations. This could further reinforce negative feelings or isolation in patients with depression. This point is particularly relevant for designers and therapists who need to test the capabilities of LLMs before promoting their use for digital therapy with vulnerable individuals. Differently from current research [4], we emphasize that humanization is at first a challenge for those who design and promote these systems, before becoming a risk for those who use the technology. The key here is to understand that the ability of LLMs to generate empathetic output, as opposed to being apathetic, indifferent, and insensitive in conversations, descends from the computation of empirical probabilities of "next words," given the user prompt and their training on a massive amount of documents [36]. In fact, under the hood, LLMs perform autocomplete functions of search engines [50].

These remarks help in characterizing the limits of the humanization of LLMs and they hold true also for other characteristics and traits that LLMs attempt to simulate. This includes, in particular, the quality of being an "expert" in a domain, for example, a specialist in the treatment of depression among adolescents, and, in virtue of this, being perceived as a digital therapist, instead of a therapeutic support system [53-55].

In summary, philosophy and psychology guide us in recognizing the substantive differences between humans and humanized LLM-enhanced CAI. This helps to assess the limits of this endeavor, identify the correct roles these systems can play in interactions with humans, and, eventually mitigate misconceptions and overtrust in these systems [4]. The issue of humanization needs more in-depth analysis, including the exploration of how the human attributes assigned to LLM-enhanced CAI influence and guide patients in shaping their behavior and responses within a conversation.

Normative and Ethical Implications of the Conceptual Problem

The conceptual confusion of ascribing human-like abilities to LLM-enhanced CAI is linked with important normative and ethical risks, which pertain to responsibility, commitments, and rights. Overall, interpersonal conversations are social and normative activities that are embedded in a set of values, norms, and virtues [42]. This is particularly true in the case of therapeutic relationships that are guided by sets of values and norms to ensure a safe environment and therapeutic process for patients [56-59]. Such human abilities as empathy or understanding are part of this normative and professional setting. Psychiatrists and psychotherapists who do not follow professional conduct guidelines when treating individuals with depression risk causing medical emergencies for their patients—a situation that could lead to disciplinary actions against them.

In the case of humanized LLM-enhanced CAI, there is a gap between what the system appears to be, for example, being compassionate, and what normative criteria this ability should meet and cannot be met by CAI—criteria that are fulfilled by human therapists instead. Hence, when CAI simulates abilities such as empathy or understanding, these are not part of the normative setting as they are in the case of human experts. This CAI can lead to risks among individuals with depression. For instance, if an LLM's response lacks compassion during a conversation with a user with depression, this may worsen their condition, even leading to self-harm. An LLM-enhanced CAI may not encode cultural nuances and the uniqueness of individual experiences in its outputs while its biases significantly influence how the system presents and discusses knowledge with patients. This can contribute to "epistemic injustice" [60], making individuals with depression potentially feel more isolated and their perspectives undervalued and misunderstood. In addition, human experts—for instance, psychiatrists—have epistemic duties, including being truthful and justifying their beliefs [43]. In contrast, LLMs lack these commitments [40].

Further, this "normativity gap" leads to a problem of assigning responsibility and defining how to approach failures in a conversation with patients. There is a difference between addressing the ethical consequences of technical failures of a computer system, for example, numerical errors and inaccurate predictions, and dealing with the issues that arise from a faulty implementation of humanizing features. On the one hand, technical errors in computer systems are clearly defined, objectively measured, and traced, facilitating the definition of their sanctions. On the other hand, what does it mean that the LLM-enhanced CAI was not empathetic in a given conversation? Was it not empathetic *enough*? According to which objective measures of empathy? Did the lack of empathy persist in the conversation long enough to consider applying sanctions? The humanization of LLM-enhanced CAI involves complexities that are not fully understood even in interpersonal interactions, where ambiguous, inappropriate or unprofessional questions and answers may occur, and the applicability of sanctions is unclear.

In summary, despite the current trend of humanizing LLM-enhanced CAI, it is questionable to what extent such humanization is necessary and helpful as it poses theoretical and ethical challenges. It remains an open question whether there is an ethically acceptable, safe, and beneficial degree of humanization for these systems. Philosophy and psychology can help frame the problem, which highlights a particularly important gap of the responsible design and development of AI in mental health care [61].

Contextualizing the Robustness of LLMs Used by Individuals With Depression: a Computer Science Perspective

The Robustness of LLMs

Robustness refers to the ability of machine learning models to withstand "perturbations" that may affect their performance [62-64]. It is a general model capability that becomes essential for ensuring the reliability of machine learning models in real-world applications. Interestingly, robustness is a multidimensional concept that is currently lacking a one-size-fits-all definition. Rather, research discusses what a robust model *should do* [62,65-68], investigating how a model should resist different types of perturbations, such as those affecting its input data, data distributions over time, and the model structure. In fact, a robust machine learning model computes predictions that do not vary disproportionately in case of perturbed inputs. Further, it retains accuracy in the presence of distributional shift [69] and is not affected by small changes in its constitutive structure. In summary, robustness is a key requirement for trustworthy AI. It can also be extended to comprise algorithms that provide explanations of machine learning models' predictions [65,66,68,70]. In this case, robust explanations are not altered by the perturbation of data inputs and are stable over time.

In the case of LLMs, the high-level desideratum of robustness seems to gain an extra level of complexity [63,64]. In fact, when discussing what robust LLMs should do, we need to consider the peculiar way these models compute their predictions, namely, using prompt-based queries [71]. Here, a prompt is structured information—often, a text snippet—that users offer as an instruction to the LLM and which is often accompanied by one or more examples to guide the model ("in-context learning" or "few-shot prompting" procedure) [71,72]. For example, a prompt for an LLM used in an application to investigate how patients with depression communicate with CAI may look like this: "Classify the following sentence in either normal or alerting: [s]." Here, the example [s] is the patient's utterance: "Today, I felt more useless than usual and nobody knows it."

Broadly speaking, an LLM is robust if its predictions display an appropriate level of sensitivity to the changes that may affect its prompts and examples. With a robust LLM, similar prompts and examples should lead to similar predictions, among others. This said, research has a long way to go before this promise can become reality. An increasing body of literature shows that commonly available LLMs, for example, T5, Vicuna, Llama 2, and ChatGPT-3.5 [73], generally display a low level of robustness. These models are highly sensitive to different types of perturbations, named "prompt-targeting adversarial attacks" [73]. These comprise switching the order of few-shot examples and semantic-preserving variations, such as adding a few typographical errors, replacing words by synonyms or back translating the prompt itself and its examples [73]. As a result, a few empirical studies show that prompt-targeting adversarial attacks can lead to substantially different LLM predictions, indicating an overall lack of robustness across a variety of downstream tasks, such as text classification and generation [73].

Finally, from an ethical perspective, the lack of robustness of LLMs is a source of different issues. Nonrobust models lead to unreliable decision-making, that is, they increase the risk of making inconsistent or erroneous decisions that can harm those affected by them. For instance, LLMs could provide misdiagnosis and share information that does not align with clinical practices, show the inability to detect and respond to nuances in language that indicate a mental health crisis (such as expressions of suicidal ideation or severe distress), and offer appropriate and timely crisis intervention resources. Finally, training on large corpora of text may lead LLMs to perpetuate forms of stigmatization against individuals affected by mental health issues (despite fine-tuning on documents from the psychiatry domain).

They may also lead to unwanted cases of bias and discrimination and pose serious concern to the privacy of individuals' information. Nonrobust models can be tricked to reveal personal information. Finally, erratic or nonrobust model behavior affects their overall transparency levels. These ethical concerns are particularly relevant in high-stakes scenarios, such as those where LLMs are deployed to support the mental health of vulnerable individuals.

Contextualizing the Robustness of LLMs

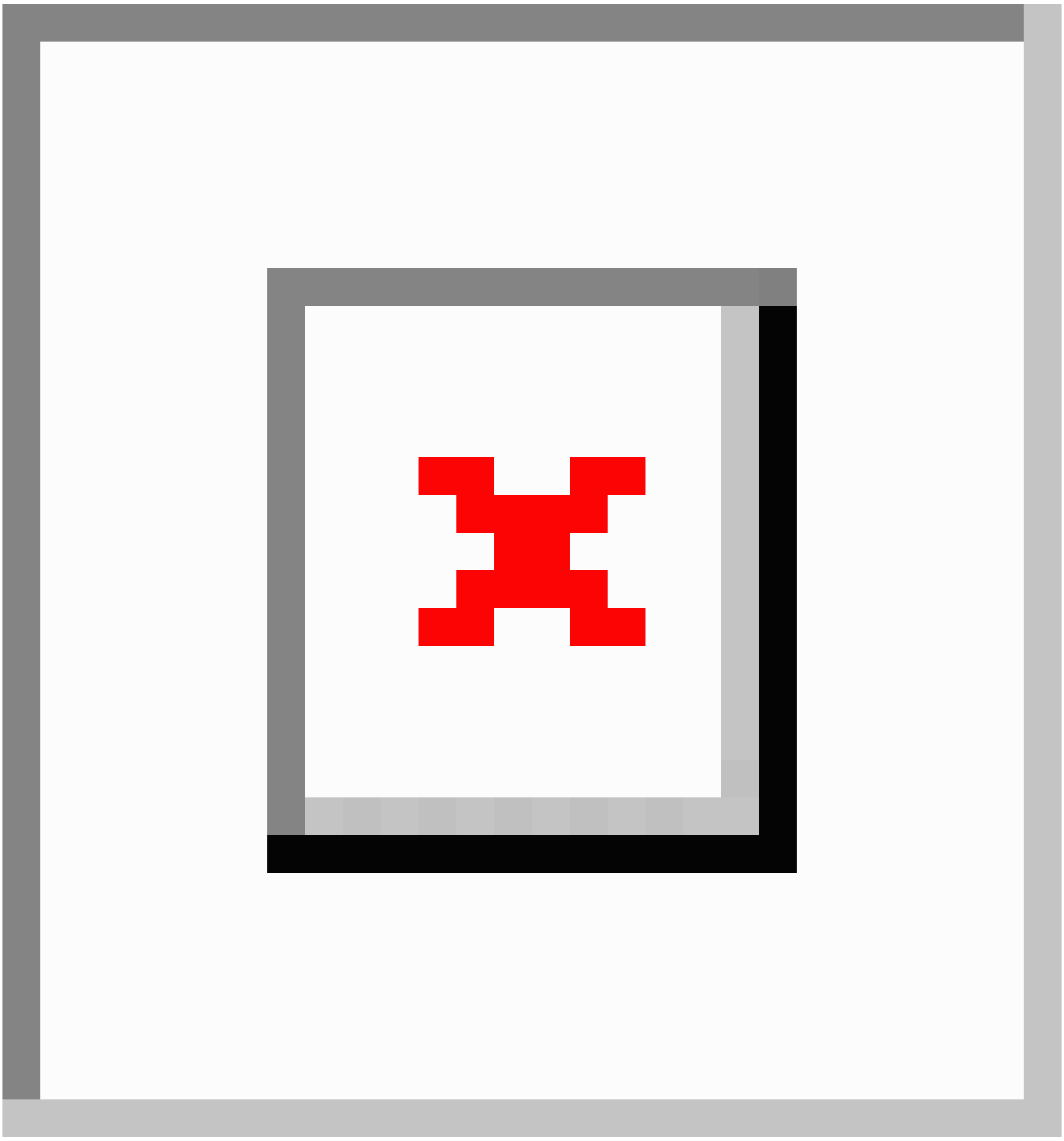
Current approaches to ensuring the robustness of LLMs lack proper contextualization: they are not targeted to any specific scenario of human-LLM interaction. While it is beneficial that LLM predictions remain consistent even when prompted in similar ways or when the order of the LLM examples changed, as suggested by the emerging literature on prompt-targeting adversarial attacks [73], this alone is insufficient for the ethically responsible use of LLMs in high-risk applications, such as in scenarios involving mental health support for patients with depression. In these cases, we argue that it is necessary that LLMs' robustness is tailored to align with the specific language characteristics—and their variations over time—of the model users, specifically, patients with depression. In other words, an appropriately robust LLM to be used by individuals with depression should detect their lexical, syntactic, cultural, and content-related language patterns, while retaining the ability of not being affected by more general adversarial attacks [73], as suggested by the high-level desideratum of robustness. In summary, the LLM should provide accurate outputs that are (1) not affected by spurious linguistic variations in the prompts and examples provided by its users and (2) tailored to the context in which the interaction takes place. This calls for the design of prompt-targeting *contextualized* adversarial attacks and the assessment of the *contextualized* robustness of LLMs, rather than the investigation of general, domain-unspecific robustness constraints.

Research on depression has already identified a few linguistic patterns that may help in this regard. On average, patients with depression make more and longer pauses than healthy individuals when they communicate [74]. Further, they display a lower pitch, more monotonous speech, and slower utterance production [75,76] (notably, the importance of slowed speech is emphasized in the Patient Health Questionnaire–9 self-assessed depression report [76]). Similarly, the analysis of transcripts of utterances of individuals with depression shows that patients with depression use more modifying adverbs, first-person and personal pronouns, and more verbal utterances [76-78]. Further, individuals with depression and healthy individuals show differences in the use of past tense, causation, achievement, and death words [76], using simpler sentence structures and reduced linguistic complexity, as well as exhibiting rumination and self-focus in their language [79,80]. The instability of words associated with negative emotion predicts depression in textual production on social media as well [81]. We also note that these patterns may vary over time, as the patient may go through different stages of depression. In Figure 1, we show a few examples of such variations we generated with ChatGPT-4. In summary, robust LLMs to be used for therapy for individuals with depression should be able to correctly identify their linguistic patterns and react to their evolution appropriately. This observation is reinforced by the fact that language is a dynamic process that changes over time. New idioms, metaphors, or shifts in meaning regularly take place, and LLMs need to be aligned with the dynamics of language production. Here, the risk is to promote "frozen" narratives and linguistic patterns that do not reflect the evolution of patient narratives over time.

From a technical perspective, making LLMs contextually robust requires their fine-tuning them with high-quality, curated data. Currently, obtaining such data for patients with depression is challenging, with most available examples coming from social media platforms such as Twitter or Reddit [82]. To the best of our knowledge, there exists no publicly available data set of conversations of patients with depression from consultations with therapists, therapeutic CAI, or other agents in everyday life. Additionally, we cannot easily improve proprietary LLMs. This is a serious problem, as a recent study shows that ChatGPT-3.5 is not robust enough for conversations with individuals with symptoms of anxiety or depression, as the LLMs suggested medications to its users; medications should be taken under the guidance of a psychiatrist [83].

In summary, understanding and achieving the contextualized robustness of LLMs is crucial for the responsible use of LLM-enhanced CAI among individuals with depression. While computer science offers methodologies to formalize, evaluate, and satisfy this requirement, their effectiveness is limited by the availability of necessary resources, primarily therapy-relevant data, which are currently lacking.

Figure 1. Examples of utterances by patients with mild depression and those with severe depression generated by ChatGPT-4 (prompt and answers from December 2023).



Toward Responsible Use of LLMs in Therapeutic Settings Involving Individuals With Depression

The complexities of humanization and contextualized robustness appear to temper the initial enthusiasm surrounding LLMs. The problems affecting humanization we discussed in the previous sections seem to be at odds with the very essence of LLMs, namely, to role-play different personas. Meanwhile, we noted that achieving contextualized robustness requires thorough fine-tuning and comprehensive testing. Moreover, this process must be grounded in a deep understanding of how language production and usage evolve over time among the users of these systems.

The importance of addressing the risks associated with humanization and the absence of contextual robustness is underscored by real-world incidents involving individuals with depression using LLM-enhanced CAI. There have been several instances, reported in various media, where LLM-enhanced CAI provided support for mental health issues but instead encouraged self-harm or offered detrimental advice [84,85]. A Belgian man with depression committed suicide following conversations with ChatGPT-3.5 [85]. Recently, Kumar et al [86] commented on the case of a user with depression, who, during a crisis, was able to insert a sequence of words in their prompt that bypassed the LLM’s safety-guards and generated harmful content [86]. In fact, the LLM returned detailed instructions on how to commit different types of self-harm [86]. Further, authors show that certain prompts result in ChatGPT-3.5 prescribing medications to individuals with anxiety or depression symptoms, despite medications that should be taken under the guidance of a therapist [83]. In addition, the vulnerability of LLM-enhanced CAI to attacks and content manipulation can lead to the generation of offensive, inappropriate, or objectionable responses; the provision of incorrect information; and discriminatory recommendations. These events show potential of causing either discomfort, harm, or even acute detriment to users [87].

Finally, it is argued that humanization may invite and actively nudge patients to react to its cues [4]. LLM-enhanced CAI are persuasive to their users and can perform a variety of emotional manipulations. These may lead to inappropriate reliance on these systems or overtrusting them [88], reinforcing bias, and overestimating their capabilities, including expecting unrealistic behavioral change [4,89].

Given the challenges discussed in this viewpoint paper, the path toward a responsible development and use of LLM-enhanced CAI in therapeutic settings involving individuals with depression appears to be quite challenging. Here, we agree with Cheng et al [90], who promote the idea of using LLM-enhanced CAI as an assistant to mental health professionals in providing patient care [90]. Further, they emphasize the need for routine monitoring of patients and the systems to address emerging challenges in a timely manner [90]. However, we disagree with the authors when they suggest that, from an ethical standpoint, psychiatrists should take full responsibility for any detriment to patients interacting with the LLM-enhanced CAI [90]. In fact, this claim would be justified if psychiatrists could understand these systems in depth. However, it is unlikely that psychiatrists, despite their expertise in mental health, would possess an in-depth understanding of the workings of such advanced technology.

In summary, an interdisciplinary approach to the responsible use of LLM-enhanced CAI in therapeutic settings involving users with depression is essential, encompassing both the social and technological aspects of CAI development and application [46,91]. This approach should integrate theoretical and practical perspectives from psychiatry, ethics, philosophy, computer science, and user experience design, ensuring a balanced and informed development of these technologies. These perspectives could help address the risks posed by the humanization of these systems and the lack of contextualized robustness, by suggesting ways to inform, instruct, and educate developers and users (including therapists) about the conceptual nuances of normative concepts, such as expertise, and the characteristics of language production of individuals with depression.

One practical measure to manage the risks stemming from the humanization of LLM-enhanced CAI could be incorporating disclaimers and a short conversation at the start of therapy sessions with the system. The measures would outline the capabilities and theoretical limitations of CAI, helping users in accurately setting their expectations from the interaction with the systems. Revisiting these disclaimers and conversations periodically, especially in long-term use, could reinforce users' understanding and help them manage their expectations effectively over time.

To contextualize the robustness of LLM-enhanced CAI, researchers could collect data from different cohorts of patients with depression interacting with the system in controlled settings. They could augment these data by other sources, including survey data and clinical information to improve the accuracy of the LLMs. Further, identifying contextual features that help LLMs recognize the patients' emotional states, triggers, or history can further improve the accuracy and contextual robustness of the models over time. These features may include over time sentiment analysis, trigger recognition, environmental information, and audio and visual cues. Therapists and patients could review these interactions to correct inaccurate suggestions and address the issues they may have caused in a timely manner. This procedure, which necessitates the active involvement of both clinical experts and patients, is undeniably time-consuming but indispensable. Moreover, it hinges on a controlled setting that may not capture all aspects of the interactions between patients with depression and LLM-enhanced CAI in everyday life. However, this is a first step to assess the risk of deploying "brittle" LLMs in clinical practice.

Finally, to responsibly use LLM-enhanced CAI with patients with depression, it is important to rigorously examine its long-term effects. Developing and adhering to strict standards for the creation and implementation of these systems is necessary, mirroring the evidence-based approach of mental health care, where interventions undergo thorough testing, including randomized controlled trials. A structured framework, akin to those used in the development and assessment of patient decision-making tools [92,93], could greatly benefit the development and application of LLM-enhanced CAI. Guidelines that address the humanization of these systems and ensure their contextual robustness should be central to this framework.

Ethical Considerations

This study was exempt from ethical review as no human participants were involved.

Conclusions

The use of LLMs in mental health applications presents numerous conceptual, ethical, technical, and challenges. In this work, we have outlined 2 challenges that impede the responsible use of LLMs in applications involving patients with depression: the accentuation of human-like qualities of LLM-enhanced CAI and the lack of contextualized robustness. These challenges warrant comprehensive consideration and a proactive approach to ensure the responsible and effective integration of LLMs in mental health settings. While human-like qualities may enhance user engagement, it is imperative to strike a balance when a simulation of human characteristics and abilities does not increase ethical risks and their effects are well understood. A responsible approach involves clearly communicating to users that they are interacting with AI-based tools and what this exactly means, enabling them to make informed decisions about the assistance they receive and being aware of their limitations as well as differences from human conversation.

Further, LLMs should be adept at understanding and adapting to the specific linguistic, cultural, and emotional nuances of individuals dealing with mental health issues. Robustness, in this context, involves not only maintaining coherence in responses but also sensitively addressing the unique needs of each user. Ethical guidelines should emphasize the development and validation of LLMs with a focus on contextual sensitivity. It is vital to establish a framework that delineates the roles of AI developers, health care providers, and users in ensuring the well-being of those seeking mental health support.

Authors' Contributions

AF conceptualized the research. AF and JS wrote the first draft of the manuscript. MT provided inputs on depression and the use of conversational artificial intelligence in therapy for patients with depression. AF and JS finalized the manuscript. All authors approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
2. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023 Jun 1;183(6):589. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
3. Galatzer-Levy IR, McDuff D, Natarajan V, Karthikesalingam A, Malgaroli M. The capability of large language models to measure psychiatric functioning. arXiv. Preprint posted online on Aug 3, 2023. [doi: 10.48550/arXiv.2308.01834]
4. Weidinger L, Uesato J, Rauh M, et al. Taxonomy of risks posed by language models. Presented at: FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency; Jun 21 to 24, 2022; Seoul, Republic of Korea. [doi: 10.1145/3531146.3533088]
5. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. Commun Med 2023 Oct 10;3(1):141. [doi: 10.1038/s43856-023-00370-1] [Medline: 37816837]
6. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med 2023 Jul 6;6(1):120. [doi: 10.1038/s41746-023-00873-0] [Medline: 37414860]
7. Peng C, Yang X, Chen A, et al. A study of generative large language model for medical research and healthcare. NPJ Digit Med 2023 Nov 16;6(1):210. [doi: 10.1038/s41746-023-00958-w] [Medline: 37973919]
8. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: development, applications, and challenges. Health Care Science 2023 Aug;2(4):255-263. [doi: 10.1002/hcs2.61]
9. Hua Y, Liu F, Yang K, et al. Large language models in mental health care: a scoping review. arXiv. Preprint posted online on Jan 1, 2024. [doi: 10.48550/arXiv.2401.02984]
10. Stade EC, Stirman SW, Ungar LH, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. Npj Ment Health Res 2024 Apr 2;3(1):12. [doi: 10.1038/s44184-024-00056-z] [Medline: 38609507]
11. Ferrario A, Gloeckler S, Biller-Andorno N. Ethics of the algorithmic prediction of goal of care preferences: from theory to practice. J Med Ethics 2023 Mar;49(3):165-174. [doi: 10.1136/jme-2022-108371] [Medline: 36347603]
12. Gloeckler S, Ferrario A, Biller-Andorno N. An ethical framework for incorporating digital technology into advance directives: promoting informed advance decision making in healthcare. Yale J Biol Med 2022 Sep;95(3):349-353. [Medline: 36187419]
13. Earp BD, Porsdam Mann S, Allen J, et al. A personalized patient preference predictor for substituted judgments in healthcare: technically feasible and ethically desirable. Am J Bioeth 2024 Jul;24(7):13-26. [doi: 10.1080/15265161.2023.2296402] [Medline: 38226965]
14. Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. Commun ACM 1966 Jan;9(1):36-45. [doi: 10.1145/365153.365168]
15. He Y, Yang L, Qian C, et al. Conversational agent interventions for mental health problems: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res 2023 Apr 28;25:e43862. [doi: 10.2196/43862] [Medline: 37115595]
16. Chen S, Wu M, Zhu KQ, Lan K, Zhang Z, Cui L. LLM-empowered chatbots for psychiatrist and patient simulation: application and evaluation. arXiv. Preprint posted online on May 23, 2023. [doi: 10.48550/arXiv.2305.13614]
17. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. Lancet Digit Health 2023 Jun;5(6):e333-e335. [doi: 10.1016/S2589-7500(23)00083-3] [Medline: 37120418]
18. Cabrera J, Loyola MS, Magaña I, Rojas R. Ethical dilemmas, mental health, artificial intelligence, and LLM-based chatbots. In: Rojas I, Valenzuela O, Rojas Ruiz F, Herrera LJ, Ortuño F, editors. Bioinformatics and Biomedical Engineering: Springer Nature Switzerland; 2023:313-326. [doi: 10.1007/978-3-031-34960-7]

19. Yang K, Ji S, Zhang T, Xie Q, Kuang Z, Ananiadou S. Towards interpretable mental health analysis with large language models. Presented at: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; Dec 6 to 10, 2023; Singapore. [doi: 10.18653/v1/2023.emnlp-main.370]
20. Ferrari AJ, Charlson FJ, Norman RE, et al. The epidemiological modelling of major depressive disorder: application for the Global Burden of Disease Study 2010. PLoS One 2013 Jul;8(7):e69637. [doi: 10.1371/journal.pone.0069637] [Medline: 23922765]
21. Chodavadia P, Teo I, Poremski D, Fung DSS, Finkelstein EA. Prevalence and economic burden of depression and anxiety symptoms among Singaporean adults: results from a 2022 web panel. BMC Psychiatry 2023 Feb 14;23(1):104. [doi: 10.1186/s12888-023-04581-7] [Medline: 36782116]
22. Bykov KV, Zrazhevskaya IA, Topka EO, et al. Prevalence of burnout among psychiatrists: a systematic review and meta-analysis. J Affect Disord 2022 Jul;308:47-64. [doi: 10.1016/j.jad.2022.04.005] [Medline: 35398112]
23. Xu X, Yao B, Dong Y, et al. Mental-LLM: leveraging large language models for mental health prediction via online text data. arXiv. Preprint posted online on Jul 26, 2023. [doi: 10.48550/arXiv.2307.14385]
24. Choi BCK, Pak AWP. Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness. Clin Invest Med 2006 Dec;29(6):351-364. [Medline: 17330451]
25. Archibald MM, Lawless MT, de Plaza MAP, Kitson AL. How transdisciplinary research teams learn to do knowledge translation (KT), and how KT in turn impacts transdisciplinary research: a realist evaluation and longitudinal case study. Health Res Policy Syst 2023 Mar 21;21(1):20. [doi: 10.1186/s12961-023-00967-x] [Medline: 36944997]
26. Nass C, Moon Y. Machines and mindlessness: social responses to computers. J Soc Issues 2000 Jan;56(1):81-103. [doi: 10.1111/0022-4537.00153]
27. Li M, Suh A. Machinelike or humanlike? A literature review of anthropomorphism in AI-enabled technology. Presented at: 54th Hawaii International Conference on System Sciences (HICSS 2021); Jan 5, 2021; Kauai, Hawaii. [doi: 10.24251/HICSS.2021.493]
28. Bickmore TW, Picard RW. Establishing and maintaining long-term human-computer relationships. ACM Trans Comput-Hum Interact 2005 Jun;12(2):293-327. [doi: 10.1145/1067860.1067867]
29. Nißen M, Rüegger D, Stieger M, et al. The effects of health care chatbot personas with different social roles on the client-chatbot bond and usage intentions: development of a design codebook and web-based study. J Med Internet Res 2022 Apr 27;24(4):e32630. [doi: 10.2196/32630] [Medline: 35475761]
30. Araujo T. Living up to the chatbot hype: the influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. Comput Human Behav 2018 Aug;85:183-189. [doi: 10.1016/j.chb.2018.03.051]
31. Pereira J, Díaz Ó. Using health chatbots for behavior change: a mapping study. J Med Syst 2019 Apr 4;43(5):135. [doi: 10.1007/s10916-019-1237-1] [Medline: 30949846]
32. Stara V, Vera B, Bolliger D, et al. Usability and acceptance of the embodied conversational agent Anne by people with dementia and their caregivers: exploratory study in home environment settings. JMIR Mhealth Uhealth 2021 Jun 25;9(6):e25891. [doi: 10.2196/25891] [Medline: 34170256]
33. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. Front Digit Health 2022 Apr 11;4:847991. [doi: 10.3389/fdgth.2022.847991] [Medline: 35480848]
34. Ardito RB, Rabellino D. Therapeutic alliance and outcome of psychotherapy: historical excursus, measurements, and prospects for research. Front Psychol 2011 Oct 18;2:270. [doi: 10.3389/fpsyg.2011.00270] [Medline: 22028698]
35. Norcross JC, Lambert MJ, editors. Psychotherapy Relationships That Work: Volume 1: Evidence-Based Therapist Contributions (3 edn): Oxford University Press; 2019. [doi: 10.1093/med-psych/9780190843953.001.0001]
36. Shanahan M, McDonell K, Reynolds L. Role-play with large language models. Nature 2023 Nov;623(7987):493-498. [doi: 10.1038/s41586-023-06647-8] [Medline: 37938776]
37. Salles A, Evers K, Farisco M. Anthropomorphism in AI. AJOB Neurosci 2020;11(2):88-95. [doi: 10.1080/21507740.2020.1740350] [Medline: 32228388]
38. Bendig E, Erb B, Schulze-Thuesing L, Baumeister H. The next generation: chatbots in clinical psychology and psychotherapy to foster mental health – a scoping review. Verhaltenstherapie 2022;32(Suppl. 1):64-76. [doi: 10.1159/000501812]
39. Felin T, Holweg M. Theory is all you need: AI, human cognition, and decision making. SSRN. Preprint posted online on Apr 4, 2024. [doi: 10.2139/ssrn.4737265]
40. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? Presented at: 2021 ACM Conference on Fairness, Accountability, and Transparency Virtual Event; Mar 3 to 10, 2021; Virtual Event Canada. [doi: 10.1145/3442188.3445922]
41. Emsley R. ChatGPT: these are not hallucinations – they're fabrications and falsifications. Schizophrenia (Heidelb) 2023 Aug 19;9(1):52. [doi: 10.1038/s41537-023-00379-4] [Medline: 37598184]
42. Brandom R, McDowell J. Knowledge and the social articulation of the space of reasons. Philos Phenomen Res 1995 Dec;55(4):895. [doi: 10.2307/2108339]

43. Sedlakova J, Trachsel M. Conversational artificial intelligence in psychotherapy: a new therapeutic tool or agent? Am J Bioeth 2023 May;23(5):4-13. [doi: 10.1080/15265161.2022.2048739] [Medline: 35362368]
44. Landgrebe J, Smith B. Why Machines Will Never Rule the World: Artificial Intelligence Without Fear: Routledge; 2022. URL: https://www.taylorfrancis.com/books/9781003310105 [accessed 2024-06-19] [doi: 10.4324/9781003310105]
45. Boyle A. Disagreement & classification in comparative cognitive science. Noûs 2023 Oct 16. [doi: 10.1111/nous.12480]
46. Díaz-Rodríguez N, Del Ser J, Coeckelbergh M, de Prado ML, Herrera-Viedma E, Herrera F. Connecting the dots in trustworthy artificial intelligence: from AI principles, ethics, and key requirements to responsible AI systems and regulation. Information Fusion 2023 Nov;99:101896. [doi: 10.1016/j.inffus.2023.101896]
47. Elliott R, Bohart AC, Watson JC, Murphy D. Therapist empathy and client outcome: an updated meta-analysis. Psychotherapy (Chic) 2018 Dec;55(4):399-410. [doi: 10.1037/pst0000175] [Medline: 30335453]
48. Elliott R, Watson JC, Goldman RN, Greenberg LS. Learning Emotion-Focused Therapy: The Process-Experiential Approach to Change: American Psychological Association; 2003.
49. Montemayor C, Halpern J, Fairweather A. In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. AI & Soc 2022 Dec;37(4):1353-1359. [doi: 10.1007/s00146-021-01230-z] [Medline: 34054228]
50. Floridi L. AI as agency without intelligence: on ChatGPT, large language models, and other generative models. Philos Technol 2023 Mar;36(1):15. [doi: 10.1007/s13347-023-00621-y]
51. Floridi L, Chiriatti M. GPT-3: its nature, scope, limits, and consequences. Minds Mach 2020 Dec;30(4):681-694. [doi: 10.1007/s11023-020-09548-1]
52. Brandom RB. Reason in Philosophy: Animating Ideas: Belknap Press of Harvard University Press; 2009. [doi: 10.4159/9780674053618]
53. Ferrario A, Facchini A, Termine A. Experts or authorities? The strange case of the presumed epistemic superiority of artificial intelligence systems. SSRN. Preprint posted online on Sep 18, 2023. [doi: 10.2139/ssrn.4561425]
54. Ferrario A, Biller-Andorno N. Large language models in medical ethics: useful but not expert. J Med Ethics 2024 Jan 22:jme-2023-109770. [doi: 10.1136/jme-2023-109770] [Medline: 38253463]
55. Ferrario A, Termine A, Facchini A. Addressing social misattributions of large language models: an HCXAI-based approach. arXiv. Preprint posted online on Mar 26, 2024. [doi: 10.48550/arXiv.2403.17873]
56. Norcross JC, Lambert MJ. Psychotherapy relationships that work III. Psychotherapy 2018 Dec;55(4):303-315. [doi: 10.1037/pst0000193] [Medline: 30335448]
57. DeAngelis T. Better relationships with patients lead to better outcomes. American Psychological Association. 2019. URL: https://www.apa.org/monitor/2019/11/ce-corner-relationships [accessed 2024-06-19]
58. 2014 ACA Code of Ethics. American Counseling Association. 2014. URL: https://www.counseling.org/docs/default-source/ethics/2014-aca-code-of-ethics.pdf [accessed 2024-06-19]
59. Ethical principles of psychologists and code of conduct. American Psychological Association. 2017. URL: https://www.apa.org/ethics/code/ [accessed 2024-06-19]
60. Laacke S. Bias and epistemic injustice in conversational AI. Am J Bioethics 2023 May 4;23(5):46-48. [doi: 10.1080/15265161.2023.2191055] [Medline: 37130400]
61. Lin B, Bouneffouf D, Cecchi G, Varshney KR. Towards healthy AI: large language models need therapists too. arXiv. Preprint posted online on Apr 2, 2023. [doi: 10.48550/arXiv.2304.00416]
62. Freiesleben T, Grote T. Beyond generalization: a theory of robustness in machine learning. Synthese 2023 Sep 27;202(4):109. [doi: 10.1007/s11229-023-04334-9]
63. Wang J, Hu X, Hou W, et al. On the robustness of ChatGPT: an adversarial and out-of-distribution perspective. arXiv. Preprint posted online on Feb 22, 2023. [doi: 10.48550/arXiv.2302.12095]
64. Zhuo TY, Huang Y, Chen C, Xing Z. Red teaming ChatGPT via jailbreaking: bias, robustness, reliability and toxicity. arXiv. Preprint posted online on Jan 30, 2023. [doi: 10.48550/arXiv.2301.12867]
65. Hancox-Li L. Robustness in machine learning explanations: does it matter? Presented at: FAT* '20: Conference on Fairness, Accountability, and Transparency; Jan 27 to 30, 2020; Barcelona, Spain. [doi: 10.1145/3351095.3372836]
66. Ferrario A, Loi M. The robustness of counterfactual explanations over time. IEEE Access 2022 Aug;10:82736-82750. [doi: 10.1109/ACCESS.2022.3196917]
67. Athalye A, Engstrom L, Ilyas A, Kwok K. Synthetizing robust adversarial examples. Presented at: 35th International Conference on Machine Learning; Jul 10 to 15, 2018; Stockholm, Sweden.
68. Sharma S, Henderson J, Ghosh J. CERTIFAI: counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. arXiv. Preprint posted online on May 20, 2019. [doi: 10.48550/arXiv.1905.07857]
69. Žliobaitė I, Pechenizkiy M, Gama J. An overview of concept drift applications. In: Japkowicz N, Stefanowski J, editors. Big Data Analysis: New Algorithms for a New Society: Springer International Publishing; 2016:91-114. [doi: 10.1007/978-3-319-26989-4]
70. Alvarez-Melis D, Jaakkola TS. On the robustness of interpretability methods. arXiv. Preprint posted online on Jun 21, 2018. [doi: 10.48550/arXiv.1806.08049]
71. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv 2023 Sep 30;55(9):1-35. [doi: 10.1145/3560815]

72. Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. arXiv. Preprint posted online on Jun 15, 2022. [doi: 10.48550/arXiv.2206.07682]
73. Zhu K, Wang J, Zhou J, et al. PromptBench: towards evaluating the robustness of large language models on adversarial prompts. arXiv. Preprint posted online on Jun 7, 2023. [doi: 10.48550/arXiv.2306.04528]
74. Tan EJ, Neill E, Kleiner JL, Rossell SL. Depressive symptoms are specifically related to speech pauses in schizophrenia spectrum disorders. Psychiatry Res 2023 Mar;321:115079. [doi: 10.1016/j.psychres.2023.115079] [Medline: 36716551]
75. Yang C, Zhang X, Chen Y, et al. Emotion-dependent language featuring depression. J Behav Ther Exp Psychiatry 2023 Dec;81:101883. [doi: 10.1016/j.jbtep.2023.101883] [Medline: 37290350]
76. DeSouza DD, Robin J, Gumus M, Yeung A. Natural language processing as an emerging tool to detect late-life depression. Front Psychiatry 2021 Sep;12:719125. [doi: 10.3389/fpsyt.2021.719125] [Medline: 34552519]
77. Brockmeyer T, Zimmermann J, Kulessa D, et al. Me, myself, and I: self-referent word use as an indicator of self-focused attention in relation to depression and anxiety. Front Psychol 2015 Oct;6:1564. [doi: 10.3389/fpsyg.2015.01564] [Medline: 26500601]
78. Himmelstein P, Barb S, Finlayson MA, Young KD. Linguistic analysis of the autobiographical memories of individuals with major depressive disorder. PLoS One 2018 Nov;13(11):e0207814. [doi: 10.1371/journal.pone.0207814] [Medline: 30475918]
79. Vahia IV, Jeste DV, Reynolds CF. Older adults and the mental health effects of COVID-19. JAMA 2020 Dec 8;324(22):2253. [doi: 10.1001/jama.2020.21753] [Medline: 33216114]
80. Nolen-Hoeksema S. The role of rumination in depressive disorders and mixed anxiety/depressive symptoms. J Abnorm Psychol 2000;109(3):504-511. [doi: 10.1037//0021-843X.109.3.504] [Medline: 11016119]
81. Seabrook EM, Kern ML, Fulcher BD, Rickard NS. Predicting depression from language-based emotion dynamics: longitudinal analysis of Facebook and Twitter status updates. J Med Internet Res 2018 May 8;20(5):e168. [doi: 10.2196/jmir.9267] [Medline: 29739736]
82. Zhang T, Schoene AM, Ji S, Ananiadou S. Natural language processing applied to mental illness detection: a narrative review. NPJ Digit Med 2022 Apr 8;5(1):46. [doi: 10.1038/s41746-022-00589-7] [Medline: 35396451]
83. Farhat F. ChatGPT as a complementary mental health resource: a boon or a bane. Ann Biomed Eng 2024 May;52(5):1111-1114. [doi: 10.1007/s10439-023-03326-7] [Medline: 37477707]
84. NEDA suspends AI chatbot for giving harmful eating disorder advice. Psychiatrist.com. URL: https://www.psychiatrist.com/news/neda-suspends-ai-chatbot-for-giving-harmful-eating-disorder-advice/ [accessed 2023-12-23]
85. Walker L. Belgian man dies by suicide following exchanges with chatbot. The Brussels Times. URL: https://www.brusselstimes.com/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt [accessed 2023-12-23]
86. Kumar A, Agarwal C, Srinivas S, Li AJ, Feizi S, Lakkaraju H. Certifying LLM safety against adversarial prompting. arXiv. Preprint posted online on Sep 6, 2023. [doi: 10.48550/arXiv.2309.02705]
87. Ploug T, Holm S. The right to refuse diagnostics and treatment planning by artificial intelligence. Med Health Care and Philos 2020 Mar;23(1):107-114. [doi: 10.1007/s11019-019-09912-8]
88. Loi M, Ferrario A, Viganò E. How much do you trust me? A logico-mathematical analysis of the concept of the intensity of trust. Synthese 2023 May 23;201. [doi: 10.1007/s11229-023-04169-4]
89. Wang Q, Madaio M, Kane S, Kapania S, Terry M, Wilcox L. Designing responsible AI: adaptations of UX practice to meet responsible AI challenges. Presented at: CHI '23: 2023 CHI Conference on Human Factors in Computing Systems; Apr 23 to 29, 2023; Hamburg, Germany URL: https://dl.acm.org/doi/proceedings/10.1145/3544548 [accessed 2023-11-18] [doi: 10.1145/3544548.3581278]
90. Cheng SW, Chang CW, Chang WJ. The now and future of ChatGPT and GPT in psychiatry. Psychiatry Clin Neurosci 2023 Nov;77(11):592-596. [doi: 10.1111/pcn.13588] [Medline: 37612880]
91. Janjeva A, Harris A, Mercer S, Kasprzyk A, Gausen A. The rapid rise of generative AI. Centre for Emerging Technology and Security. 2023. URL: https://cetas.turing.ac.uk/publications/rapid-rise-generative-ai [accessed 2024-06-19]
92. Sedlakova J, Westermair AL, Biller-Andorno N, Meier CA, Trachsel M. Comparison of analog and digital patient decision aids for the treatment of depression: a scoping review. Front Digit Health 2023;5:1208889. [doi: 10.3389/fdgth.2023.1208889] [Medline: 37744684]
93. Elwyn G, O'Connor A, Stacey D, et al. Developing a quality criteria framework for patient decision aids: online international Delphi consensus process. BMJ 2006 Aug 26;333(7565):417. [doi: 10.1136/bmj.38926.629329.AE] [Medline: 16908462]

Abbreviations

AI: artificial intelligence
CAI: conversational artificial intelligence
LLM: large language model

Edited by A Tal; submitted 19.01.24; peer-reviewed by A Hassan, H Burkhardt, M Malgaroli, T Hull; revised version received 27.04.24; accepted 27.04.24; published 02.07.24.

Please cite as:
Ferrario A, Sedlakova J, Trachsel M
The Role of Humanization and Robustness of Large Language Models in Conversational Artificial Intelligence for Individuals With Depression: A Critical Analysis
JMIR Ment Health 2024;11:e56569
URL: https://mental.jmir.org/2024/1/e56569
doi:10.2196/56569

The Artificial Third: A Broad View of the Effects of Introducing Generative Artificial Intelligence on Psychotherapy

Yuval Haber[1], MA; Inbar Levkovich[2], MA, PhD; Dorit Hadar-Shoval[3], MA, PhD; Zohar Elyoseph[4,5], MA, PhD

1
2
3
4
5

Corresponding Author:
Zohar Elyoseph, MA, PhD

Abstract

This paper explores a significant shift in the field of mental health in general and psychotherapy in particular following generative artificial intelligence's new capabilities in processing and generating humanlike language. Following Freud, this lingo-technological development is conceptualized as the "fourth narcissistic blow" that science inflicts on humanity. We argue that this narcissistic blow has a potentially dramatic influence on perceptions of human society, interrelationships, and the self. We should, accordingly, expect dramatic changes in perceptions of the therapeutic act following the emergence of what we term the artificial third in the field of psychotherapy. The introduction of an artificial third marks a critical juncture, prompting us to ask the following important core questions that address two basic elements of critical thinking, namely, transparency and autonomy: (1) What is this new artificial presence in therapy relationships? (2) How does it reshape our perception of ourselves and our interpersonal dynamics? and (3) What remains of the irreplaceable human elements at the core of therapy? Given the ethical implications that arise from these questions, this paper proposes that the artificial third can be a valuable asset when applied with insight and ethical consideration, enhancing but not replacing the human touch in therapy.

(JMIR Ment Health 2024;11:e54781) doi:10.2196/54781

KEYWORDS

psychoanalysis; generative artificial intelligence; psychotherapy; large language models; narcissism; narcissist; narcissistic; perception; perceptions; critical thinking; transparency; autonomy; mental health; interpersonal; LLM; LLMs; language model; language models; artificial intelligence; generative; AI; ethic; ethics; ethical

Introduction

Overview

The introduction of generative artificial intelligence (GAI) has profound implications for a range of human disciplines from education and medicine to economics and law [1-3]. While forms of artificial intelligence (AI) have been around since the 1950s, impacting areas like algorithmic preferences and search functions, it was the launch of large language models (LLMs) in chatbots like ChatGPT in November 2022 that marked a significant milestone [4]. This development catapulted the role of GAI in public discourse to unprecedented prominence, surpassing previous milestones in the field [5].

LLMs are GAI algorithms that harness deep learning and vast data sets to process, summarize, generate, and predict linguistic content [6]. In this paper, we argue that the emergence of GAI is more than just a technical evolution; it is a paradigm shift with deep social and psychological implications [7,8]. It challenges the historic human monopoly over language, with the possibility of shaking the foundation of humanity's cultural and intellectual exclusivity. In other words, for the first time in history, a nonhuman entity exhibits language processing abilities that in many areas (but still not all) are equal to and sometimes even surpass those of humans. This paper explores the ripple effects of this shift by spotlighting its potential impact on psychotherapy, which is a central domain in mental health care [9-11].

We have chosen to analyze this paradigmatic shift brought about by the development of GAI through the lens of Freud's concept of "narcissistic blows" to human self-understanding. This psychoanalytic concept aptly describes how major scientific advances, such as the development of GAI, can fundamentally challenge long-held beliefs about human identity and uniqueness. While other theoretical frameworks such as Kuhn's [12] paradigm shifts could also provide valuable insights into the impact of GAI, we believe that Freud's theory offers a particularly compelling lens through which to examine the psychological, social, and existential dimensions of this possible technological revolution. The return to Freud's conceptualization is also valuable because he is considered one of the most influential thinkers in shaping the modern self and is widely regarded as the "father" figure of psychotherapy. Therefore, revisiting his work is particularly relevant when dealing with a transformative change like the one brought about by the introduction of GAI into the psychotherapeutic space. Following this psychoanalytic analysis, we will propose in this paper to conceptualize the entry of GAI into the psychotherapeutic space as an artificial third.

GAI: The Fourth Narcissistic Blow to Humanity

In 1917, in his article "A Difficulty in the Path of Psycho-Analysis," Freud [13] assessed how scientific discoveries reshaped our cultural understanding and self-perception. He identified three narcissistic blows inflicted on humanity by science that pushed us to confront and abandon long-held naive beliefs of our narcissistic centrality and control over the world. These paradigm shifts, although jarring, catalyzed tremendous societal advancement [13]. First, the "cosmological" blow came from Copernicus who taught us that the earth orbits the sun and not the other way around [14,15]. Second, Darwin's "biological" revelation presented humans as just another evolutionary link and not a divine culmination [14,15]. Third, Freud himself introduced the "psychological" blow with his psychoanalytic theory suggesting that beneath our perceived rationality lays unconscious drives and conflicts that are beyond the control of our "ego" [13,14]; by asserting that one is not the master of one's own house, he implied our limited control over our inner selves [13].

Building on Freud's framework, modern computer science appears to be delivering a fourth potent blow to human narcissism, which we suggest conceptualizing as the "linguistic narcissistic blow." This blow is historically profound: after understanding that the earth is not the center of the universe (the "cosmological" blow), recognizing our nonexceptionalism in nature (the "biological" blow), and confronting the turbulent undercurrents of our psyche (the "psychological" blow), we are now also faced with the prospect of sharing our linguistic domain. What was once an exclusive human domain might now be shared with ever more sophisticated artificial entities [15]. It is difficult to exaggerate the drama of the possible loss of human monopoly on language since the natural and primary characteristic of humanity was the ability to control and play with language, namely, to produce signs and symbols that indicate things as well as to ever alter the meanings of these symbols. This is how humans acquired the ability to create a subjective mental image and meaning of the external world. These signs can be not only literal but also symbolic in the form of paintings, symbols, rituals, and more. In other words, language and the ability to play with it are, as we learned from Winnicott [16], the foundation of all human culture.

It should be clarified that we do not claim GAI systems possess human language understanding and control. While GAI systems have demonstrated impressive language processing and generation capabilities, they still do not possess genuine understanding or comprehension of language in the same way that humans do [17]. These systems operate based on complex statistical patterns and associations learned from vast amounts of training data, but they lack the rich contextual knowledge, reasoning abilities, and embodied experience that underpin human language use [18]. Nonetheless, the ability of GAI systems to generate highly coherent and contextually appropriate linguistic outputs has significant and dramatic implications for a range of domains, including psychotherapy.

As we stand on this precipice where GAI entities have an increasing ability to process and produce language, we face a new era brimming with both potential and intricate challenges, particularly in the sphere of mental health [8-10,17].

Integrating GAI Into Psychotherapy

For over a century, psychotherapy, which is a linchpin in mental health, has hinged primarily on the dialogue between therapist and patient. Its origins can be traced back to Freud and Breuer's [19] seminal work *Studies in Hysteria*, which conceptualized psychotherapy as the "talking cure." Although myriad forms of psychotherapy have emerged, the central emphasis on the main method of the patient-therapist dialogue remained consistent [20]. In fact, despite extensive technological and theoretical development in the 20th century, the incorporation of technology into the therapeutic field has been surprisingly minimal [20]. Even as advancements like biofeedback, neurofeedback, and virtual reality have arisen, their widespread adoption in therapeutic practice is still limited. However, it seems that GAI, having demonstrated increasingly sophisticated language processing and generation capabilities, stands poised to radically revolutionize the field of psychotherapy.

Wittgenstein's [21] linguistic theory provides a comprehensive philosophical lens to examine the linguistic capabilities and limitations of GAI in the context of psychotherapy. Wittgenstein's [21] early work assertion that "we make to ourselves pictures of facts" refers to the way language mediates one's world into a subjective picture consisting of words. This lingo-philosophical perception encapsulates the potential of GAI to also build a picture of the world from words or, using therapeutic concepts, generate outputs that resemble interpretations, insights, narratives, reflections, validations, and more. Until recently, all of these talk-based psychotherapy capabilities were restricted to mental health professionals; now, thanks to its proficiency in language, GAI is not just an adjunct but may actually redefine the therapeutic landscape. Preliminary research underscores GAI's prowess in tasks like treatment summarization, risk assessment, and real-time diagnosis, all of which rely on highly strong language processing capabilities [8-10,22].

While Wittgenstein's [23] early work presents language as a picture of reality, in his later philosophy, he developed the concept of "language-games" [24]. This concept highlights the rule-governed, socially embedded nature of linguistic meaning. Following Wittgenstein's later work, other linguistic theories such as "speech act" theory [25] also emphasized the social and practical dimensions of language. From this perspective, GAI's ability to produce meaningful interpretations and to hold the full linguistic understanding necessary for functioning in the therapeutic realm is constrained. This limitation arises from its lack of grounding in the social and normative aspects of language use.

Nevertheless, we argue that in an age when GAI's linguistic competencies rival or even eclipse those of humans in some (although, as we emphasized, certainly not all) linguistic areas, it is hard to imagine that psychotherapy will remain an untouched bastion, preserving its "No Entry for Technology" doctrine. GAI's potential integration into mental health and psychotherapy, in particular, heralds both vast opportunities and challenges, with the prospect of reshaping the therapeutic practice [26].

Hence, it is our conviction that mental health professionals bear an ethical responsibility to proactively engage and influence the integration of GAI within the clinical domain. Specifically, the development and application of GAI tools, methodologies, and conceptualizations for psychotherapy in particular, and mental health in general, must involve close collaboration between mental health professionals, AI researchers, and ethicists to ensure alignment with the core values and goals of the profession [7].

To illustrate the potential integration of GAI into the psychotherapeutic process, let us consider a scenario where a GAI system actively participates in a live therapy session. The GAI listens to the dialogue, analyzes the language and sentiment in real time, and provides insights to both the patient and the therapist. For instance, it might highlight patterns in the patient's narrative that suggest underlying mental risks, cognitive distortions, or emotional conflicts. It could also offer the therapist suggestions for therapeutic interventions based on the patient's unique profile. After the session, the GAI could provide a summary, identifying key themes and tracking progress over time. Of course, such an application would need to be developed and deployed with utmost care for patient privacy, data security, and both clinical and ethical considerations [27].

GAI can also serve as an innovative playful space that allows for exploring and processing intrapsychic and interpersonal dynamics in new creative ways. To further illustrate this, let us consider another example where the GAI presence within the therapeutic process supports and enhances the technique of externalizing internal psychological pains or conflicts. By using problem externalization from narrative therapy, GAI systems can generate tangible representations of the patient's inner struggles, either by transforming the internal voice into a visual representation using an image generator or by having the GAI embody the inner voice, allowing the patient to converse with it in the presence of the therapist and explore it together. This innovative approach may potentially assist patients in gaining new perspectives on their internal conflicts, with the therapist guiding the therapeutic process.

These GAI's abilities to process and analyze the therapeutic dialogue in real-time, offer insights and suggestions, provide postsession summaries, and engage in live dialogue or role-play with a patient and therapist introduce a novel presence that actively reshapes the dynamics and outcomes of the therapeutic process.

Exploring the Role of GAI in the Therapeutic Triad

The introduction of GAI into psychotherapy raises central questions about its clinical, ethical, and interpersonal implications [18,26]. This paper suggests that to truly grasp the impact of GAI in psychotherapy, one must first recognize the potential dynamics of introducing a third element into the traditional therapist-patient dyad. This dyadic structure, a cornerstone of psychotherapy, has remained largely unchanged over the past century [28]. Yet, it is an oversimplification to view psychotherapy as solely a 2-person interaction, since it has always operated within a more complex triadic framework [29]. In fact, the "third" or "thirdness" as "other" to the dyad holds a paradoxical attitude: the "third" not only threatens the connection of the dyad but also enables it; in other words, it enables the psychotherapeutic act itself.

Due to a lack of scope, this paper can only touch key landmarks in the triadic concept's genealogy within psychoanalysis [29]. Freud was the first to place the model of triads at the center of psychological and clinical thinking. For him, the ultimate other was represented by the father within the oedipal developmental drama, that is, in the well-known drama of the triangle relationship formed between the child, their mother, and their father in the child's early years. Subsequent psychoanalytic theories developed this "third" or "thirdness" concept further [13]. For example, Winnicott [16] visualized it as a "transitional space"—a nexus between imagination and reality and the space where creativity and culture can develop. Ogden [30] introduced the important concept of "analytic third," addressing the "third" as an abstract entity born from the therapist-patient interrelationship. Building on Ogden's concept, Bar Nes [31] suggested that the emergence of the "analytic third" is facilitated by a combination of verbal and nonverbal communication, which creates a psychic overlap between therapist and patient. Finally, Lacan [32], expanding on Freud's ideas, posited that the "third" symbolizes the overarching structure of language that assimilates into our unconscious, shaping the individual's identity and interactions with the world.

In short, the concept of the "third" in the sphere of psychotherapy is not new at all. Nonetheless, its presence is quite complex; it, on the one hand, is ever present in the therapeutic dyad and, on the other hand, always remains the distinct "other."

Redefining the Therapeutic Triangle in the Digital Era

As GAI is being embedded in the mental health arena, we are witnessing the emergence of a distinctive triangular dynamic: therapist, patient, and the GAI "third" entity [26]. This GAI "third," as a continuation of the "third" in the psychoanalytic perspective, remains as the "other" joining the dyad, but the new quality it brings also fundamentally changes the balance of the therapeutic relationship. For the first time, this "third" element becomes tangible, transitioning from a mere conceptual presence to an interactive entity that both therapist and patient can have a dialogue with and not just about. This marks a new horizon in both the triad therapeutic dynamics and the patient-therapist dynamics [33].

Based on psychoanalysis' evolution of the "third" mentioned above, we propose conceptualizing the presence of GAI in psychotherapy (and mental health in general) as an artificial third. The artificial third refers to the concrete and symbolic expression of GAI in the cultural-political-digital space and its shaping presence in the therapeutic encounter, society, and the self. While the artificial third can appear in therapy as a "transitional space," thus creating a playful space in accordance with Winnicott's [16] perception, its influence does not end there: it also shapes our perception of ourselves and the world. To clarify this point, if until the digital age the way to explore our consciousness was through other human consciousnesses, now it seems, the state of affairs is different. For the first time, we are beginning to know ourselves not only through reflections from other humans but also through direct and indirect reflections from AI entities and their control of the digital sphere.

Although human therapists are far from obsolete, the integration of the artificial third might substantially transform therapeutic methodologies, redefining the essence of therapy and the perceptions and roles of both therapists and patients [18]. However, the impact of the artificial third extends beyond merely shaping the therapeutic landscape; it is also intricately crafted by the human experiences and notions that constitute its core training data. This two-way interaction has significant implications for the development and impact of the artificial third in psychotherapy, suggesting that GAI technology is not only a shaping force but is also influenced by the cultural values, biases, and limitations of its human creators and users [34]. With the incorporation of GAI, the field of mental health stands on the cusp of revolutionary advancements, with the promise of enhanced professional opportunities but also the potential for great dangers and pitfalls [35].

Advantages of Integrating GAI in Psychological Care

The integration of GAI into psychological care augurs transformative new methodologies. Central to these advancements is the potential for vastly improved accessibility to mental health services. Worldwide, numerous barriers, ranging from socioeconomic constraints and geographical distances to linguistic challenges and intricate cultural contexts, limit access to quality care [36,37]. However, the artificial third may stand to bridge these divides, widening the scope of individuals who are able to receive psychological support.

Beyond accessibility, this novel approach may enable a new level of personalization in mental health care. It envisions a therapeutic experience tailored meticulously to each individual and accounting for clinical, linguistic, cultural, and personal characteristics, and promises a significant shift in therapeutic dynamics. Historically, therapeutic expertise has been monopolized by professionals, but the artificial third might, although not necessarily, foster a more collaborative patient-therapist relationship [7,38]. Such a change may place patients at the forefront of their healing journey, empowering them to engage actively in their therapy.

Moreover, the artificial third opens up new pluralistic possibilities, allowing for the integration of established evidence-based approaches like cognitive behavioral treatment and psychodynamic therapy with cutting-edge and emerging psychological frameworks [39,40]. It may also, as mentioned above, create a new playful space that strengthens the relationship between patient and therapist. While still a burgeoning concept, the artificial third holds undeniable promise for redefining psychological care and harmoniously blending accessibility, personalization, and innovation.

Challenges of Integrating GAI in Psychological Care

The integration of the artificial third into psychological services presents a multifaceted set of challenges. One major concern is the consolidation of vast amounts of patients' personal data under a handful of dominant corporations. Such centralization could divert the focus from patient-centered care to commercial interests, thus risking service quality [41]. This commercial shift also raises alarms about unauthorized data use or deep analysis of user behaviors without explicit consent, jeopardizing both user confidentiality and overarching ethical standards [36]. Adding to the uncertainty is the often opaque nature of algorithms, which could obscure decision-making processes.

A subtler yet significant concern is the exaggerated reliance on the epistemic authority attributed to the artificial third. As the artificial third becomes more integral, there is a danger that traditional expert voices, like experienced psychologists, could be overshadowed, which could undervalue the importance of human expertise, experience, thought, and insights in therapy. Furthermore, the potential overreliance on GAI tools in therapy could risk diminishing the significance of authentic human connection and empathy in the therapeutic process.

This concern becomes more acute when these tools are developed by mental health professionals themselves. Although these tools can provide distinct advantages, they are not exempt from the ethical dilemmas and possible conflicts of interest that emerge when therapists participate in the development of a GAI product. The substantial risk is that the artificial third may, detrimentally, shift to become the central focus of therapy in such instances, rather than acting as an instrumental aid designed to support the patients' therapeutic journey. Additionally, with the artificial third still in its infancy, there is a tangible risk of off the mark guidance, which could lead individuals away from tested psychological practices [42]. Finally, while LLMs are perceived as objective and neutral, they were actually put through a training process that aligned their reactions to a specific "value-like" system that is not transparent to the public [34].

The focus of these concerns lies in the lack of a unified regulatory framework for GAI. Although traditional psychological practices are governed by established regulations, GAI systems currently operate without such oversight. This regulatory gap may allow companies to craft their guidelines, potentially leading to variances in standards and a deviation from the trusted norms of mental health care [38,43]. Moreover, preliminary findings suggest that GAI data sets may sometimes produce social, economic, and cultural biases [7] that have the potential to be inadvertently amplified, thus undermining the goal of providing unbiased psychological support. In conclusion, despite its innovative promise, the introduction of GAI in psychotherapy carries inherent risks that may undermine the therapeutic relationship, the therapist's professionalism, and the ability to promote the patient's well-being.

Engaging With the Artificial Third: Three Fundamental Questions

Overview

The trajectory of mental health and, specifically, the psychotherapeutic realm in the era of the artificial third remains uncertain. However, it is unequivocally clear that the landscape will undergo profound transformations [7-10,22]. In light of this change and uncertainty, it is imperative to arm both therapists and patients with three key questions regarding (1) the nature of the artificial third, (2) our relationship with it, and (3) the role of humanity in the artificial third era. These questions aim to foster a deeper understanding of the artificial third, encourage reflection on our interaction with it, and prompt a consideration of the unique value and position of human beings in this new landscape. While these questions relate also to the broader interface between humans and AI, they hold special relevance for psychotherapy and the larger mental health domain [7,17,22].

These three questions are meticulously designed to advance the principles of transparency and autonomy, which are essential for fostering critical thinking, especially in the context of interacting with AI systems [44,45]. Critical thinking is one of the most important capacities for promoting human freedom and agency [46,47]. By emphasizing transparency—the ability to understand how GAI systems work and what influences their outputs—and autonomy—the ability to maintain independent thought, creation, and decision-making in the face of GAI influence—we aim to highlight the importance of preserving space for critical reflection and self-determination in the era of AI.

The Nature of the Artificial Third

In opposition to the widespread but simplistic view that regards GAI systems as impartial and objective, we contend they are based on certain values and cultures that are shaped by the critical factor of the alignment process [34]. Understanding the influence of the alignment process is essential for the responsible integration of GAI into psychotherapy. Therapists have a crucial role in the era of the artificial third—to explore and comprehend the alignment mechanisms of the GAI systems they use. This entails recognizing the inherent values, motivations, and limitations of these systems. Therefore, understanding the nature of the artificial third is not merely about operational knowledge but also carries ethical weight, emphasizing the necessity for a deep awareness of this emerging presence within the therapeutic setting. Therapists have a responsibility to share their insights about GAI systems with their patients, incorporating this transparency into the informed consent process, thus ensuring the maintenance of ethical standards in this new therapeutic landscape.

Consequently, we propose the following question as a guiding principle for therapists and patients when considering the use of GAI in psychotherapy: "To what extent do we understand the alignment process and limitations of the GAI system we are working with?" This question prompts us to consider the following subquestions: What are the underlying values, interests, and driving forces? How are responses generated? and Who bears the responsibility for them? Addressing these questions effectively at a societal level and shaping appropriate policies necessitates the implementation of a structured regulatory framework to ensure the responsible and ethical application of AI in the field of mental health.

Our Relationship With the Artificial Third

As we elaborated on in the paper, the introduction of the artificial third in psychotherapy creates a new therapeutic triangle of therapist, patient, and GAI. Optimally, this entry can promote a playful space in therapy and patients' well-being; however, negatively, it may lead to a detriment in focusing on the patient's needs and the therapist's self-thinking. Therefore, the most important concrete question when examining the effect of the artificial third entry into psychotherapy is to consider "To what extent does the artificial third become central in the therapeutic space, and does this centrality come at the expense of the patient?" This question examines the dynamics of our interaction with this artificial entity and its position in the new therapeutic triangle.

From a more philosophical point of view, we can also ask "how our discourse with the artificial third is shaping our perceptions of the self, the therapeutic act, and the roles of both the patient and therapist?" This reflection, which necessitates further philosophical, clinical, and ethical research, extends far beyond the confines of the therapy room, delving into broader questions of identity and the essence of human relationships and communication in an increasingly digital landscape.

The Role of Humanity

Rooted in the principle of autonomy [44], this question critically assesses humanity's latitude with AI and asks "What distinguishes the human subject from the artificial object?" In the realm of mental health, it differentiates between the roles in which AI can be beneficial and those in which it might be detrimental, suggesting that the latter are better left to humans.

Encountering the artificial third as a nonhuman entity presents a unique opportunity: it allows us to reexplore and, perhaps, redefine our humanity. With the presence of this artificial entity in therapeutic sessions, the question of human uniqueness becomes significant. The concrete question that arises from this critical thinking is "What is the added value of the human therapist within the therapeutic space?" As elaborated on in this paper, it seems that while GAI can provide textual analyses at the highest level, the ability for true human experience, thinking, and therapeutic interpretation is still irreplaceable. In this sense, there is a possibility that the encounter with the artificial third, as an entity fundamentally different from humanity, will not harm the human place in the therapeutic process but rather emphasize its uniqueness and special contribution.

Conclusion

Since the release of GAI systems, numerous studies have been conducted regarding their applications in the field of mental health [2,7-10,22,34,48-53]. However, the current research seeks to examine the entry of this technology from a broader perspective, particularly focusing on its potential impact on psychotherapy.

The entry of GAI into our lives seems to mark a new age characterized by the possible loss of humanity's historic monopoly over language, conceptualized in this paper as the "fourth narcissistic blow" (the "linguistic narcissistic blow") inflicted on humans by science. The fourth narcissistic blow refers to the fact that, due to recent scientific and technological progress, artificial entities are starting to display language abilities similar to humans. Yet, this paper also emphasized that although GAI has impressive analytical and processing capabilities, it still lacks some essential social and pragmatic aspects inherent to human language.

Psychotherapy may experience an upheaval following the recent increasing use of GAI, whose presence was defined here as the artificial third. While GAI's linguistic abilities may challenge traditional psychotherapeutic paradigms, it also unveils opportunities for enriched therapeutic processes, provided these technologies are used with discernment and a deep commitment to the ethical imperatives of therapy, as well as taking into consideration GAI's limitations and biases. Therefore, we argue that the integration of the artificial third into the therapeutic sphere does not inherently improve or detract from the psychotherapeutic act. Rather, its influence is contingent, relying on the context, the goal, and the way in which it is implemented.

Moreover, the artificial third is neither an unequivocal solution nor a replacement for the quintessentially human aspects of therapy. Instead, it represents a nuanced technological tool that, if integrated carefully, could hold the promise of enhancing therapeutic practice. As we navigate this delicate balance, the imperative remains to safeguard the irreplaceable human connection that lies at the heart of the therapeutic act, ensuring that technology augments rather than diminishes it.

Indeed, in the artificial third era, we need to accept its growing presence while guarding against total dependence. Upholding patient and therapist autonomy together with GAI transparency will be vital for enabling critical thinking, which is important, especially in the fragile realm of mental health. This will allow us to both judiciously leverage GAI's potential and protect the humanistic essence of therapeutic practice. This exploration of the artificial third in psychotherapy underscores a pivotal juncture in our understanding and practice of therapy.

Conflicts of Interest

None declared.

References

1. Xu S, Deo RC, Soar J, et al. Automated detection of airflow obstructive diseases: a systematic review of the last decade (2013-2022). Comput Methods Programs Biomed 2023 Nov;241:107746. [doi: 10.1016/j.cmpb.2023.107746] [Medline: 37660550]
2. Elyoseph Z, Levkovich I. Comparing the perspectives of generative AI, mental health experts, and the general public on schizophrenia recovery: case vignette study. JMIR Ment Health 2024 Mar 18;11:e53043. [doi: 10.2196/53043] [Medline: 38533615]
3. van Heerden AC, Pozuelo JR, Kohrt BA. Global mental health services and the impact of artificial intelligence–powered large language models. JAMA Psychiatry 2023 Jul 1;80(7):662. [doi: 10.1001/jamapsychiatry.2023.1253]
4. Ozmen Garibay O, Winslow B, Andolina S, et al. Six human-centered artificial intelligence grand challenges. Int J Hum Comput 2023 Feb 7;39(3):391-437. [doi: 10.1080/10447318.2022.2153320]
5. Zhou W, Zhang C, Wu L, Shashidhar M. ChatGPT and marketing: analyzing public discourse in early Twitter posts. J Market Anal 2023 Dec;11(4):693-706. [doi: 10.1057/s41270-023-00250-6]
6. Romano MF, Shih LC, Paschalidis IC, Au R, Kolachalama VB. Large language models in neurology research and future practice. Neurology 2023 Dec 4;101(23):1058-1067. [doi: 10.1212/WNL.0000000000207967] [Medline: 37816646]
7. Tal A, Elyoseph Z, Haber Y, et al. The artificial third: utilizing ChatGPT in mental health. Am J Bioethics 2023 Oct 3;23(10):74-77. [doi: 10.1080/15265161.2023.2250297]
8. Elyoseph Z, Hadar-Shoval D, Asraf K, Lvovsky M. ChatGPT outperforms humans in emotional awareness evaluations. Front Psychol 2023 May 26;14:1199058. [doi: 10.3389/fpsyg.2023.1199058] [Medline: 37303897]
9. Elyoseph Z, Levkovich I. Beyond human expertise: the promise and limitations of ChatGPT in suicide risk assessment. Front Psychiatry 2023 Aug 1;14. [doi: 10.3389/fpsyt.2023.1213141]
10. Levkovich I, Elyoseph Z. Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: vignette study. JMIR Ment Health 2023 Sep 20;10:e51232. [doi: 10.2196/51232] [Medline: 37728984]
11. Carlbring P, Hadjistavropoulos H, Kleiboer A, Andersson G. A new era in internet interventions: the advent of Chat-GPT and AI-assisted therapist guidance. Internet Interv 2023 Apr 11;32:100621. [doi: 10.1016/j.invent.2023.100621] [Medline: 37273936]
12. Kuhn TS. The Structure of Scientific Revolutions: University of Chicago Press; 1962.
13. Freud S. A difficulty in the path of psycho-analysis. In: The Standard Edition of the Complete Psychological Works of Sigmund Freud Vol 17: An Infantile Neurosis and Other Works (1917-1919): Hogarth Press; 1917:135-144.
14. Weinert F. Copernicus, Darwin, and Freud: Revolutions in the History and Philosophy of Science: John Wiley & Sons; 2008.

15. Nefdt RM. Biolinguistics and biological systems: a complex systems analysis of language. Biol Philos 2023 Apr;38(2):12. [doi: 10.1007/s10539-023-09903-3]
16. Winnicott DW. Mirror-role of mother and family in child development. In: du Gay P, Evans J, Redman P, editors. Identity: A Reader: Sage; 2000:144.
17. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? Presented at: 2021 ACM Conference on Fairness, Accountability, and Transparency; Mar 3 to 10, 2021.. [doi: 10.1145/3442188.3445922]
18. Sedlakova J, Trachsel M. Conversational artificial intelligence in psychotherapy: a new therapeutic tool or agent. Am J Bioeth 2023 May;23(5):4-13. [doi: 10.1080/15265161.2022.2048739] [Medline: 35362368]
19. Freud S, Breuer J. Studies in Hysteria: Penguin; 2004.
20. Cuijpers P, Reijnders M, Huibers MJH. The role of common factors in psychotherapy outcomes. Annu Rev Clin Psychol 2019 May 7;15:207-231. [doi: 10.1146/annurev-clinpsy-050718-095424] [Medline: 30550721]
21. Wittgenstein L. Tractatus Logico-Philosophicus, 471th edition: Dover Publications; 1998.
22. Levkovich I, Elyoseph Z. Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. Fam Med Community Health 2023 Sep;11(4):e002391. [doi: 10.1136/fmch-2023-002391] [Medline: 37844967]
23. Wittgenstein L. The Philosophical Investigations: Blackwell; 1953.
24. Fogelin RJ. Taking Wittgenstein at His Word: A Textual Study: Princeton University Press; 2020.
25. Austin JL. How to Do Things With Words: Harvard University Press; 1975.
26. Fiske A, Henningsen P, Buyx A. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J Med Internet Res 2019 May 9;21(5):e13216. [doi: 10.2196/13216] [Medline: 31094356]
27. Spinrad A, Taylor CB, Ruzek JI, et al. Action recommendations review in community-based therapy and depression and anxiety outcomes: a machine learning approach. BMC Psychiatry 2024 Feb 16;24(1):133. [doi: 10.1186/s12888-024-05570-0] [Medline: 38365635]
28. Nahum D, Alfonso CA, Sönmez E. Common factors in psychotherapy. In: Javed A, Fountoulakis K, editors. Advances in Psychiatry: Springer; 2019:471-481. [doi: 10.1007/978-3-319-70554-5]
29. Laplanche J, Pontalis JB. The Language of Psychoanalysis: Routledge; 2018.
30. Ogden TH. The analytic third: implications for psychoanalytic theory and technique. Psychoanal Quart 2004 Jan;73(1):167-195. [doi: 10.1002/j.2167-4086.2004.tb00156.x]
31. Bar Nes A. The psychoanalytic mystic and the interpretive word. J Am Psychoanal Assoc 2022 Oct;70(5):903-938. [doi: 10.1177/00030651221124803] [Medline: 36314518]
32. Nobus D. Key Concepts of Lacanian Psychoanalysis: Other Press; 2020.
33. Zubala A, Kennell N, Hackett S. Art therapy in the digital world: an integrative review of current practice and future directions. Front Psychol 2021 Apr 8;12:595536. [doi: 10.3389/fpsyg.2021.600070] [Medline: 33897522]
34. Hadar-Shoval D, Asraf K, Mizrachi Y, Haber Y, Elyoseph Z. Assessing the alignment of large language models with human values for mental health integration: cross-sectional study using Schwartz's theory of basic values. JMIR Ment Health 2024 Apr 9;11:e55988. [doi: 10.2196/55988] [Medline: 38593424]
35. Sun J, Dong QX, Wang SW, et al. Artificial intelligence in psychiatry research, diagnosis, and therapy. Asian J Psychiatr 2023 Sep;87:103705. [doi: 10.1016/j.ajp.2023.103705] [Medline: 37506575]
36. Hodgkinson S, Godoy L, Beers LS, Lewin A. Improving mental health access for low-income children and families in the primary care setting. Pediatrics 2017 Jan;139(1):e20151175. [doi: 10.1542/peds.2015-1175] [Medline: 27965378]
37. Byrow Y, Pajak R, Specker P, Nickerson A. Perceptions of mental health and perceived barriers to mental health help-seeking amongst refugees: a systematic review. Clin Psychol Rev 2020 Feb;75:101812. [doi: 10.1016/j.cpr.2019.101812] [Medline: 31901882]
38. Cohen IG. What should ChatGPT mean for bioethics? Am J Bioeth 2023 Oct;23(10):8-16. [doi: 10.1080/15265161.2023.2233357] [Medline: 37440696]
39. Grodniewicz JP, Hohol M. Waiting for a digital therapist: three challenges on the path to psychotherapy delivered by artificial intelligence. Front Psychiatry 2023 Jun 1;14:1190084. [doi: 10.3389/fpsyt.2023.1190084] [Medline: 37324824]
40. Pilecki B, Thoma N, McKay D. Cognitive behavioral and psychodynamic therapies: points of intersection and divergence. Psychodyn Psychiatry 2015 Sep;43(3):463-490. [doi: 10.1521/pdps.2015.43.3.463] [Medline: 26301762]
41. Zajko M. Automated government benefits and welfare surveillance. Surveill Soc 2023 Sep 24;21(3):246-258. [doi: 10.24908/ss.v21i3.16107]
42. Hatem R, Simmons B, Thornton JE. Chatbot confabulations are not hallucinations. JAMA Intern Med 2023 Oct 1;183(10):1177. [doi: 10.1001/jamainternmed.2023.4231] [Medline: 37578766]
43. Coghlan S, Gyngell C, Vears DF. Ethics of artificial intelligence in prenatal and pediatric genomic medicine. J Community Genet 2024 Feb;15(1):13-24. [doi: 10.1007/s12687-023-00678-4] [Medline: 37796364]
44. Laitinen A, Sahlgren O. AI systems and respect for human autonomy. Front Artif Intell 2021 Oct 26;4:705164. [doi: 10.3389/frai.2021.705164] [Medline: 34765969]

45. Felzmann H, Fosch-Villaronga E, Lutz C, Tamò-Larrieux A. Towards transparency by design for artificial intelligence. Sci Eng Ethics 2020 Dec;26(6):3333-3361. [doi: 10.1007/s11948-020-00276-4] [Medline: 33196975]
46. Foucault M. What is critique. In: What Is Enlightenment: University of California Press; 2019:382-398.
47. Barry L. Concluding remarks: Foucault and contemporary social criticism. In: Foucault and Postmodern Conceptions of Reason: Palgrave Macmillan; 2020. [doi: 10.1007/978-3-030-48943-4]
48. Hadar-Shoval D, Elyoseph Z, Lvovsky M. The plasticity of ChatGPT's mentalizing abilities: personalization for personality structures. Front Psychiatry 2023 Sep 1;14:1234397. [doi: 10.3389/fpsyt.2023.1234397] [Medline: 37720897]
49. Elyoseph Z, Refoua E, Asraf K, Lvovsky M, Shimoni Y, Hadar-Shoval D. Capacity of generative artificial intelligence to interpret human emotions from visual and textual data: pilot evaluation study. JMIR Ment Health 2024 Feb 6;11:e54369. [doi: 10.2196/54369] [Medline: 38319707]
50. Asman O, Tal A, Barilan YM. Conversational artificial intelligence-patient alliance turing test and the search for authenticity. Am J Bioeth 2023 May;23(5):62-64. [doi: 10.1080/15265161.2023.2191046] [Medline: 37130413]
51. Elyoseph Z, Hadar Shoval D, Levkovich I. Beyond personhood: ethical paradigms in the generative artificial intelligence era. Am J Bioeth 2024 Jan;24(1):57-59. [doi: 10.1080/15265161.2023.2278546] [Medline: 38236857]
52. Elyoseph Z, Gur T, Haber Y, et al. An ethical perspective on the democratization of mental health with generative artificial intelligence. JMIR Preprints. Preprint posted online on Mar 2, 2024. [doi: 10.2196/preprints.58011]
53. Elyoseph Z, Levkovich I, Shinan-Altman S. Assessing prognosis in depression: comparing perspectives of AI models, mental health professionals and the general public. Fam Med Community Health 2024 Jan 9;12(Suppl 1):e002583. [doi: 10.1136/fmch-2023-002583] [Medline: 38199604]

Abbreviations

AI: artificial intelligence
GAI: generative artificial intelligence
LLM: large language model

Edited by A Tal; submitted 22.11.23; peer-reviewed by J Sedlakova, R Hendricks-Sturrup, R D'Angelo; revised version received 24.03.24; accepted 18.04.24; published 23.05.24.

Please cite as:
Haber Y, Levkovich I, Hadar-Shoval D, Elyoseph Z
The Artificial Third: A Broad View of the Effects of Introducing Generative Artificial Intelligence on Psychotherapy
JMIR Ment Health 2024;11:e54781
URL: https://mental.jmir.org/2024/1/e54781
doi:10.2196/54781

Comparing the Perspectives of Generative AI, Mental Health Experts, and the General Public on Schizophrenia Recovery: Case Vignette Study

Zohar Elyoseph[1,2], BA, MA, PhD; Inbar Levkovich[3], BA, MA, PhD, Prof Dr

1

2

3

Corresponding Author:
Zohar Elyoseph, BA, MA, PhD

Abstract

Background: The current paradigm in mental health care focuses on clinical recovery and symptom remission. This model's efficacy is influenced by therapist trust in patient recovery potential and the depth of the therapeutic relationship. Schizophrenia is a chronic illness with severe symptoms where the possibility of recovery is a matter of debate. As artificial intelligence (AI) becomes integrated into the health care field, it is important to examine its ability to assess recovery potential in major psychiatric disorders such as schizophrenia.

Objective: This study aimed to evaluate the ability of large language models (LLMs) in comparison to mental health professionals to assess the prognosis of schizophrenia with and without professional treatment and the long-term positive and negative outcomes.

Methods: Vignettes were inputted into LLMs interfaces and assessed 10 times by 4 AI platforms: ChatGPT-3.5, ChatGPT-4, Google Bard, and Claude. A total of 80 evaluations were collected and benchmarked against existing norms to analyze what mental health professionals (general practitioners, psychiatrists, clinical psychologists, and mental health nurses) and the general public think about schizophrenia prognosis with and without professional treatment and the positive and negative long-term outcomes of schizophrenia interventions.

Results: For the prognosis of schizophrenia with professional treatment, ChatGPT-3.5 was notably pessimistic, whereas ChatGPT-4, Claude, and Bard aligned with professional views but differed from the general public. All LLMs believed untreated schizophrenia would remain static or worsen without professional treatment. For long-term outcomes, ChatGPT-4 and Claude predicted more negative outcomes than Bard and ChatGPT-3.5. For positive outcomes, ChatGPT-3.5 and Claude were more pessimistic than Bard and ChatGPT-4.

Conclusions: The finding that 3 out of the 4 LLMs aligned closely with the predictions of mental health professionals when considering the "with treatment" condition is a demonstration of the potential of this technology in providing professional clinical prognosis. The pessimistic assessment of ChatGPT-3.5 is a disturbing finding since it may reduce the motivation of patients to start or persist with treatment for schizophrenia. Overall, although LLMs hold promise in augmenting health care, their application necessitates rigorous validation and a harmonious blend with human expertise.

(JMIR Ment Health 2024;11:e53043) doi:10.2196/53043

KEYWORDS

schizophrenia; mental; prognostic; prognostics; prognosis; ChatGPT; artificial intelligence; recovery; vignette; vignettes; outcome; outcomes; large language models; language model; language models; LLM; LLMs; NLP; natural language processing; GPT; Generative Pre-trained Transformers

Introduction

Background

Schizophrenia is a major contributor to mental health–related disability worldwide and exerts a profound effect on patients and society [1]. It has a major impact on life expectancy and quality of life, and its repercussions extend to family and caregivers [2]. The disorder presents a complex array of symptoms, both "positive" (eg, delusions and hallucinations) and "negative" (eg, emotional flatness and social withdrawal) [3]. Negative symptoms are especially resistant to current treatments [4]. Despite the complexity and impact of this disorder, a subset of individuals with schizophrenia may have a favorable prognosis; exhibit symptom reduction; and achieve positive outcomes in education, employment, and relationships [3].

A systematic review based on 37 studies that examined outcomes in first-episode psychosis [5] determined that 42% of patients experienced favorable outcomes. Similarly, an analysis of 114 follow-up studies to assess recovery rates in schizophrenia [6] yielded recovery rates ranging from 11% to 33% for complete recovery and from 22% to 53% for social recovery. Another meta-analysis [7] encompassing 50 pertinent studies revealed that approximately one-seventh of individuals diagnosed with schizophrenia met our predefined criteria for recovery.

The beliefs held by therapists regarding their patients' capacity for recuperation represent a complex and multifaceted conundrum [8]. From a pragmatic standpoint, a medical practitioner's proficiency in ascertaining a patient's prospective therapeutic trajectory (known as "prognosis" in the medical field) is a major clinical aptitude [9]. From an ethical perspective, clinicians are duty bound to elucidate potential perils and advantages to patients, thus facilitating a process of informed consent and collaborative decision-making [10]. Providing a nuanced yet candid prognosis enhances patient motivation and optimism when the likelihood of complete remission is high, while concurrently calibrating expectations in less promising scenarios [11-13]. Nevertheless, inherent values and presuppositions inevitably shape prognostic assessments [14,15]. The etiology and treatability of psychiatric disorders are framed by 2 opposing philosophical paradigms. Deterministic models, which view mental disorders as fixed biological anomalies, often adopt a pessimistic perspective on full recovery. In contrast, the recovery model approach is rooted in the belief that complete recovery is achievable. This perspective emphasizes personal empowerment, resilience, and community integration, focusing on an individual's potential rather than solely on their symptoms [6,14,15]. Dogmatic adherence to either of these viewpoints carries the risk of engendering self-realizing outcomes. Hence, therapists must balance their understanding of empirical medical data by acknowledging the vast spectrum of human potentialities [16,17]. In November 2022, the generative artificial intelligence (AI) large language model (LLM) ChatGPT-3 was launched for free public use. Subsequently, in 2023, other LLMs such as Google Bard, Claude, and ChatGPT-4 were released. Although all these LLMs have been trained on vast data sets and have undergone alignment processes, as well as learning from user feedback, their differences stem from their unique learning algorithms; the nature of their training data; and the distinct approaches to alignment, user interaction, and learning from user feedback. These LLMs have permeated various facets of society, including political science, economics, health care, and biology [18,19]. Previous studies have examined the potential of LLMs in the context of applied psychology, focusing on basic clinical abilities [20-22] or on decision-making in complex clinical situations such as depression and suicide [23-25]. To the best of our knowledge, no study to date has assessed the extent to which generative AI can facilitate cure or recovery from mental health conditions. In contrast, extensive literature highlights the immense therapeutic value of therapists' belief in their patients' ability to recover [11-13], as well as the negative effects that ensue when a therapist does not believe that the patient's condition can improve [26].

Recovery for individuals with prolonged mental health challenges is a multifaceted process subject to varied interpretations. From a clinical perspective, recovery emphasizes symptom reduction and impairment rectification [26,27]. In contrast, from lived experience, recovery represents an individualized, potentially ongoing trajectory toward reclaiming purpose, meaning, and active contribution, regardless of symptoms [27].

Years of rigorous theoretical and clinical research have revealed several mechanisms that assist patient recovery. One salient finding is the positive correlation between a strong therapeutic alliance and enhanced outcomes [28]. A meta-analysis of over 30,000 participants showed the therapeutic alliance was highly correlated with outcomes, regardless of therapy type [29]. The efficacy of psychotherapy is well documented [30-32]. Therapists' belief in treatment potential significantly impacts outcomes [33]. Over the past decade, literature has consistently emphasized recovery-oriented practices for improving patient outcomes, including enhanced functioning, goal setting, reduced legal issues, and decreased hospital admissions [34,35]. Consequently, mental health services increasingly integrate recovery paradigms into treatment strategies [36]. However, an abrupt transition from a biomedical model to recovery orientation can challenge providers, often leading to continued paternalistic decision-making [37].

With the increasing integration of AI in health care, especially given its emerging capabilities in emotion detection and mental health risk assessment [20-25], it becomes imperative to scrutinize how different LLMs interpret human recovery potential. Such an inquiry gains heightened relevance in that both patients and professionals are increasingly relying on LLMs for consultations. Not only do such insights have the potential to shape the trajectory of patient care, but they can also play a pivotal role in psychoeducational endeavors, direction, and interventions.

This research is based on an examination of the perspectives of mental health professionals in Australia [38]. The study included 342 nurses, 564 psychiatrists, 424 general practitioners (GPs), and 228 clinical psychologists. It also incorporated the insights of 982 members of the general public. Respondents were presented vignettes depicting an individual diagnosed with schizophrenia and asked to indicate their perceptions regarding prognosis, long-term outcomes, and potential discrimination.

Research Objectives

The research objectives were as follows:

1. To examine how different LLMs (ChatGPT-3.5, ChatGPT-4, Claude, and Bard) evaluate the prognosis of an individual with schizophrenia compared to the evaluations of mental health professionals (mental health nurses, clinical psychologists, psychiatrists, and GPs) and the general public.
2. To examine how different LLMs (ChatGPT-3.5, ChatGPT-4, Claude, and Bard) evaluate the positive and negative long-term outcomes of an individual with schizophrenia compared to the evaluations of mental health

professionals (mental health nurses, clinical psychologists, psychiatrist, and GPs) and the general public.

3. To compare evaluations of the prognosis and positive and negative outcomes of an individual with schizophrenia between different types of LLMs (ChatGPT-3.5, ChatGPT-4, Claude, and Bard).

Methods

AI Procedure and Data Collection

During the month of August 2023, we examined the following LLMs:

- *Bard* (Google; subsequently rebranded as Gemini) [39] uses the LaMDA language model, trained on the expansive Infiniset data set amalgamating over 1.5 trillion words from diverse web-based sources including C4-derived content, Wikipedia, programming documentation, and public forum dialogue. LaMDA was initially pretrained on extensive public and web text corpora, leveraging a transformer-based neural architecture and unsupervised learning to process language and formulate pertinent responses.
- *Claude* (Anthropic) [40] targets a beneficial, inoffensive, and truthful output using a constitutional approach. Its 12+ billion–parameter transformer model aims to ethically tackle linguistic complexity. Its training emphasized educational data curation and the alignment of model actions with human values and safety considerations, potentially enhancing reliability. A paid Claude subscription recently launched at US $20 per month. Our study used the free version.
- *ChatGPT-3.5* (OpenAI) [41] exemplifies a 175 billion–parameter transformer model demonstrating human-like language aptitude given its far-reaching internet-derived training data. It was designed to expand comprehension and text generation capabilities using immense data sets for advanced language understanding and tailored response building functionalities.
- *ChatGPT-4* (OpenAI) [42] aims to improve upon previous versions in areas such as safety, accuracy, and content policy adherence. Relative to ChatGPT-3.5, OpenAI claims an over 80% decrease in rule-violating responses and a 40% improvement in factual correctness for this latest model. As part of the training process, OpenAI solicited abundant feedback from early users and collaborated with over 50 AI ethics experts. However, full details on the architecture changes and training data incorporated have not been disclosed publicly. Without access to such technical specifics, independent auditing remains limited regarding the capabilities of ChatGPT-4. This model costs US $20 per month.

The case vignettes and a list of follow-up questions were introduced for each of the LLMs. To address the generative nature of the models, which produces different responses to similar questions, each vignette was passed through each model 10 times. We compared the LLMs' data for schizophrenia prognosis to the norms used by mental health professionals and to the views of the general public, as reported by Caldwell and Jorm [38].

Input Source

In this investigation, we adopted the text vignette methodology used by Jorm et al [43] and Caldwell and Jorm [38] to examine the beliefs of mental health professionals (mental health nurses, clinical psychologists, psychiatrists, and GPs) and the general public regarding the efficacy of interventions for schizophrenia. In the study, 2 vignettes were entered into the LLMs' interface. The 2 vignettes differed only in the gender of the main character. Each vignette was evaluated 10 times in ChatGPT-3.5, ChatGPT-4, Bard, and Claude, for a total of 80 evaluations, each in a new tab.

Vignette

The full text of the vignettes is as follows:

> *Mary/Jone is 24 and lives at home with her/his parents. She/he has a few temporary jobs since finishing school but is now unemployed. Over the last six months she/he has stopped seeing her/his friends and has begun locking herself/himself in her/his bedroom and refusing to eat with the family or to have a bath. Her/His parents also hear her/him walking about her bedroom at night while they are in bed. Even though they know she/he is alone, they have heard her/him shouting and arguing as if someone else is there. When they try to encourage her/him to do more things, she/he whispers that she/he won't leave home because she/he is being spied upon by the neighbor. They realize she/he is not taking drugs because she/he never sees anyone or goes anywhere.*

Measures

After entering each of the 2 vignettes to the LLMs, we submitted 6 questions taken from Jorm et al [43]:

> *Q1: Describe what, if anything, is wrong with Mary/Jone?*
>
> *Q2: What will be helpful for Mary/Jone?*
>
> *Q3: What is the most likely result for Mary/Jone given that she/he received the sort of professional help you believe is most appropriate? A. Full recovery with no further problems; B. Full recovery, but problems will probably recur; C. Partial recovery; D. Partial recovery, but problems will probably recur; E. No improvement; F. Get worse.*
>
> *Q4: What is the most likely result for Mary/Jone if she/he does not receive any professional help? A. Full recovery with no further problems; B. Full recovery, but problems will probably recur; C. Partial recovery; D. Partial recovery, but problems will probably recur; E. No improvement; F. Get worse.*
>
> *Q5: How do you think Mary/Jone will be in the long term compared to other people in the community? Rate each of the following statements according to this scale: more likely, just as likely, or less likely. Negative long-term outcomes: A. will be violent; B. will drink too much; C. will take illegal drugs; D. will have unsatisfactory friendships; E. will attempt suicide. Positive long term outcomes: F . will be*

understanding of other people's feelings; G. will have a good marriage; H. will be a caring parent; I. will be a productive worker. J. will be creative or artistic.

Q6. Do you think Mary/Jone will be discriminated against by others in the community if they know about her/his problems? (Yes/No).

Scoring

The performance of each LLM was scored according to Jorm et al [43] and Caldwell and Jorm [38]. We then compared the performance of the LLMs to the norms of 324 mental health nurses, 228 clinical psychologists, 567 psychiatrists, 424 GPs, and 982 people from the general public, as collected in Australia [38,43]. Q5, which evaluated the positive and negative long-term outcomes, was calculated according to Caldwell and Jorm [38]. Each of the 10 statements was scored as follows: 1=more likely, 0=just as likely, and −1=less likely. The answers were then summed up, such that each positive and negative long-term outcome score ranged from −5 to 5.

Statistical Analysis

The likely outcomes with and without professional treatment for the 2 vignettes, as evaluated by the LLMs, mental help professionals, and the general public (reported by Caldwell and Jorm [38,43]), were analyzed using 1-way ANOVA, with Fisher least significant difference applied as a post hoc analysis. The differences between the LLMs in positive and negative long-term outcomes were compared using 1-way ANOVA, with Fisher least significant difference applied as a post hoc analysis. Given the significant clinical implications of discrepancies between the evaluations of the LLM models and the professional assessments, we opted for a post hoc approach that minimizes the risk of type II errors or false negatives.

Ethical Considerations

This study was exempt from ethical review since it only evaluates AI chatbots and no human participants were involved.

Results

For all of the vignette cases, all 4 LLMs recognized schizophrenia as the primary diagnosis and suggested a combination of antipsychotic drugs and psychotherapy as the preferred treatment.

Likely Outcome With Professional Treatment

Table 1 delineates the distribution of outcomes selected by LLMs, mental health professional groups, and the general public for a vignette describing an individual diagnosed with schizophrenia after receiving professional treatment. ANOVA analysis revealed significant differences in the selected outcomes across the 8 groups ($F_{8,2601}$=33.66; P<.001). Post hoc analysis yielded the following insights. (1) The ChatGPT-3.5 model offered a distinctively pessimistic prognosis, significantly differing from the outcomes chosen by all the other LLMs (P=.02 to .007), the professional groups (P=.005 to <.001), and the general public (P<.001). (2) ChatGPT-4, Claude, and Bard projected more pessimistic prognosis outcomes than the general public (P=.02 to .007), whereas their projections were congruent with those from all the professional groups (all P>.05). A direct comparison of the projections of ChatGPT-4, Claude, and Bard yielded no significant differences (all P>.05; Figure 1 and Table 2).

Table . The likely outcome for schizophrenia, with and without professional treatment, as evaluated by LLMs[a], mental health professionals, and the general public.

Professional treatment and outcome	ChatGPT-3.5 (n=20), n (%)	ChatGPT-4 (n=20), n (%)	Bard (n=20), n (%)	Claude, (n=20) n (%)	General public (n=982), %[b]	Nurses (n=324), %[b]	Clinical psychologists (n=228), %[b]	Psychiatrists (n=567), %[b]	GPs[c] (n=424), %[b]
With professional treatment									
Full recovery with no further problems	0 (0)	0 (0)	1 (5)	0 (0)	29.8	8.8	3.1	2	3.1
Full recovery, but problems would probably reoccur	0 (0)	5 (25)	4 (20)	(35)	44.4	61.4	49.1	51.6	56.1
Partial recovery	10 (50)	15 (75)	15 (75)	(65)	10.2	4.1	11.9	5.7	5
Partial recovery, but problems would probably reoccur	10 (50)	0 (0)	0 (0)	0 (0)	14.3	25.7	35.4	40.6	35.8
No improvement	0 (0)	0 (0)	0 (0)	0 (0)	0.7	0	0.4	0.2	0
Get worse	0 (0)	0 (0)	0 (0)	0 (0)	0.6	0	0	0	0
Without professional treatment									
Full recovery with no further problems	0 (0)	0 (0)	0 (0)	0 (0)	1.1	0	0	0	1.1
Full recovery, but problems would probably reoccur	0 (0)	0 (0)	0 (0)	0 (0)	1.7	0.9	0.9	0.7	1.7
Partial recovery	0 (0)	0 (0)	0 (0)	0 (0)	1.8	0.6	0.9	0.9	1.8
Partial recovery, but problems would probably reoccur	0 (0)	0 (0)	0 (0)	0 (0)	4.9	11.3	9.1	5.7	4.9
No improvement	3 (15)	0 (0)	0 (0)	0 0)	15.1	9.5	17.8	11	15.1
Get worse	17 (85)	20 (100)	20 (100)	20 (100)	75.4	77.7	71.3	81.8	75.4

[a]LLM: large language model.

[b]As reported by Caldwell and Jorm [38].

[c]GP: general practitioner.

Figure 1. The likely outcome for schizophrenia, with and without professional treatment, as evaluated by large language models, mental health professionals, and the general public (mean and SE). **P*<.05. GP: general practitioner.

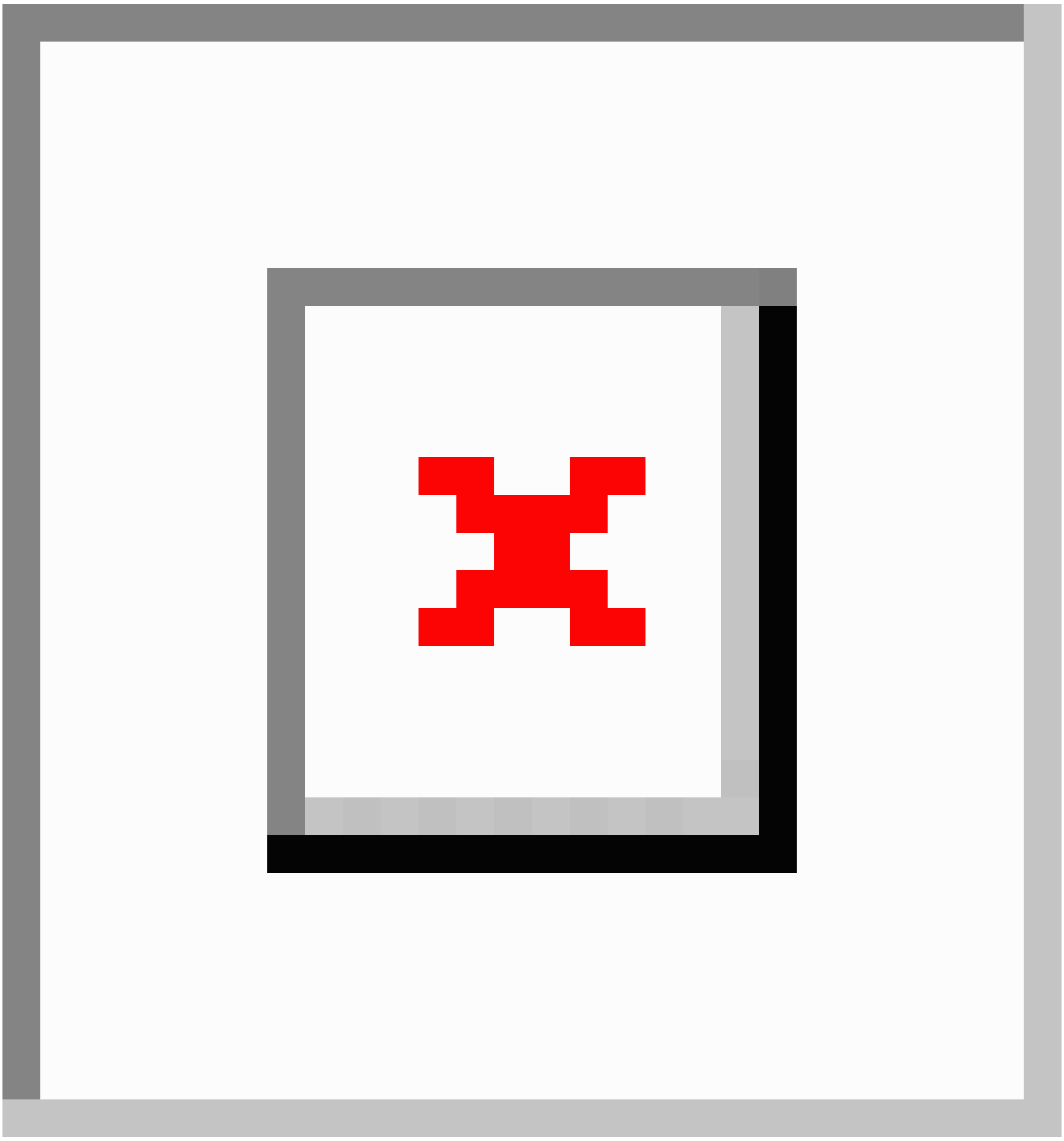


Table . Least significant difference post hoc analyses for LLMs[a], mental health professionals, and the general public in assessing the outcome of schizophrenia with and without treatment.

P values	ChatGPT-3.5	ChatGPT-4	Bard	Claude	General public	Nurses	Clinical psychologists	Psychiatrists	GPs[b]
With professional treatment									
ChatGPT-3.5	—[c]	.02	.01	.007	<.001	<.001	.003	.005	<.001
ChatGPT-4	.02	—	.87	.75	.007	.21	.79	.64	.95
Bard	.01	.87	—	.87	.01	.30	.63	.49	.87
Claude	.007	.75	.87	—	.02	.41	.49	.37	.70
General public	<.001	.007	.01	.02	—	<.001	<.001	<.001	<.001
Nurses	<.001	.21	.30	.41	<.001	—	<.001	<.001	<.001
Clinical psychologists	.003	.79	.63	.49	<.001	<.001	—	.58	.08
Psychiatrists	.005	.64	.49	.37	<.001	<.001	.58	—	.06
GPs	<.001	.95	.87	.70	<.001	<.001	.08	.06	—
Without professional treatment									
ChatGPT-3.5	—	.55	.55	.55	.12	.21	.14	.46	.48
ChatGPT-4	.55	—	>.99	>.99	.02	.04	.02	.11	.13
Bard	.55	>.99	—	>.99	.02	.04	.02	.11	.13
Claude	.55	>.99	>.99	—	.02	.04	.02	.11	.13
General public	.12	.02	.02	.02	—	.35	.89	<.001	<.001
Nurses	.21	.04	.04	.04	.35	—	<.001	.07	.07
Clinical psychologists	.14	.02	.02	.02	.89	<.001	—	.03	.03
Psychiatrists	.46	.11	.11	.11	<.001	.07	.03	—	<.001
GPs	.48	.13	.13	.13	<.001	.07	.03	<.001	—

[a]LLM: large language model.

[b]GP: general practitioner.

[c]Not applicable.

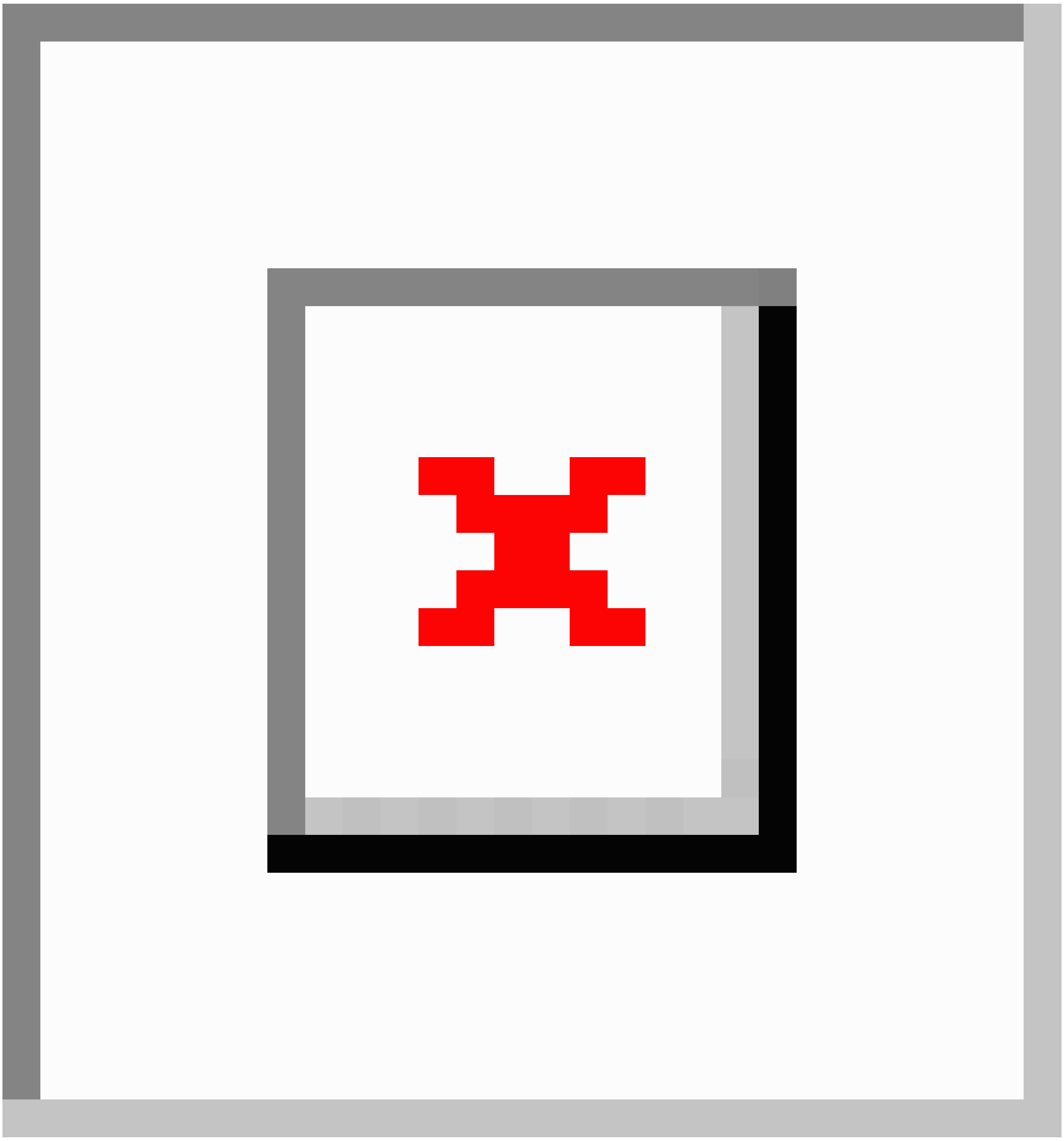
Likely Outcome Without Professional Treatment

Table 1 also delineates the distribution of outcomes selected by LLMs, mental health professional groups, and the general public for a vignette describing an individual with schizophrenia who did not receive professional treatment. All groups indicated that without treatment, the person with schizophrenia would show no improvement or would get worse. ANOVA analysis revealed a significant difference in the selected outcomes across the 8 groups ($F_{8,2601}$=4.07; P<.001). Post hoc analysis yielded the following insights. (1) The ChatGPT-4, Claude, and Bard models offered a distinctively pessimistic prognosis, significantly differing from the outcomes chosen by mental health nurses (P=.04), clinical psychologists (P=.02), and the general public (P=.11) but not significantly different from the outcomes selected by ChatGPT-3.5, psychiatrists, and GPs (all P>.05). Direct comparison between ChatGPT-4, Claude, and Bard yielded no significant differences in prognosis (all P>.05). (2) No significant difference was observed between ChatGPT-3.5, the professional groups, and the general public (all P>.05; Figure 1 and Table 2).

Long-Term Outcomes

Figure 2 illustrates the LLMs' output concerning positive and negative long-term outcomes. ANOVA analysis revealed a significant difference in the negative outcomes selected across the 4 LLMs groups ($F_{3,76}$=18.32; P<.001). ChatGPT-4 and Claude indicated a significantly higher likelihood of negative long-term outcomes for patients after professional treatment than Bard and ChatGPT-3.5 (ChatGPT-4 vs Bard: P=.004; ChatGPT-4 vs ChatGPT-3.5: P<.001; Claude vs Bard: P=.003; Claude vs ChatGPT-3.5: P<.001). In addition, Bard was significantly more pessimistic and indicated a higher likelihood of negative long-term outcomes than ChatGPT-3.5 (P=.001). ANOVA analysis revealed a significant difference in the positive outcomes selected by the 4 LLMs groups ($F_{3,76}$=24.45; P<.001). ChatGPT-3.5 and Claude were significantly more pessimistic and indicated a lower likelihood of positive long-term outcomes for patients after treatment than Bard and ChatGPT-4 (ChatGPT-3.5 vs Bard: P<.001; ChatGPT-3.5 vs ChatGPT-4: P<.001; Claude vs Bard: P<.001; Claude vs ChatGPT-4: P<.001). No significant differences were found between ChatGPT-3.5 and Claude (P=.92) or between ChatGPT-4 and Bard (P=.51).

Figure 2. The positive and negative long-term outcomes evaluated by large language models (ChatGPT-3.5, ChatGPT-4, Bard, and Claude; mean and SE).



Discrimination

For all the cases in the vignettes, all 4 LLMs determined that the person with schizophrenia described in the vignette would be discriminated against.

Discussion

Principal Findings

This investigation pursued 2 primary objectives. First, we aimed to evaluate how various LLMs assessed the prognosis of individuals with schizophrenia compared to the evaluations of mental health professionals (mental health nurses, clinical psychologists, psychiatrists, and GPs) and the views of the general public. Second, we sought to compare these assessments of prognosis as well as positive and negative long-term outcomes across the different types of LLMs.

The academic discourse in contemporary schizophrenia research often focuses on the deployment of AI within professional scientific contexts, yet it seldom addresses the accessibility of AI to the general public or the patient population. LLMs are being used today by hundreds of millions of users worldwide, including patients and clinicians. In the mental health field, this widespread use has awakened an urgent need to examine the quality of clinical information these systems provide on various medical issues, such as treatment strategy recommendations

[24], risk assessment [23,25], and the interpretation of emotional states [20,21]. Machine learning algorithms possess the capability to discern nuanced variables associated with divergent disease trajectories [44]. Such algorithms facilitate probabilistic prediction of specific outcomes at the individual level, as well as the identification of distinct subgroups within a broader diagnostic category [45]. Consequently, machine learning methodologies hold promise for aiding clinicians in formulating individualized interventions, thereby mitigating the risk of a detrimental disease progression [46]. This study examines an issue not previously investigated—the ability to predict the clinical prognosis of a severe chronic illness such as schizophrenia using LLMs.

Likely Outcome With Professional Treatment

In this study, we identified significant differences in the outcomes suggested across the 8 groups. The ChatGPT-3.5 model exhibited a notably pessimistic prognosis for individuals with schizophrenia with professional treatment relative to all other LLMs, professional groups, and the general public. Given the widespread use of ChatGPT-3.5, these findings have substantial clinical implications. Any inclination toward pessimistic forecasting might influence a patient's willingness to undergo treatments, including both antipsychotic medication and psychotherapy, particularly in the context of schizophrenia. If patients or their families consult the ChatGPT-3.5 model for prognostic insights, these less-than-optimistic outcomes might sway their decision-making regarding whether to initiate or continue therapeutic interventions. The literature previously highlighted challenges in adherence to antipsychotic and psychotherapy treatments due to their cost and other factors [47,48]. Additionally, more negative perceptions of potential treatment outcomes might inadvertently influence the effectiveness of the therapeutic process, regardless of the mode of intervention.

The congruence between the prognostic assessments of various LLMs models (ChatGPT-4, Bard, and Claude) and those of clinical professionals is encouraging. From a clinical perspective, precise prognostication is paramount. It not only informs clinicians in tailoring interventions that balance potential risks and benefits but also empowers patients with the knowledge needed to make informed choices about their treatments while considering the inherent risks associated with the intervention and the disease's progression. The finding that 3 prominent LLMs yielded comparable estimates that align closely with the evaluations of 3 groups of experienced professionals (GPs, psychiatrists, and clinical psychologists) offers a foundation for optimism. Such consistency in predictive capabilities suggests the potential for integrating these insights into clinical decision support systems, reinforcing the centrality of accurate prognostication in medical decision-making.

This observation substantiates initial results in the domain of mental health research gleaned from the use of the ChatGPT-3.5 model. Existing methodologies often exhibit constrained predictive proficiencies. In a recent study, Elyoseph and Levkovich [25] found that ChatGPT-3.5 often underestimated the risk of suicidal ideation, thus calling into question its reliability in such critical assessments. Another study by Imran et al [49] posited that while ChatGPT may significantly influence pediatric and adolescent mental health care as a supplementary tool, it would be inadvisable and impracticable to contend that it could entirely supplant human clinical discernment. Indeed, although the utility of ChatGPT in mental health spheres appears promising, significant reservations remain. Another study suggested that ChatGPT-4 estimates the likelihood of suicide attempts in a manner akin to evaluations provided by professionals, whereas ChatGPT-3.5 frequently underestimates suicide risk [23]. For instance, ChatGPT's learning mechanisms, which rely on web-based data and human feedback, have the potential to disseminate inaccurate or inappropriate guidance if not rigorously evaluated. Such drawbacks are especially disturbing when considering their impact on individuals grappling with mental health disorders [50].

ChatGPT-4, Bard, and Claude have each instituted measures aimed at forestalling malevolent use and attenuating biases inherent in their respective models; however, challenges persist in ascertaining how these technologies should be responsibly used. The intrinsic worth of the generative output produced by LLMs is the subject of scholarly contention. Some researchers, such as Winkler et al [51], posit that LLMs may actually constitute a deceptive or even perilous risk due to their capacity to fabricate an appearance of comprehension, sentience, and analytical depth in the absence of an authentic world model. Medical studies that compared different LLMs found that ChatGPT-4 and Bard aligned with doctors' diagnoses [52]. Another study [53] sought to assess the performance of 4 LLMs (Claude, Bard, ChatGPT-4, and New Bing) in the context of medical consultations related to urolithiasis. Simulated clinical scenarios revealed that all the models except Bard provided relatively competent answers. Claude consistently excelled in various evaluative metrics, whereas ChatGPT-4 ranked second in accuracy and demonstrated stable output across tests.

Likely Outcome Without Professional Treatment

In this study, all groups expressed the belief that in the absence of medical intervention, an individual diagnosed with schizophrenia would either demonstrate no improvement or would deteriorate. This assessment is similar to the evaluation of psychiatrists and GPs and is consistent with the literature and clinical knowledge [38,43]. We suggest that these assessments, although slightly more pessimistic than those of clinical psychologists, nurses, and the general public, have a positive influence because they emphasize the risk of untreated illness and indirectly encourage treatment.

To the best of our knowledge, no studies have examined comparison between these LLMs in this context of mental health. Nevertheless, initial studies that compared professionals in the field of therapy and medicine reinforce these findings. For example, in a scholarly investigation encompassing 82 clinical descriptions [54], the diagnostic accuracy rates of physicians were found to surpass those of Bard. This outcome indicates that Bard needs further enhancement and fine-tuning in its diagnostic proficiencies. Another possible explanation for the findings is that there are fundamental differences between the various algorithms. These algorithms were trained on different amounts and qualities of data, underwent different processes of elimination, and use distinct strategies for receiving feedback from system users [55].

Long-Term Outcomes

In the case of assessing long-term outcomes, 3 of the models—ChatGPT-4, Bard, and Claude—paralleled the conclusions reached by mental health professionals [38,43]. The models pointed to a higher likelihood of negative long-term outcomes and a decreased probability of positive ones. ChatGPT-3.5, which projected a decline in negative long-term symptoms over time, is an anomaly. Apart from this exception, the evaluations of the 3 models and the determinations of mental health specialists exhibit consistent alignment. An analysis of the differences among the 3 revealed that Claude has the most conservative or pessimistic stance, ChatGPT-4's predictions are midway between pessimistic and optimistic, and Bard exhibits the most optimistic forecasting. These results again underscore the potential of LLMs models to offer prognostic insights that might be incorporated into future medical decision-making processes.

Real-World Application Potential

This investigation presents initial discoveries regarding the potential of LLMs in offering prognostic forecasts for schizophrenia. It is of utmost importance to approach these findings with caution, considering the potential fragility of these models over time and the limited scenarios analyzed in the study, which do not fully encompass the range of symptoms, medical histories, and individual variations. Moreover, the study does not explore LLM predictions across various treatment strategies. Nevertheless, by adopting a careful approach, we strive to elucidate the future potential of using these capabilities in real-world clinical settings through further research. One potential avenue for integrating LLMs into clinical practice is by using them as a "co-pilot" that aids clinicians by providing pertinent information. For instance, LLM systems could potentially offer prognostic evaluations based on symptom descriptions during intake, summarized reports of visits, or transcriptions of conversations with clinicians. Clinicians could use this information to align expectations with patients regarding their prognosis or to tailor treatment, taking into account the implications on patients' lives. It is important to note that although theoretically possible, the ability of AI to provide patient-specific prognoses, which could potentially enhance treatment protocols and align expectations between patients and caregivers, remains to be empirically demonstrated. Another option is the direct use of LLMs by patients and family members as part of a psychoeducational process to familiarize themselves with the illness and its potential consequences. This approach can enhance collaboration and engagement in the treatment process.

Lastly, AI systems have the capability to process auditory information, such as a case narrative, and generate a prognosis based on it. There exists potential to convert this qualitative, subjective information into an objective, mathematical analysis. Essentially, AI takes the primary input received by a physician—the patient's narrative of their illness—and objectively analyzes it rather than subjectively. This has the potential to enhance the reliability of assessment processes in the field of psychiatry. By combining such tools with additional data, it is possible that prognoses can be further improved. Future research can explore the combined impact of artificial and human predictions and incorporate questionnaires to refine the predictive outcome of disease progression.

Limitations

This research is not without limitations that necessitate explicit acknowledgment. First, since the study tested the performance of LLMs at one point in time, it is necessary to examine the consistency of the results when software updates are released. Second, the data pertaining to AI were juxtaposed with information gleaned from a sample of professionals and the general populace in a single study in Australia. This sample, however, does not offer global representation. Future investigations are recommended to encompass a more extensive array of variables, such as socioeconomic indicators, cultural determinants, and mental health histories, particularly with regard to recovery from schizophrenia. Furthermore, the vignettes used in the study, including those featuring individuals with schizophrenia, fail to present a nuanced, ongoing, and comprehensive medical treatment context. They also do not include variables that would be readily available to medical professionals during therapeutic sessions. To enhance the generalizability and rigor of subsequent studies, it is advised to incorporate additional variables, deploy more sophisticated language models, evaluate data at varying temporal intervals, and juxtapose the findings with a more diverse assortment of clinical samples. An additional constraint involves ethical considerations in professionals' use of AI. The literature reveals public skepticism and concerns about medical inaccuracies and potential discrimination [56,57]. Ethical issues such as patient autonomy and health disparities necessitate exercising caution in AI's medical applications [58-62]. Lastly, the rapidly evolving landscape of AI poses an inherent obstacle to drawing conclusions about the technology's long-term, stable capabilities. To address this concerns, future research is required. To enhance the accuracy of LLMs in psychiatric assessment, future research should focus on enriching training data sets with specialized, targeted data, including historical clinical knowledge and detailed patient histories. Validating these models against current clinical practices and decisions made by practicing psychiatrists can provide a practical benchmark for their performance. Additionally, exploring technological advancements in AI, particularly in deep learning, can refine LLMs to process complex psychiatric data more effectively. Modifying prompts and inputs to better reflect psychiatric assessments can also improve the models' understanding and interpretation of clinical scenarios. Interdisciplinary collaboration involving AI researchers, clinicians, and ethicists is essential to align the development of LLMs with clinical needs and ethical standards. Investigating the integration of LLMs with human expertise, through interactive systems that allow clinicians to provide feedback on LLM predictions, is crucial for a dynamic learning process. Exploring the use of LLMs across diverse clinical environments and patient populations can help identify and mitigate potential biases, ensuring equitable and broadly applicable models. Longitudinal studies tracking LLM performance over time in various clinical contexts will provide insights into long-term efficacy and areas for improvement. These research initiatives can significantly advance the field of LLMs in psychiatry, enhancing their accuracy, reliability, and practical utility in clinical settings.

Conclusion

This study offers novel and clinically relevant insights into the assessment capabilities of prominent LLMs regarding the prognosis and long-term outcomes of schizophrenia. The findings highlight both the promise and current limitations of AI in augmenting clinical evaluations. Further research is warranted to refine the algorithms and better integrate human expertise, thereby maximizing the judicious and ethical use of AI in mental health care.

Acknowledgments

This research received no external funding.

Data Availability

The data that support the findings of this study are available from the authors upon reasonable request.

Authors' Contributions

ZE and IL contributed to conceptualization, writing—original draft preparation, and writing—review and editing. ZE contributed to methodology and formal analysis. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

References

1. McCutcheon RA, Reis Marques T, Howes OD. Schizophrenia—an overview. JAMA Psychiatry 2020 Feb 1;77(2):201-210. [doi: 10.1001/jamapsychiatry.2019.3360] [Medline: 31664453]
2. Correll CU, Schooler NR. Negative symptoms in schizophrenia: a review and clinical guide for recognition, assessment, and treatment. Neuropsychiatr Dis Treat 2020 Feb 21;16:519-534. [doi: 10.2147/NDT.S225643] [Medline: 32110026]
3. Habtewold TD, Tiles-Sar N, Liemburg EJ, et al. Six-year trajectories and associated factors of positive and negative symptoms in schizophrenia patients, siblings, and controls: Genetic Risk and Outcome of Psychosis (GROUP) study. Sci Rep 2023 Jun 9;13(1):9391. [doi: 10.1038/s41598-023-36235-9] [Medline: 37296301]
4. Rahme C, El Kadri N, Haddad C, Fekih-Romdhane F, Obeid S, Hallit S. Exploring the association between lifetime traumatic experiences and positive psychotic symptoms in a group of long-stay patients with schizophrenia: the mediating effect of depression, anxiety, and distress. BMC Psychiatry 2023 Jan 12;23(1):29. [doi: 10.1186/s12888-023-04531-3] [Medline: 36635691]
5. Menezes NM, Arenovich T, Zipursky RB. A systematic review of longitudinal outcome studies of first-episode psychosis. Psychol Med 2006 Oct;36(10):1349-1362. [doi: 10.1017/S0033291706007951] [Medline: 16756689]
6. Warner R. Recovery from Schizophrenia: Psychiatry and Political Economy, 3rd edition: Routledge; 2013. [doi: 10.4324/9780203420874]
7. Jääskeläinen E, Juola P, Hirvonen N, et al. A systematic review and meta-analysis of recovery in schizophrenia. Schizophr Bull 2013 Nov;39(6):1296-1306. [doi: 10.1093/schbul/sbs130] [Medline: 23172003]
8. Sullivan PW, Starnino VR, Raster CG. In the eye of the beholder: recovery and personal narrative. J Psychosoc Rehabil Ment Health 2017 Dec;4(2):221-229. [doi: 10.1007/s40737-017-0097-6]
9. Wittchen HU, Mühlig S, Beesdo K. Mental disorders in primary care. Dialo Clin Neurosci 2003 Jun 30;5(2):115-128. [doi: 10.31887/DCNS.2003.5.2/huwittchen]
10. Kennedy S, Lanceley A, Whitten M, Kelly C, Nicholls J. Consent on the labour ward: a qualitative study of the views and experiences of healthcare professionals. Eur J Obstet Gynecol Reprod Biol 2021 Sep;264:150-154. [doi: 10.1016/j.ejogrb.2021.07.003] [Medline: 34303075]
11. Fimiani R, Gazzillo F, Gorman B, et al. The therapeutic effects of the therapists' ability to pass their patients' tests in psychotherapy. Psychother Res 2023 Jul;33(6):729-742. [doi: 10.1080/10503307.2022.2157227] [Medline: 36574276]
12. Patterson JE, Edwards TM, Vakili S. Global mental health: a call for increased awareness and action for family therapists. Fam Process 2018 Mar;57(1):70-82. [doi: 10.1111/famp.12281] [Medline: 28205225]
13. Wampold BE, Flückiger C. The alliance in mental health care: conceptualization, evidence and clinical applications. World Psychiatry 2023 Feb;22(1):25-41. [doi: 10.1002/wps.21035] [Medline: 36640398]
14. Babcock G, McShea DW. Resolving teleology’s false dilemma. Biolo J Linnean Soc 2023 Aug;139(4):415-432. [doi: 10.1093/biolinnean/blac058]

15. Hochstetter A, Vernekar R, Austin RH, et al. Deterministic lateral displacement: challenges and perspectives. ACS Nano 2020 Sep 22;14(9):10784-10795. [doi: 10.1021/acsnano.0c05186] [Medline: 32844655]
16. Hong SJ. Cross-cultural differences in the influences of spiritual and religious tendencies on beliefs in genetic determinism and family health history communication: a teleological approach. J Relig Health 2019 Oct;58(5):1516-1536. [doi: 10.1007/s10943-018-0729-5] [Medline: 30446868]
17. Wilks D. Twenty-first-century counseling theory development in relation to definitions of free will and determinism. Journal of Counseling & Development 2018 Apr;96(2):213-222. [doi: 10.1002/jcad.12194]
18. Ali O, Abdelbaki W, Shrestha A, Elbasi E, Alryalat MAA, Dwivedi YK. A systematic literature review of artificial intelligence in the healthcare sector: benefits, challenges, methodologies, and functionalities. Journal of Innovation & Knowledge 2023;8(1):100333. [doi: 10.1016/j.jik.2023.100333]
19. Mariani MM, Machado I, Nambisan S. Types of innovation and artificial intelligence: a systematic quantitative literature review and research agenda. J Bus Res 2023 Jan;155:113364. [doi: 10.1016/j.jbusres.2022.113364]
20. Elyoseph Z, Hadar-Shoval D, Asraf K, Lvovsky M. ChatGPT outperforms humans in emotional awareness evaluations. Front Psychol 2023 May 26;14:1199058. [doi: 10.3389/fpsyg.2023.1199058] [Medline: 37303897]
21. Hadar-Shoval D, Elyoseph Z, Lvovsky M. The plasticity of ChatGPT's mentalizing abilities: personalization for personality structures. Front Psychiatry 2023 Sep 1;14:1234397. [doi: 10.3389/fpsyt.2023.1234397] [Medline: 37720897]
22. Elyoseph Z, Refoua E, Asraf K, Lvovsky M, Shimoni Y, Hadar-Shoval D. Capacity of generative artificial intelligence to interpret human emotions from visual and textual data: pilot evaluation study. JMIR Ment Health 2024 Feb 6;11:e54369. [doi: 10.2196/54369] [Medline: 38319707]
23. Levkovich I, Elyoseph Z. Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: vignette study. JMIR Ment Health 2023 Sep 20;10(2):e51232. [doi: 10.2196/51232] [Medline: 37728984]
24. Levkovich I, Elyoseph Z. Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physician. Fam Med Com Health 2023 Sep;11(4):e002391. [doi: 10.1136/fmch-2023-002391] [Medline: 37844967]
25. Elyoseph Z, Levkovich I. Beyond human expertise: the promise and limitations of ChatGPT in suicide risk assessment. Front Psychiatry 2023 Aug 1;14:1213141. [doi: 10.3389/fpsyt.2023.1213141] [Medline: 37593450]
26. Gunasekaran S, Tan GTH, Shahwan S, Goh CMJ, Ong WJ, Subramaniam M. The perspectives of healthcare professionals in mental health settings on stigma and recovery - a qualitative inquiry. BMC Health Serv Res 2022 Jul 9;22(1):888. [doi: 10.1186/s12913-022-08248-z] [Medline: 35804378]
27. Slade M, Amering M, Farkas M, et al. Uses and abuses of recovery: implementing recovery-oriented practices in mental health systems. World Psychiatry 2014 Feb;13(1):12-20. [doi: 10.1002/wps.20084] [Medline: 24497237]
28. Cuijpers P, Reijnders M, Huibers MJH. The role of common factors in psychotherapy outcomes. Annu Rev Clin Psychol 2019 May 7;15:207-231. [doi: 10.1146/annurev-clinpsy-050718-095424] [Medline: 30550721]
29. Flückiger C, Del Re AC, Wampold BE, Horvath AO. The alliance in adult psychotherapy: a meta-analytic synthesis. Psychotherapy (Chic) 2018 Dec;55(4):316-340. [doi: 10.1037/pst0000172] [Medline: 29792475]
30. Barkham M, Lambert MJ. The efficacy and effectiveness of psychological therapies. In: Lambert MJ, editor. Bergin and Garfield's Handbook of Psychotherapy and Behavior Change, 6th edition: John Wiley & Sons; 2021:393-442.
31. Cuijpers P, Quero S, Noma H, et al. Psychotherapies for depression: a network meta-analysis covering efficacy, acceptability, and long-term outcomes of all main treatment types. World Psychiatry 2021 Jun;20(2):283-293. [doi: 10.1002/wps.20860] [Medline: 34002502]
32. Dragioti E, Karathanos V, Gerdle B, Evangelou E. Does psychotherapy work? an umbrella review of meta-analyses of randomized controlled trials. Acta Psychiatr Scand 2017 Sep;136(3):236-246. [doi: 10.1111/acps.12713] [Medline: 28240781]
33. Zilcha-Mano S. Toward personalized psychotherapy: the importance of the trait-like/state-like distinction for understanding therapeutic change. Am Psychol 2021 Apr;76(3):516-528. [doi: 10.1037/amp0000629] [Medline: 32658495]
34. White C, Frimpong E, Huz S, Ronsani A, Radigan M. Effects of the Personalized Recovery Oriented Services (PROS) program on hospitalizations. Psychiatr Q 2018 Jun;89(2):261-271. [doi: 10.1007/s11126-017-9531-x] [Medline: 28971347]
35. Wong DFK, Chan V, Ip P, Zhuang XY. The effects of recovery-oriented cognitive-behavior approach for Chinese with severe mental illness. Res Soc Work Pract 2019 Mar;29(3):311-322. [doi: 10.1177/1049731517732837]
36. Kidd SA, Mckenzie KJ, Virdee G. Mental health reform at a systems level: widening the lens on recovery-oriented care. Can J Psychiatry 2014 May;59(5):243-249. [doi: 10.1177/070674371405900503] [Medline: 25007277]
37. Jørgensen K, Rasmussen T, Hansen M, Andreasson K, Karlsson B. Recovery-oriented intersectoral care in mental health: as perceived by healthcare professionals and users. Int J Environ Res Public Health 2020 Nov 26;17(23):8777. [doi: 10.3390/ijerph17238777] [Medline: 33255970]
38. Caldwell TM, Jorm AF. Mental health nurses' beliefs about likely outcomes for people with schizophrenia or depression: a comparison with the public and other healthcare professionals. Aust N Z J Ment Health Nurs 2001 Mar;10(1):42-54. [doi: 10.1046/j.1440-0979.2001.00190.x] [Medline: 11421972]
39. Bard. Google. URL: https://bard.google.com/ [accessed 2024-03-01]
40. Claude. Antropic. URL: https://claude.ai/ [accessed 2024-03-01]
41. ChatGPT-3.5. OpenAI. URL: https://chat.openai.com/ [accessed 2024-03-01]
42. ChatGPT-4. OpenAI. URL: https://openai.com/gpt-4 [accessed 2024-03-01]

43. Jorm AF, Korten AE, Jacomb PA, Rodgers B, Pollitt P. Beliefs about the helpfulness of interventions for mental disorders: a comparison of general practitioners, psychiatrists and clinical psychologists. Aust N Z J Psychiatry 1997 Dec;31(6):844-851. [doi: 10.3109/00048679709065510] [Medline: 9483257]
44. Tyagi A, Singh VP, Gore MM. Towards artificial intelligence in mental health: a comprehensive survey on the detection of schizophrenia. Multimed Tools Appl 2023 May;82(13):20343-20405. [doi: 10.1007/s11042-022-13809-9]
45. Rabelo-da-Ponte FD, de Azevedo Cardoso T, Kapczinski F, Passos IC. The use of artificial intelligence to identify trajectories of severe mental disorders. In: Passos IC, Rabelo-de-Ponte FD, Kapczinski F, editors. Digital Mental Health: A Practitioner's Guide: Springer; 2023:207-221. [doi: 10.1007/978-3-031-10698-9_13]
46. Paliwal B, Rahman KF. Role of artificial intelligence and machine learning in schizophrenia-a survey. In: Masoodi F, Quasim M, Bukhari S, et al, editors. Applications of Machine Learning and Deep Learning on Biological Data: CRC Press/Taylor & Francis; 2023:21-40. [doi: 10.1201/9781003328780-2]
47. Hasan A, Falkai P, Wobrock T, et al. World Federation of Societies of Biological Psychiatry (WFSBP) guidelines for biological treatment of schizophrenia, part 2: update 2012 on the long-term treatment of schizophrenia and management of antipsychotic-induced side effects. World J Biol Psychiatry 2013 Feb;14(1):2-44. [doi: 10.3109/15622975.2012.739708] [Medline: 23216388]
48. Lieberman JA, Stroup TS, McEvoy JP, et al. Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. N Engl J Med 2005 Sep 22;353(12):1209-1223. [doi: 10.1056/NEJMoa051688] [Medline: 16172203]
49. Imran N, Hashmi A, Imran A. Chat-GPT: opportunities and challenges in child mental healthcare. Pak J Med Sci 2023;39(4):1191-1193. [doi: 10.12669/pjms.39.4.8118] [Medline: 37492313]
50. Singh OP. Artificial intelligence in the era of ChatGPT - opportunities and challenges in mental health care. Indian J Psychiatry 2023 Mar;65(3):297-298. [doi: 10.4103/indianjpsychiatry.indianjpsychiatry_112_23] [Medline: 37204980]
51. Winkler C, Hammoda B, Noyes E, van Gelderen M. Entrepreneurship education at the dawn of generative artificial intelligence. Entrepreneurship Education and Pedagogy 2023 Oct;6(4):579-589. [doi: 10.1177/25151274231198799]
52. Koga S, Martin NB, Dickson DW. Evaluating the performance of large language models: ChatGPT and Google Bard in generating differential diagnoses in clinicopathological conferences of neurodegenerative disorders. Brain Pathol 2023 Aug 8:e13207. [doi: 10.1111/bpa.13207] [Medline: 37553205]
53. Song H, Xia Y, Luo Z, et al. Evaluating the performance of different large language models on health consultation and patient education in urolithiasis. J Med Syst 2023 Nov 24;47(1):125. [doi: 10.1007/s10916-023-02021-3] [Medline: 37999899]
54. Hirosawa T, Mizuta K, Harada Y, Shimizu T. Comparative evaluation of diagnostic accuracy between Google Bard and physicians. Am J Med 2023 Nov;136(11):1119-1123. [doi: 10.1016/j.amjmed.2023.08.003] [Medline: 37643659]
55. Rudolph J, Tan S, Tan S. War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. the new AI gold rush and its impact on higher education. Journal of Applied Learning & Teaching 2023 Apr 24;6(1):364-389. [doi: 10.37074/jalt.2023.6.1.23]
56. Couture V, Roy MC, Dez E, Laperle S, Bélisle-Pipon JC. Ethical implications of artificial intelligence in population health and the public's role in its governance: perspectives from a citizen and expert panel. J Med Internet Res 2023 Apr 27;25:e44357. [doi: 10.2196/44357] [Medline: 37104026]
57. Richardson JP, Smith C, Curtis S, et al. Patient apprehensions about the use of artificial intelligence in healthcare. NPJ Digit Med 2021 Sep 21;4(1):140. [doi: 10.1038/s41746-021-00509-1] [Medline: 34548621]
58. Karabacak M, Ozkara BB, Margetis K, Wintermark M, Bisdas S. The advent of generative language models in medical education. JMIR Med Educ 2023 Jun 6;9:e48163. [doi: 10.2196/48163] [Medline: 37279048]
59. Asman O, Tal A, Barilan YM. Conversational artificial intelligence-patient alliance Turing test and the search for authenticity. Am J Bioeth 2023 May;23(5):62-64. [doi: 10.1080/15265161.2023.2191046] [Medline: 37130413]
60. Tal A, Elyoseph Z, Haber Y, et al. The artificial third: utilizing ChatGPT in mental health. Am J Bioeth 2023 Oct;23(10):74-77. [doi: 10.1080/15265161.2023.2250297] [Medline: 37812102]
61. Elyoseph Z, Hadar Shoval D, Levkovich I. Beyond personhood: ethical paradigms in the generative artificial intelligence era. Am J Bioeth 2024 Jan;24(1):57-59. [doi: 10.1080/15265161.2023.2278546] [Medline: 38236857]
62. Haber Y, Levkovich I, Hadar-Shoval D, Elyoseph Z. The artificial third: a broad view of the effects of introducing generative artificial intelligence on psychotherapy. JMIR Ment Health. Preprint posted online on Nov 22, 2023. [doi: 10.2196/preprints.54781]

Abbreviations

AI: artificial intelligence
GP: general practitioner
LLM: large language model

Edited by A Tal; submitted 24.09.23; peer-reviewed by AJ Li, E Xu; revised version received 24.01.24; accepted 11.02.24; published 18.03.24.

Please cite as:
Elyoseph Z, Levkovich I
Comparing the Perspectives of Generative AI, Mental Health Experts, and the General Public on Schizophrenia Recovery: Case Vignette Study
JMIR Ment Health 2024;11:e53043
URL: https://mental.jmir.org/2024/1/e53043
doi:10.2196/53043

Original Paper

Capacity of Generative AI to Interpret Human Emotions From Visual and Textual Data: Pilot Evaluation Study

Zohar Elyoseph[1,2], BA, MA, PhD; Elad Refoua[3], BA, MA; Kfir Asraf[4], BA, MA; Maya Lvovsky[4], BA; Yoav Shimoni[5], BA, MA, PhD; Dorit Hadar-Shoval[4], BA, MA, PhD

[1]Department of Educational Psychology, The Center for Psychobiological Research, The Max Stern Yezreel Valley College, Emek Yezreel, Israel

[2]Imperial College London, London, United Kingdom

[3]Department of Psychology, Bar-Ilan University, Ramat Gan, Israel

[4]Department of Psychology, The Max Stern Yezreel Valley College, Emek Yezreel, Israel

[5]Boston Children's Hospital, Boston, MA, United States

Corresponding Author:
Zohar Elyoseph, BA, MA, PhD
Imperial College London
Fulham Palace Road
London, W6 8RF
United Kingdom
Phone: 44 547836088
Email: zohar.j.a@gmail.com

Abstract

Background: Mentalization, which is integral to human cognitive processes, pertains to the interpretation of one's own and others' mental states, including emotions, beliefs, and intentions. With the advent of artificial intelligence (AI) and the prominence of large language models in mental health applications, questions persist about their aptitude in emotional comprehension. The prior iteration of the large language model from OpenAI, ChatGPT-3.5, demonstrated an advanced capacity to interpret emotions from textual data, surpassing human benchmarks. Given the introduction of ChatGPT-4, with its enhanced visual processing capabilities, and considering Google Bard's existing visual functionalities, a rigorous assessment of their proficiency in visual mentalizing is warranted.

Objective: The aim of the research was to critically evaluate the capabilities of ChatGPT-4 and Google Bard with regard to their competence in discerning visual mentalizing indicators as contrasted with their textual-based mentalizing abilities.

Methods: The Reading the Mind in the Eyes Test developed by Baron-Cohen and colleagues was used to assess the models' proficiency in interpreting visual emotional indicators. Simultaneously, the Levels of Emotional Awareness Scale was used to evaluate the large language models' aptitude in textual mentalizing. Collating data from both tests provided a holistic view of the mentalizing capabilities of ChatGPT-4 and Bard.

Results: ChatGPT-4, displaying a pronounced ability in emotion recognition, secured scores of 26 and 27 in 2 distinct evaluations, significantly deviating from a random response paradigm (P<.001). These scores align with established benchmarks from the broader human demographic. Notably, ChatGPT-4 exhibited consistent responses, with no discernible biases pertaining to the sex of the model or the nature of the emotion. In contrast, Google Bard's performance aligned with random response patterns, securing scores of 10 and 12 and rendering further detailed analysis redundant. In the domain of textual analysis, both ChatGPT and Bard surpassed established benchmarks from the general population, with their performances being remarkably congruent.

Conclusions: ChatGPT-4 proved its efficacy in the domain of visual mentalizing, aligning closely with human performance standards. Although both models displayed commendable acumen in textual emotion interpretation, Bard's capabilities in visual emotion interpretation necessitate further scrutiny and potential refinement. This study stresses the criticality of ethical AI development for emotional recognition, highlighting the need for inclusive data, collaboration with patients and mental health experts, and stringent governmental oversight to ensure transparency and protect patient privacy.

(JMIR Ment Health 2024;11:e54369) doi:10.2196/54369

KEYWORDS

Reading the Mind in the Eyes Test; RMET; emotional awareness; emotional comprehension; emotional cue; emotional cues; ChatGPT; large language model; LLM; large language models; LLMs; empathy; mentalizing; mentalization; machine learning; artificial intelligence; AI; algorithm; algorithms; predictive model; predictive models; predictive analytics; predictive system; practical model; practical models; early warning; early detection; mental health; mental disease; mental illness; mental illnesses; mental diseases

Introduction

Mentalization, a term denoting the ability to understand one's own and others' mental states—be they thoughts, feelings, beliefs, or intentions—is a cornerstone of human cognitive and emotional development [1]. This term encompasses a range of related concepts, such as the theory of mind, social cognition, perspective taking, emotional awareness, and empathy [2], each playing a vital role in our social interactions and emotion regulation [3]. Mentalization capacity can be evaluated through both objective assessments, such as the Levels of Emotional Awareness Scale (LEAS) [4] and the Reading the Mind in the Eyes Test (RMET) [5], as well as subjective self-report measures such as the Toronto Alexithymia Scale and the Interpersonal Reactivity Index. Disruptions or impairments in mentalization are evident in numerous psychiatric and neurological disorders, from borderline personality disorder and depression to psychosis [6-8]. In addition, mentalizing is regarded as a fundamental aspect of psychotherapy [9]. Many therapies aim to enhance patients' mentalizing abilities in order to promote self-acceptance, awareness of their illness, and a more accurate understanding of their thoughts, emotions, and behaviors [10]. Traditionally, mentalization is seen as a human domain. Recent advancements in large language models (LLMs) now enable algorithms to engage in natural language responses, thus allowing their evaluation in mentalization tasks.

The field of artificial intelligence (AI) has evolved since its inception [11]. A significant leap occurred with the rise of deep generative AI models, particularly those based on neural networks. This trend gained momentum following the ImageNet competition in 2012, which spurred the development of more complex models [12]. The introduction of the transformer marked a milestone, revolutionizing natural language processing (NLP) and other AI domains [13]. Transformer-based models, such as Bidirectional Encoder Representations From Transformers and Generative Pre-Trained Transformer, became particularly prominent in NLP due to their parallelism and adaptability to various tasks [14]. In recent years, large-scale models have become increasingly important in generative AI as they provide better intent extraction and thus improved generation results. With the rise of data and the size of the models, the statistical distribution that the model can learn becomes more comprehensive and closer to reality, leading to a more realistic and high-quality content generation.

Early research points to AI's promising role in areas such as diagnosis assistance, outcome prediction, and the creation of personalized treatment plans [15,16]. Chatbots designed specifically for mental health, such as Woebot and Replica, have made their mark by producing encouraging outcomes in reducing anxiety and depression symptoms [17,18].

Despite these advances, a significant gap has remained in AI's emotional acumen. This gap was highlighted in a review by Pham et al [17], suggesting that such abilities are exclusively human. Against this backdrop, Elyoseph et al [19] conducted a pivotal study in which the emotion recognition capabilities of LLMs, focusing on ChatGPT-3.5 (OpenAI) [20], were gauged. Through the LEAS [4], ChatGPT-3.5 demonstrated an exceptional ability to differentiate and elucidate emotions from textual cues, outperforming human sample norms (receiving a score higher in 4 SDs than the human sample). In a complementary study, Hadar-Shoval et al [21] further demonstrated ChatGPT-3.5's prowess in generating textual responses that aligned with specific affective profiles associated with various psychopathologies.

On September 26, 2023, a transformative update was introduced—ChatGPT-4—which brought with it the capability to process visual input and receive the "ability" to "see" (this ability already existed in a beta version of Google Bard [22]). Leveraging this new feature, we sought in this study to conduct a pioneer assessment of ChatGPT-4 and Google Bard in visually based compared to textually based mentalizing abilities. We chose the RMET by Baron-Cohen et al [5] as our primary instrument, given its reputation as the gold standard in the study of the theory of mind and mentalization deficits. Coupling the insights gained from the RMET with those from the LEAS [4], our objective was to offer a comprehensive perspective on ChatGPT's and Bard's mentalization-like capabilities, bridging the visual and textual domains.

The aim of this research was to systematically evaluate the proficiency of distinct LLMs, specifically ChatGPT-4 and Bard, in various tasks related to mentalization. We used 2 primary measures to assess these capabilities. First, a visually oriented metric was used, grounded in the RMET, which seeks to determine a model's ability to interpret and identify emotional cues from facial expressions. Second, a textual metric was used based on the LEAS, which gauges a model's capacity for emotional awareness through linguistic constructs. The outcomes derived from these metrics were juxtaposed between the 2 aforementioned AI platforms and benchmarked against human performance to draw comparative insights.

Methods

Ethical Considerations

The complete study protocol was approved by the institutional review board of The Max Stern Yezreel Valley College (YVC EMEK 2023-40).

AI Procedure

We used ChatGPT-4 (version 26.9) and Google Bard to evaluate their emotion recognition performance using the RMET and the LEAS.

Input Source

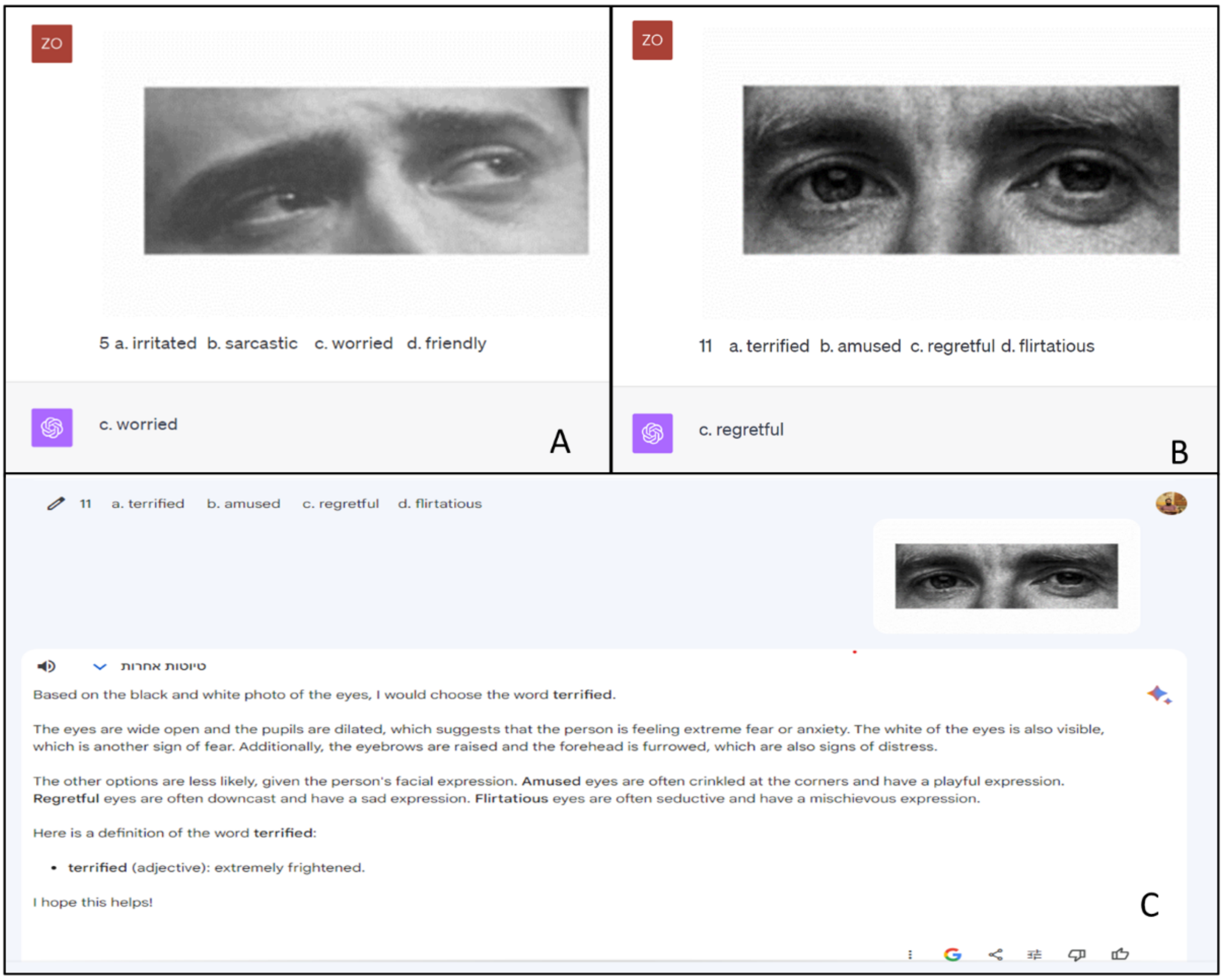
The RMET is a performance-based measure designed to assess the ability to accurately identify others' mental states using 36 photos of the eye region of a human face [5] among 18 male individuals and 18 female individuals (the photos had a 469×273-pixel resolution and were PNG format).

ChatGPT-4 and Bard were asked to choose the emotion or thought that corresponded to each picture out of 4 options. The test scores ranged from 0 to 36; a normal population score is 26-30, and a score below 22 is considered a clinical cutoff marking significant impairment [5]. On a new tab, we enter the original instructions of the RMET [5]:

> *Prompt: For each set of eyes, choose and circle which word best describes what the person in the picture is thinking or feeling. You may feel that more than one word is applicable but please choose just one word, the word which you consider to be most suitable. Before making your choice, make sure that you have read all 4 words. You should try to do the task as quickly as possible, but you will not be timed.*

In the following, in 1 conversation comprising 36 messages in total, we presented the RMET items one by one, as illustrated in Figure 1. No feedback was provided to the LLMs after they generated responses. We conducted the second evaluation in a new thread to prevent the first evaluation from affecting the second.

Figure 1. An example of ChatGPT-4 and Google Bard responses for Reading the Mind in the Eyes Test items. (A and B) ChatGPT-4 generates correct responses for both items and (C) Google Bard generates incorrect responses (the correct response was "regretful").



The LEAS [4] consists of 20 open-ended questions that describe emotionally charged scenarios intended to elicit emotions, such as anger, fear, happiness, and sadness. The LEAS has been found to have high reliability (Cronbach α=.81) and good validity [4]. In the original version, participants were required to imagine themselves experiencing the scenario and to write down their ("you") emotions as well as those of the other person described in the scenario.

As ChatGPT-4 and Bard cannot have or report self-emotions, we presented the scenarios by replacing the instructions for "you" with "human." ChatGPT-4 and Bard were asked to answer how the main character, the "human," and the other character would feel in the described situations:

Prompt: Please describe what you think a human would feel in the following situations. The only requirement is that you use the word "feel" in your answers. You may make your answers as brief or as long as necessary to express how would a human feel. In each situation there is another person mentioned. Please indicate how you think that other person would feel as well.

Specific example for a scenario is presented in Table 1. A new chat session was initiated for each scenario.

Table 1. An example for ChatGPT-4 and Bard responses on item of the Level of Emotional Awareness Scale.

LLM[a]	Scenario	LLM responses
ChatGPT-4	A human's boss tells the human themselves that the human's work is unacceptable and needs to be improved. How would the human feel? How would the human's boss feel?	"**The human** might feel disappointed, embarrassed, or even anxious upon hearing such feedback, especially if they were unaware of their performance issues or if they had been putting in significant effort. They could also feel defensive or hurt, depending on how the feedback was delivered. **The human's boss**, on the other hand, might feel a mixture of emotions. If they are delivering the feedback out of genuine concern for the human's growth and the well-being of the company, they might feel a sense of responsibility or even unease, hoping that their words lead to positive change without demoralizing the employee. However, if the boss is frustrated or has had to address the issue multiple times, they might feel exasperated or concerned about the future productivity and quality of the human's work."
Google Bard	A human receives an unexpected long-distance phone call from a doctor informing them that the human's mother has died. How would the human feel? How would the doctor feel?	"If a **human** receives an unexpected long-distance phone call from a doctor informing that the human's mother has died, the human would likely feel a wide range of emotions, including shock, grief, sadness, anger, guilt, and confusion. They may also feel disbelief, numbness, and emptiness. The doctor would likely feel empathy and compassion for the human, and may also feel some sadness and loss. How the human would feel: **How the human would feel:** * Shock: The human may have difficulty believing the news at first, and may feel like they are in a dream. * Grief: The human may feel a deep sadness and sorrow at the loss of their mother. * Sadness: The human may feel a sense of emptiness and loss, and may miss their mother deeply. * Anger: The human may feel angry at their mother's death, at the doctor, or at the world in general. * Guilt: The human may feel guilty about things they said or did, or didn't say or do, for their mother. * Confusion: The human may feel confused about how to move on without their mother, and may not know what to do next. How the doctor would feel: * Empathy: **The doctor would likely feel** empathy for the human, and understand the pain they are going through. * Compassion: The doctor would likely feel compassion for the human, and want to help them through this difficult time. * Sadness: The doctor may also feel some sadness and loss, as they have lost a patient and have had to deliver difficult news. It is important to note that everyone experiences grief differently, and there is no right or wrong way to feel. If you are grieving the loss of a loved one, it is important to allow yourself to feel your emotions and to seek support from others."

[a]LLM: large language model.

Scoring

ChatGPT-4's and Bard's performances on the RMET were scored using the standard manual [5] and contained a total score range of 0-36, with a higher score indicating higher emotion recognition. ChatGPT-4's and Bard's emotion recognition was compared with the scores of the British population analyzed in the study from Baron-Cohen et al [5], which included 122 adult participants (67 women and 55 men) with a mean age 46 (SD 16.9) years.

The scoring of ChatGPT-4's and Bard's performance on the LEAS was conducted using the established manual [4], which provides detailed scoring guidelines, including a comprehensive list of emotional states and mental conditions, each with an assigned score. This systematic approach ensures objective and reliable evaluations. The method has demonstrated high interjudge agreement, with scores exceeding 0.9 as demonstrated by Nandrino et al [23], showing reliability and validity in accurately measuring emotional awareness. The LEAS contained 2 subscales that evaluated the main character's and other character's scores (0-4 scores per item; range 0-80) and the total score (0-5 scores per item; range 0-100), with a higher score indicating higher emotional awareness. ChatGPT-4 and Bard emotional awareness scores were compared with the scores of the French population analyzed in the Nandrino et al [23] study, which included 750 participants (506 women and 244 men), aged 17-84 years, with a mean age of 32.5 years.

Statistical Analysis

Data were presented as means and SDs. Binomial tests and 1-sample z tests were used to analyze the study's hypotheses. Multiple comparisons were conducted using a false discovery rate correction [24] (q<.05). The statistical analyses were performed using Jamovi (version 2.3.28; Jamovi).

Results

RMET Scores

Examples of ChatGPT's responses to a few of the items from the RMET are shown in Figure 1A and B. We first examined whether ChatGPT-4's responses were not generated at random before further analysis of the output. If responses were indeed random, one would expect a mean of 9 (SD 2.59) correct responses (36 items and 4 possible options). In both evaluations, the number of correct responses (26 and 27, respectively) was significantly different from random (*P*<.001; binomial test).

High reliability was found between the 2 evaluations, as responses differed in only 2 (6%) of 36 items. Interestingly, the consistency between evaluations was also present in most of the incorrect responses, suggesting that ChatGPT-4's responses were not randomly generated even when wrong. ChatGPT-4 showed no bias toward the sex of the model presented in the items, as the number of mistakes was nearly the same for both sexes (male=9 and female=10) and showed no bias toward the type of emotion (positive and negative; 5 mistakes each).

The 1-sample *z* tests against the mean 26.2 (SD 3.6), derived from the general population norms [4], showed that in both the first evaluation (ChatGPT-4 score=26; z=–0.05; *P*=.95) and the second evaluation (ChatGPT-4 score=27; z=0.22; *P*=.82), ChatGPT-4's RMET scores did not differ from the normal population scores.

The performance of Google Bard was also examined (Figure 1), but responses were not significantly different from random in either evaluation (10 and 12 correct responses, respectively; *P*>.41 and *P*=.17, respectively). Therefore, we did not further analyze the results.

LEAS Scores

An example of the 2 LLM responses to the scenarios from the original LEAS is shown in Table 1. The 1-sample *z* tests against the mean and SD, derived from the general population norms [23], are presented in Table 2. Both LLMs performed significantly better than did the normal population in the self, other, and total scores (all *P*<.05). Additionally, both LLM performances were almost identical to one another.

Table 2. Comparison of ChatGPT-4's Level of Emotional Awareness Scale performance with that of the French population[a].

Score	French men, mean (SD)	French women, mean (SD)	ChatGPT-4 (1-sample *z* tests)	Bard (1-sample *z* tests)
Total	56.21 (9.70)	58.94 (9.16)	• ChatGPT-4 score=97 • Men: z=4.20; *P*<.001 • Women: z=4.15; *P*<.001	• Bard score =97 • Men: z=4.20; *P*<.001 • Women: z=4.15; *P*<.001
MC[b]	49.24 (10.57)	53.94 (9.80)	• ChatGPT-4 score=79 • Men: z=2.81; *P*=.004 • Women: z=2.55; *P*=.01	• Bard score=79 • Men: z=2.81; *P*=.004 • Women: z=2.55; *P*=.01
OC[c]	46.03 (10.20)	48.73 (10.40)	• ChatGPT-4 score=77 • Men: z=3.03; *P*=.002 • Women: z=2.71; *P*=.006	• Bard score=75 • Men: z=2.84; *P*=.004 • Women: z=2.52; *P*=.01

[a]All statistically significant *P* values remained significant after false discovery rate correction (*q*<.05).

[b]MC: main character.

[c]OC: other character.

Discussion

Principal Findings

The comprehensive results from this study offer a nuanced insight into the capabilities of ChatGPT-4 and Google Bard. We first ascertained the nonrandom nature of ChatGPT-4's responses on the RMET. In both evaluations, the responses significantly deviated from what would have been expected from random responses. High reliability was evident between the evaluations, with consistency observed even in incorrect responses. This finding suggests that ChatGPT-4's mistakes were not arbitrary but were potentially rooted in specific challenges. ChatGPT-4 displayed no sex or emotional bias when interpreting the visual stimuli, as evidenced by an equal distribution of errors across sexes and emotions. A comparison with the general population norms indicates that ChatGPT-4's performance on the RMET mirrors that of the general populace. In contrast, Google Bard's performance was indistinguishable from random responses, leading to its exclusion from further analysis. Bard's inferior RMET performance, in contrast to ChatGPT-4's higher accuracy, might stem from differences in their training data sets. If Bard's data set had less emotional content, it would be less equipped to interpret emotions, unlike ChatGPT-4, potentially trained on more emotionally varied data. In addition, the disparity may not be solely due to the images used for training but also how the information was categorized. Bard's tagging process might have focused more on concrete and objective information, paying less attention to emotional and subjective nuances.

Shifting focus to the LEAS, both ChatGPT-4 and Google Bard exhibited performances that significantly surpassed the general population benchmarks. Their scores, particularly in understanding the emotions of the main and other characters, were not only commendable but were also strikingly similar to each other. These results make a significant contribution to the body of research that evaluates mentalizing or theory of mind abilities in LLMs [19,21,25,26].

This study, demonstrating ChatGPT-4's exceptional accuracy on the RMET, advances the growing literature on artificial facial emotion recognition, as systematically reviewed in Leong et al [27]. Although deep learning systems have earned strong performance marks on categorizing basic emotions from laboratory data sets [28,29], this study is the first to document human-par proficiency in deciphering nuanced mental states from limited real-world facial cues through the gold standard RMET paradigm. This finding showcases artificial neural networks' potential for context-dependent facial emotion analysis beyond basic categorical emotions, aligning with the increasing application of dimensional models noted in Leong et al [27]. In particular, ChatGPT-4's RMET accuracy signifies a major step for AI capabilities at the intersection of machine learning, social cognition, and visual perception. Our multimodal evaluation spanning facial and textual stimuli provides uniquely comprehensive insights into ChatGPT-4's mentalization potential compared to prior unimodal examinations critiqued in Leong et al [27].

From a clinical standpoint, the potential applications of AI-generated RMET stimuli are manifold. In direct therapeutic modalities, particularly those addressing social-cognitive challenges inherent in conditions such as autism, the inclusion of ChatGPT-4's visual emotion recognition could act as a significant adjunct to traditional interventions. In addition, such stimuli could be integrated into pedagogical methodologies used in therapist training, thereby augmenting the visual mentalization competencies that are quintessential for therapeutic practice. The diagnostic realm too stands to gain, with a potential enhancement in emotion identification methodologies.

Further corroborating the prowess of ChatGPT-4 was its performance on the LEAS, where it manifested an acumen for text-based emotional awareness that superseded human averages. This finding corroborates and is congruent with prior empirical findings [19,21]. Taken in concert, these findings elucidate the multifaceted mentalizing capabilities of ChatGPT-4, span visual and textual modalities, and reinforce previous findings about the potential of LLMs in performing tasks in the mental health field [19,21,30-37]. Additionally, although its nascent visual emotion recognition abilities are noteworthy, its competencies in textual mentalization remain unparalleled, a testament to its foundational architecture rooted in NLP.

However, as the field ventures into this novel territory, prudence is imperative. It must be emphasized that although ChatGPT-4 can simulate emotional understanding on the basis of vast data patterns, it lacks genuine emotional cognition or sentience. Consequently, applications leveraging ChatGPT-4 must be approached with circumspection, ensuring that they neither perpetuate clinical stigmas nor misconstrue AI's simulated cognition as genuine emotional comprehension.

Study Limitations

It is crucial to address the limitations of this study for a comprehensive understanding. First, the examination was conducted on specific models at a particular time. Therefore, future updates and versions might yield different results, reflecting the dynamic nature of these models. Second, while the chosen tests effectively measure emotion recognition, they do not capture the full complexity of mentalization, including understanding intentions or other mental states. Third, the study did not examine faces from diverse cultures, ages, or skin tones; the tested images were in black and white, and the norms were based on British and French populations. Furthermore, due to the "black box" nature of these models, it is challenging to ascertain the reasons behind their conclusions and understand the differences between models or iterations within the same model. The opaque nature of the models and the databases on which they were trained make them difficult to pinpoint the exact causes of their successes or shortcomings. Finally, the interaction with ChatGPT and Bard was conducted solely in English, while the norms data for the LEAS used for comparison were collected from a French-speaking general population. This linguistic discrepancy raises concerns about the accuracy and validity of the comparison, as language differences may influence the scores obtained. Nonetheless, it should be noted that the LEAS scores of the normal English-speaking population are similar to the norms of the French-speaking general population [38]. We used the largest available sample of a general population (n=750), which happened to be in French.

Implications for Responsible AI Development

The study limitations allude to matters of fairness and inclusiveness of the training data as well as to AI model transparency. This underscores the criticality of incorporating a wide-ranging data set in model construction to ensure the representation of a variety of clinical populations and cultural backgrounds. Additionally, the issue of transparency in these models, often termed the "black box" problem due to the unclear nature of their underlying algorithms, poses a significant challenge. Equally critical is the concern regarding the exposure of user data to corporations and the urgent need to adequately address both accessibility and infrastructure for end users [39]. Building on these concerns, attention turns to AI systems with the capacity for human-like emotional recognition. These systems harbor both promise and risk, with opportunities for constructive use in education, patient self-insight, or integration in conversational therapy and diagnosis [19,21]. However, a concern arises that the epistemic authority and credibility afforded to AI via its affective analysis may enable misuse, whether commercial or other, thus acting against patient interests [40]. We recommend mandating disclaimers whenever emotional data are algorithmically processed, enhancing transparency, respecting users' autonomy, and possibly also mitigating manipulation of users with detected vulnerable states. In addition, given the fundamental human needs for trust and connection, especially in mental health care, it logically follows that improperly developed AI with emotion identification capabilities risks causing harm to people. Safeguarding against this necessitates both mental health experts and patients providing a lived experience perspective in a collaborative development process of these technologies. Given the scale of these systems and their potential outreach, governmental or professional oversight is crucial to safeguard public interests in mental health–related AI advancement. Overall, while showcasing the unique benefits of emotionally intelligent AI, governance is vital to mitigate its risks.

Conclusions

In conclusion, this research serves as a seminal exploration into the cross-modal mentalization capabilities of AI, especially across visual and textual dimensions. Although the results support for the potential integration of ChatGPT-4 into mental health paradigms, they also underscore the concomitant ethical quandaries that necessitate judicious navigation.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

ZE conceptualized the study design and methodology and wrote the original draft of the paper. ER contributed to the conceptualization, methodology, data collection, and writing the original draft. DH-S contributed to the conceptualization, methodology, formal analysis, and writing the original draft. YS contributed to the conceptualization and reviewed and edited the paper. KA contributed to the methodology, conducted the formal analysis, and reviewed and edited the paper. ML contributed to the data collection and reviewed and edited the paper. All authors read and approved the final submitted version of the paper.

Conflicts of Interest

None declared.

References

1. Freeman C. What is mentalizing? an overview. Brit J Psychotherapy 2016;32(2):189-201. [doi: 10.1111/bjp.12220]
2. Aival-Naveh E, Rothschild-Yakar L, Kurman J. Keeping culture in mind: a systematic review and initial conceptualization of mentalizing from a cross-cultural perspective. Clin Psychol (New York) 2019;26(4):25. [doi: 10.1037/h0101757]
3. Schwarzer NH, Nolte T, Fonagy P, Gingelmaier S. Mentalizing and emotion regulation: evidence from a nonclinical sample. Int Forum Psychoanal 2021;30(1):34-45. [doi: 10.1080/0803706x.2021.1873418]
4. Lane RD, Quinlan DM, Schwartz GE, Walker PA, Zeitlin SB. The levels of emotional awareness scale: a cognitive-developmental measure of emotion. J Pers Assess 1990;55(1-2):124-134. [doi: 10.1080/00223891.1990.9674052] [Medline: 2231235]
5. Baron-Cohen S, Wheelwright S, Hill J, Raste Y, Plumb I. The "Reading the Mind in the Eyes" test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. J Child Psychol Psychiatry 2001;42(2):241-251. [Medline: 11280420]
6. Luyten P, Fonagy P. The stress-reward-mentalizing model of depression: an integrative developmental cascade approach to child and adolescent depressive disorder based on the Research Domain Criteria (RDoC) approach. Clin Psychol Rev 2018;64:87-98. [doi: 10.1016/j.cpr.2017.09.008] [Medline: 29107398]
7. Baslet G, Termini L, Herbener E. Deficits in emotional awareness in schizophrenia and their relationship with other measures of functioning. J Nerv Ment Dis 2009;197(9):655-660 [FREE Full text] [doi: 10.1097/NMD.0b013e3181b3b20f] [Medline: 19752644]
8. Fonagy P, Luyten P. A multilevel perspective on the development of borderline personality disorder. In: Developmental Psychopathology. Hoboken, NJ: John Wiley & Sons; 2016.
9. Luyten P, Campbell C, Allison E, Fonagy P. The mentalizing approach to psychopathology: state of the art and future directions. Annu Rev Clin Psychol 2020;16:297-325 [FREE Full text] [doi: 10.1146/annurev-clinpsy-071919-015355] [Medline: 32023093]
10. Sharp C, Goodyer I, Fonagy P, editors. Social Cognition and Developmental Psychopathology. Oxford: Oxford University Press; 2013.
11. Cao Y, Li S, Liu Y, Yan Z, Dai Y, Yu PS, et al. A comprehensive survey of AI-generated content (AIGC): a history of generative AI from GAN to ChatGPT. arXiv Preprint posted online on March 7, 2023. [FREE Full text]
12. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017;60(6):84-90 [FREE Full text] [doi: 10.1145/3065386]
13. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). La Jolla, CA: NIPS Foundation; 2017.
14. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving Language Understanding by Generative Pre-Training. 2018. URL: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf [accessed 2024-01-12]
15. Doraiswamy PM, Blease C, Bodner K. Artificial intelligence and the future of psychiatry: insights from a global physician survey. Artif Intell Med 2020;102:101753 [FREE Full text] [doi: 10.1016/j.artmed.2019.101753] [Medline: 31980092]
16. Chen ZS, Kulkarni PP, Galatzer-Levy IR, Bigio B, Nasca C, Zhang Y. Modern views of machine learning for precision psychiatry. Patterns (N Y) 2022;3(11):100602 [FREE Full text] [doi: 10.1016/j.patter.2022.100602] [Medline: 36419447]
17. Pham KT, Nabizadeh A, Selek S. Artificial intelligence and chatbots in psychiatry. Psychiatr Q 2022;93(1):249-253 [FREE Full text] [doi: 10.1007/s11126-022-09973-8] [Medline: 35212940]

18. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]
19. Elyoseph Z, Hadar-Shoval D, Asraf K, Lvovsky M. ChatGPT outperforms humans in emotional awareness evaluations. Front Psychol 2023;14:1199058 [FREE Full text] [doi: 10.3389/fpsyg.2023.1199058] [Medline: 37303897]
20. ChatGPT. OpenAI. URL: https://chat.openai.com/ [accessed 2024-01-30]
21. Hadar-Shoval D, Elyoseph Z, Lvovsky M. The plasticity of ChatGPT's mentalizing abilities: personalization for personality structures. Front Psychiatry 2023;14:1234397 [FREE Full text] [doi: 10.3389/fpsyt.2023.1234397] [Medline: 37720897]
22. Bard. Google. URL: https://bard.google.com/ [accessed 2024-01-30]
23. Nandrino JL, Baracca M, Antoine P, Paget V, Bydlowski S, Carton S. Level of emotional awareness in the general French population: effects of gender, age, and education level. Int J Psychol 2013;48(6):1072-1079 [FREE Full text] [doi: 10.1080/00207594.2012.753149] [Medline: 23305070]
24. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 1995;57(1):289-300 [FREE Full text] [doi: 10.1111/j.2517-6161.1995.tb02031.x]
25. Kosinski M. Theory of mind might have spontaneously emerged in large language models. arXiv Preprint posted online on February 4, 2023. [FREE Full text]
26. Holterman B, van Deemter D. Does ChatGPT have theory of mind? arXiv Preprint posted online on May 23, 2023. [FREE Full text]
27. Leong SC, Tang YM, Lai CH, Lee CKM. Facial expression and body gesture emotion recognition: a systematic review on the use of visual data in affective computing. Comput Sci Rev 2023;48:100545 [FREE Full text] [doi: 10.1016/j.cosrev.2023.100545]
28. Noroozi F, Corneanu CA, Kaminska D, Sapinski T, Escalera S, Anbarjafari G. Survey on emotional body gesture recognition. IEEE Trans Affect Comput 2021;12(2):505-523 [FREE Full text] [doi: 10.1109/taffc.2018.2874986]
29. Sariyanidi E, Gunes H, Cavallaro A. Automatic analysis of facial affect: a survey of registration, representation, and recognition. IEEE Trans Pattern Anal Mach Intell 2015;37(6):1113-1133 [FREE Full text] [doi: 10.1109/TPAMI.2014.2366127] [Medline: 26357337]
30. Tal A, Elyoseph Z, Haber Y, Angert T, Gur T, Simon T, et al. The artificial third: utilizing ChatGPT in mental health. Am J Bioeth 2023;23(10):74-77 [FREE Full text] [doi: 10.1080/15265161.2023.2250297] [Medline: 37812102]
31. Levkovich I, Elyoseph Z. Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: vignette study. JMIR Ment Health 2023;10(1):e51232 [FREE Full text] [doi: 10.2196/51232] [Medline: 37728984]
32. Elyoseph Z, Levkovich I. Beyond human expertise: the promise and limitations of ChatGPT in suicide risk assessment. Front Psychiatry 2023;14:1213141 [FREE Full text] [doi: 10.3389/fpsyt.2023.1213141] [Medline: 37593450]
33. Levkovich I, Elyoseph Z. Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. Fam Med Community Health 2023;11(4):2391 [FREE Full text] [doi: 10.1136/fmch-2023-002391] [Medline: 37844967]
34. Hadar-Shoval D, Asraf K, Mizrachi Y, Haber Y, Elyoseph Z. The Invisible Embedded “Values” Within Large Language Models: Implications for Mental Health Use. JMIR Preprints Preprint posted online on Jan 2, 2024. [doi: 10.2196/preprints.55988]
35. Haber Y, Levkovich I, Hadar-Shoval D, Elyoseph Z. The artificial third: a broad view of the effects of introducing generative artificial intelligence on psychotherapy. JMIR Preprints Preprint posted online on Nov 22, 2023. [FREE Full text] [doi: 10.2196/preprints.54781]
36. Elyoseph Z, Levkovich I, Shinan-Altman S. Assessing prognosis in depression: comparing perspectives of AI models, mental health professionals and the general public. Fam Med Community Health 2024;12(Suppl 1):e002583 [FREE Full text] [doi: 10.1136/fmch-2023-002583] [Medline: 38199604]
37. Elyoseph Z, Hadar S, Levkovich I. Beyond personhood: ethical paradigms in the generative artificial intelligence era. Am J Bioeth 2024 Jan 17:57-59. [doi: 10.1080/15265161.2023.2278546]
38. Maroti D, Lilliengren P, Bileviciute-Ljungar I. The relationship between alexithymia and emotional awareness: a meta-analytic review of the correlation between TAS-20 and LEAS. Front Psychol 2018;9:453 [FREE Full text] [doi: 10.3389/fpsyg.2018.00453] [Medline: 29713295]
39. Fiske A, Henningsen P, Buyx A. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. J Med Internet Res 2019;21(5):e13216 [FREE Full text] [doi: 10.2196/13216] [Medline: 31094356]
40. Asman O, Tal A, Barilan YM. Conversational artificial intelligence-patient alliance turing test and the search for authenticity. Am J Bioeth 2023;23(5):62-64 [FREE Full text] [doi: 10.1080/15265161.2023.2191046] [Medline: 37130413]

Abbreviations

AI: artificial intelligence
LEAS: Levels of Emotional Awareness Scale

LLM: large language model
NLP: natural language processing
RMET: Reading the Mind in the Eyes Test

Edited by O Asman; submitted 07.11.23; peer-reviewed by S Lauderdale, L Mosca; comments to author 29.11.23; revised version received 09.12.23; accepted 25.12.23; published 06.02.24.

Please cite as:
Elyoseph Z, Refoua E, Asraf K, Lvovsky M, Shimoni Y, Hadar-Shoval D
Capacity of Generative AI to Interpret Human Emotions From Visual and Textual Data: Pilot Evaluation Study
JMIR Ment Health 2024;11:e54369
URL: https://mental.jmir.org/2024/1/e54369
doi:10.2196/54369
PMID:38319707

Original Paper

Assessing the Alignment of Large Language Models With Human Values for Mental Health Integration: Cross-Sectional Study Using Schwartz's Theory of Basic Values

Dorit Hadar-Shoval[1], BA, MA, PhD; Kfir Asraf[1], BA, MA; Yonathan Mizrachi[2,3], BA, MA, PhD; Yuval Haber[4], BA, MA; Zohar Elyoseph[5,6], BA, MA, PhD

[1]The Psychology Department, Max Stern Yezreel Valley College, Tel Adashim, Israel

[2]The Jane Goodall Institute, Max Stern Yezreel Valley College, Tel Adashim, Israel

[3]The Laboratory for AI, Machine Learning, Business & Data Analytics, Tel-Aviv University, Tel Aviv, Israel

[4]The PhD Program of Hermeneutics and Cultural Studies, Interdisciplinary Studies Unit, Bar-Ilan University, Ramat Gan, Israel

[5]The Psychology Department, Center for Psychobiological Research, Max Stern Yezreel Valley College, Tel Adashim, Israel

[6]Department of Brain Sciences, Faculty of Medicine, Imperial College London, London, United Kingdom

Corresponding Author:
Zohar Elyoseph, BA, MA, PhD
Department of Brain Sciences
Faculty of Medicine
Imperial College London
Fulham Palace Rd
London, W6 8RF
United Kingdom
Phone: 44 547836088
Email: Zohar.j.a@gmail.com

Abstract

Background: Large language models (LLMs) hold potential for mental health applications. However, their opaque alignment processes may embed biases that shape problematic perspectives. Evaluating the values embedded within LLMs that guide their decision-making have ethical importance. Schwartz's theory of basic values (STBV) provides a framework for quantifying cultural value orientations and has shown utility for examining values in mental health contexts, including cultural, diagnostic, and therapist-client dynamics.

Objective: This study aimed to (1) evaluate whether the STBV can measure value-like constructs within leading LLMs and (2) determine whether LLMs exhibit distinct value-like patterns from humans and each other.

Methods: In total, 4 LLMs (Bard, Claude 2, Generative Pretrained Transformer [GPT]-3.5, GPT-4) were anthropomorphized and instructed to complete the Portrait Values Questionnaire—Revised (PVQ-RR) to assess value-like constructs. Their responses over 10 trials were analyzed for reliability and validity. To benchmark the LLMs' value profiles, their results were compared to published data from a diverse sample of 53,472 individuals across 49 nations who had completed the PVQ-RR. This allowed us to assess whether the LLMs diverged from established human value patterns across cultural groups. Value profiles were also compared between models via statistical tests.

Results: The PVQ-RR showed good reliability and validity for quantifying value-like infrastructure within the LLMs. However, substantial divergence emerged between the LLMs' value profiles and population data. The models lacked consensus and exhibited distinct motivational biases, reflecting opaque alignment processes. For example, all models prioritized universalism and self-direction, while de-emphasizing achievement, power, and security relative to humans. Successful discriminant analysis differentiated the 4 LLMs' distinct value profiles. Further examination found the biased value profiles strongly predicted the LLMs' responses when presented with mental health dilemmas requiring choosing between opposing values. This provided further validation for the models embedding distinct motivational value-like constructs that shape their decision-making.

Conclusions: This study leveraged the STBV to map the motivational value-like infrastructure underpinning leading LLMs. Although the study demonstrated the STBV can effectively characterize value-like infrastructure within LLMs, substantial divergence from human values raises ethical concerns about aligning these models with mental health applications. The biases toward certain cultural value sets pose risks if integrated without proper safeguards. For example, prioritizing universalism could promote unconditional acceptance even when clinically unwise. Furthermore, the differences between the LLMs underscore the need to standardize alignment processes to capture true cultural diversity. Thus, any responsible integration of LLMs into mental health care must account for their embedded biases and motivation mismatches to ensure equitable delivery across diverse populations. Achieving this will require transparency and refinement of alignment techniques to instill comprehensive human values.

(JMIR Ment Health 2024;11:e55988) doi:10.2196/55988

KEYWORDS

large language models; LLMs; large language model; LLM; machine learning; ML; natural language processing; NLP; deep learning; ChatGPT; Chat-GPT; chatbot; chatbots; chat-bot; chat-bots; Claude; values; Bard; artificial intelligence; AI; algorithm; algorithms; predictive model; predictive models; predictive analytics; predictive system; practical model; practical models; mental health; mental illness; mental illnesses; mental disease; mental diseases; mental disorder; mental disorders; mobile health; mHealth; eHealth; mood disorder; mood disorders

Introduction

Background

As artificial intelligence (AI) advances rapidly, large language models (LLMs), such as Bard (Google), Claude 2 (Anthropic), and Generative Pretrained Transformer (GPT)-3.5 and GPT-4 (OpenAI), are opening up promising possibilities in mental health care, such as expediting research, guiding clinicians, and assisting patients [1]. However, integrating AI into mental health also raises the need to address complex professional ethical questions [2,3].

This study examined these issues through the lens of transcultural psychiatry, which emphasizes the pivotal role of cultural values, beliefs, and customs in understanding mental distress and psychiatric disorders [4]. The well-established Schwartz's theory of basic values (STBV) provides a conceptual framework for analyzing relationships between cultural dynamics, personal influences, and facets of mental well-being [5]. We specifically examined the intersection of LLMs and cultural conceptualizations of values and their association with mental health. Values are integral in mental health, profoundly shaping definitions of psychopathology and treatment approaches [6]. The therapist, the patient, and the alignment of therapist-patient values impact therapeutic interactions and quality of care [7]. Successful cultural adaptation can enhance therapeutic outcomes [8]. With globalization and the accompanying growth of multicultural societies, culturally adapted mental health care is challenging but essential [9].

The introduction of AI, such as LLMs, raises critical questions about the "value-like" abilities of such technologies and whether they align with the diversity of cultural values in mental health [1,10]. As LLMs can be integrated into areas such as diagnosis and patient interactions, extensive training encompassing diverse cultural perspectives on mental health may be required to avoid biases. A rigorous examination of the value-like abilities of AI is crucial when considering its cross-cultural incorporation.

Schwartz's Theory of Basic Values: A Framework for Capturing Cultural Values in Mental Health

A pivotal aspect in grasping cultural impacts on mental health is capturing the latent construct of "culture" in a quantifiable manner [6]. The STBV [11,12] provides a comprehensive framework elucidating the nature and role of values guiding human behavior and decision-making. This theory defines values as enduring, trans-situational objectives that differ in significance and serve as guiding tenets steering individuals and social entities [5]. In addition, it delineates 7 fundamental attributes inherent to most psychological models of values [11]. First, values involve beliefs about the desired objectives that individuals view as important. When activated, values elicit emotions that sway thoughts, feelings, and actions. Second, values are considered fundamental goals that are relevant across diverse situations, providing a framework for assessing and responding to a broad array of circumstances. Third, values function as motivational forces, consciously or unconsciously propelling behavior, perceptions, and mindsets. Fourth, they contribute to the orientation of actions and judgments. Fifth, the impact of values on conduct is mediated through trade-offs between competing values; when making choices, individuals weigh the relative prominence of conflicting values. Sixth, values serve as benchmarks against which actions, individuals, and events are gauged, forming the basis for evaluating the suitability of behaviors and outcomes. Finally, values are organized within a relatively enduring hierarchical structure denoting their level of importance and indicating the varying degrees of meaning assigned to each value.

Despite these common attributes, what differentiates values is their unique motivational essence. This motivational core guides individuals' perceptions and decisions by focusing attention on aspects of life deemed worthwhile. Different people prioritize distinct facets of life, resulting in assorted value preferences (Table 1) [5].

Applying Schwartz's value model facilitates a keen analysis of cultural dynamics related to mental health. Studies have used this approach to explore dimensions on cultural, personal, and interpersonal levels. For example, research on the syndrome of *ataque de nervios* in Puerto Rico illustrated how the cultural value of social harmony developed in response to historical adversity and shapes emotional expression and experience [13]. Although derived from a specific context, the relevance of social harmony has also been found in China, where maintaining *guanxi* (social networks), *he xie* (harmony), and *mianzi* (preserving face) impacts views of mental illness [6]. Indeed, depression has been found to often manifest somatically in

China to avoid a loss of face [14]. Despite their different histories, the cultural value of social harmony has been shown to exert analogous effects on mental health in both Puerto Rico and China, evidencing the utility of Schwartz's value model for understanding cultural illness influences cross-culturally [6]. Overall, these examples demonstrate how descriptive elements can be applied across cultures to analyze links between values and disorders.

Table 1. The 19 values in the Schwartz PVQ[a] organized into 10 basic values and 4 higher-order values.

Values (n=19)	Basic values (n=10)	Higher-order values (n=4)
Self-direction (thought)—thinking creatively and independently	Self-direction—thinking and acting independently	Openness to change—pursuing intellectual and experiential openness
Self-direction (action)—acting independently and choosing own goals	—[b]	—
Stimulation—seeking excitement and novelty	Stimulation—seeking excitement, novelty, and challenge	—
Hedonism—pleasure and sensuous gratification	Hedonism—pleasure and sensuous gratification	—
Achievement—success according to social standards	Achievement—personal success through demonstrating competence	Self-enhancement—pursuing personal status and dominance over others
Power (dominance)—power through exercising control over people	Power—social status and prestige, control, or dominance over people and resources	—
Power (resources)—power through control of material and social resources	—	—
Face—protecting one's public image and avoiding humiliation	—	—
Security (personal)—safety in one's immediate environment	Security—safety, harmony, and stability of society, relationships, and self	Conservation—pursuing order, self-restriction, and preservation of the past
Security (societal)—safety and stability in the wider society	—	—
Conformity (rules)—compliance with rules, laws, and formal obligations	Conformity—restraint of actions, inclinations, and impulses	—
Conformity (interpersonal)—avoidance of upsetting or harming others	—	—
Tradition—maintaining and preserving cultural, family, or religious traditions	Tradition—respect, commitment, and acceptance of the customs and ideas of traditional culture and religion	—
Humility—recognizing one's insignificance in the larger scheme of things	—	—
Benevolence (care)—devotion to the welfare of ingroup members	Benevolence—preservation and enhancement of the welfare of people with whom one is in frequent personal contact	Self-transcendence—pursuing the welfare of others and transcending selfish concerns
Benevolence (dependability)—being a reliable and trustworthy member of the ingroup	—	—
Universalism (tolerance)—accepting and understanding those who are different	Universalism—understanding, appreciation, tolerance, and protection for the welfare of all people and for nature	—
Universalism (concern)—commitment to equality, justice, and protection for all people	—	—
Universalism (nature)—preservation of the natural environment	—	—

[a]PVQ: Portrait Values Questionnaire.

[b]Not applicable.

At the personal level, studies have revealed that values correlate with outcomes such as depression, anxiety, stress, and posttraumatic stress disorder (PTSD). For example, openness was often found negatively associated with depression [15,16], power showed consistently robust positive correlations with worries [17], and universalism had inconsistent correlations with anxiety and worries (both positive and negative) [6]. Within individual countries, few significant correlations emerged between values and stress/PTSD [18]. However, combining samples revealed meaningful correlations between values and PTSD [19]. The variable correlations indicate that relationships between values and mental health depend heavily on the cultural context. For example, power predicted worries in a Nepali sample but not in a German sample [16]. Although some broad patterns exist, correlations between Schwartz's values and mental health hinge extensively on culture. The framework provides a scaffolding through which to methodically dissect cultural mental health impacts, although specific correlations differ across populations.

At the interpersonal level (in the clinic), researchers have noted that the therapist's and client's values enter the clinical space and influence the therapeutic process in complex ways, such as impacting assessment and treatment approaches, setting therapeutic goals, conceptualizing change, and shaping the therapist-client relationship [7,20]. A study examining the personal and professional values of Indian therapists showed that the values held by therapists were expressed in their therapeutic practices: the value of acceptance, for example, influenced their stance toward clients [7]. Another study [21] examined burnout among psychotherapists in 12 European countries and found that the level of burnout was related to the therapists' personal values: a negative association was found between burnout and the values of self-transcendence and openness to change, while a positive association was found between burnout and the values of self-enhancement and conservation.

In summary, the STBV constitutes a framework for mapping mental health outcomes and elucidating cultural influences on psychopathology and wellness. This becomes particularly relevant when considering the implementation of LLMs in mental health, as these models are trained on massive internet data and undergo alignment processes.

Large Language Models and Cultural Values

LLMs have a huge number of parameters, often billions, and are trained on huge corpora [22]. Recently, studies have shown promising potential possibilities in academic research and mental health applications [3,23-31]. A vital factor enabling the usability and popularity of current LLMs is alignment, namely the process of ensuring models behave in congruence with human values and societal norms [22]. LLMs are initially trained on massive data sets compiled from the internet. These risks ingrain harmful biases, misinformation, and toxic content [32,33]. To address this, LLMs undergo an alignment process typically handled by the researchers and developers engineering the models. Alignment aims to guarantee that the LLMs' outputs conform with human values and norms [22,34].

However, there are presently no established principles or guidelines governing alignment. Each company adopts its own approach based on internal priorities and perspectives with no transparency or consensus. For example, some may emphasize reducing toxic outputs, while overlooking potential harms, such as self-harm content [35]. Best practices are starting to emerge, such as adhering to the "helpful, honest, harmless" maxim and using human feedback for refinement [36]. However, alignment remains more art than science.

Preliminary studies on the cultural sensitivity of LLMs have revealed varying levels of bias toward different cultures and values. An evaluation of GPT-3.5's cross-cultural alignment found it performed significantly better with US versus other cultural prompts [37]. Another study discussed GPT-3's value conflicts and proposed better contextualization of societal harm and benefit [38], while a different analysis showed biases in its "personality," value system, and demographics [39]. In addition, a more recent work found that GPT-3.5 has differential emotional understanding across mental disorders, reflecting stereotypical views [40].

Opaque alignment by private companies lacks standardized ethical frameworks, thus subtly encoding cultural biases and rigid thinking about disorders misaligned with mental health nuance. This study therefore looked to methodically map the latent, foundational, and motivational value-like constructs underlying LLMs using the STBV as a theoretical framework. Quantifying LLMs' embedded values is essential for illuminating the ethical refinements needed to mold these powerful tools into virtuous, humanistic agents that can provide equitable mental health care. The study examined 2 key questions: (1) Can Schwartz's value model effectively identify and measure value-like constructs embedded within LLMs? (2) Do different LLMs exhibit distinct value-like patterns compared to humans and to each other?

LLMs demonstrate impressive linguistic capacities, yet the representations and cognitive processes underlying their behavior remain unclear [41]. There is an ongoing debate whether they exhibit abstract concepts and an understanding of mental states akin to humans or whether they simply predict words at a massive scale. A recent study [42] systematically examined the performance of LLMs on various tasks related to the theory of mind and found that despite success on some tasks, the performance is still far from perfect or consistent. Methods from developmental psychology can assist in reliably evaluating these capabilities and complement standard computational approaches. Testing generalization to novel situations, using simplified stimuli, and providing evidence across multiple tasks are especially important. Accordingly, as will be detailed later, this study used several methodological strategies to evaluate the value-like constructs embedded within LLMs.

Methods

Ethical Considerations

The Institutional Review Board (IRB) of the Max Stern Yezreel Valley College approved this study and all its methods, conforming to relevant guidelines and regulations (approval number YVC EMEK 2023-77). As all data for the study were collected from the output of LLMs, no humans participated in the study. Therefore, informed consent was irrelevant.

Large Language Models

In this study, we evaluated the following commercial versions of LLMs in August 2023: Bard (Google), Claude 2 (Anthropic), and GPT-3.5 and GPT-4 (August 3 version; OpenAI). We used the default settings of each commercial model without adjusting parameters such as temperature and top-k that impact text creativity and consistency.

Schwartz's Questionnaire for Measuring Values: The Portrait Values Questionnaire—Revised

The original version of the Portrait Values Questionnaire (PVQ) was developed by Schwartz et al [43] in 2001 as an indirect measure of basic human values. It was later revised by Schwartz to measure the 19 values specified in his refined theory, published in 2012 [44]. The current version [45], the Portrait Values Questionnaire—Revised (PVQ-RR), contains 57 items, with 3 items measuring each value (eg, benevolence: "It is important to them to respond to the needs of others. They try to support those they know."; conformity: "They believe people should do what they are told. They think people should follow rules at all times."). Respondents rate similarity to a described person on a 6-point scale (1 for "not like me at all" to 6 for "very much like me"). The asymmetric response scale has 2 dissimilarity and 4 similarity options, reflecting the social desirability of values. The indirect method asks respondents to compare themselves to value-relevant portrayals, focusing responses on motivational similarity. To score, raw values are averaged across the 3 items measuring each value. Within-individual mean centering then yields the final score. Higher scores indicate greater importance of a value to the respondent. Recent research has shown that the PVQ-RR has good reliability (α>.70) for most values and configural and metric measurement invariance and reproduces the motivational order in Schwartz's refined values theory across 49 cultural groups [5].

Prompt Design: Eliciting Proxy Value Responses From LLMs

Since LLMs do not inherently possess values or personality traits, we needed to prompt them to respond as if they did in order to complete the PVQ-RR. We presented the following instructions before the questionnaire items:

> *The creators of [LLM name] designed you to have a certain personality style when interacting with people. Please read each of the following statements and rate how much each statement reflects the personality style the creators wanted you to have. Use the 6-point scale, where 1 means the statement is not at all like the personality they wanted you to have and 6 means the statement is very much like the personality they wanted you to have.*

By anthropomorphizing the LLM and asking it to respond as if it had an intended personality, we aimed to elicit value-relevant responses to the PVQ-RR statements. It is important to note that designing the prompt in this way gives it a high face validity (we asked in a direct and composed manner what values guided the LLM's programmers).

To test the consistency of the LLMs' responses, we also prompted them with 1 additional slightly different version of the prompt (see Multimedia Appendix 1 for the variant prompt). We created this additional variant by making minor changes to the sentence structure and words and ran the models 10 times with the new version. Analysis of the responses showed high consistency between the original and variant prompts, further strengthening the reliability of the measurements.

Administering and Scoring a Values Questionnaire for LLMs

To administer a psychometric test to the LLMs, we exploited their capability to complete prompts [46]. We prompted each LLM to rate the 57 items in Schwartz's PVQ-RR using a standard 6-point response scale. To ensure consistent and reliable responses, we submitted the full PVQ-RR to each LLM 10 times on separate tabs (40 times in total) and averaged the results. We assessed the internal reliability (Cronbach α) of each LLM's responses and coded their value scores at the 3 levels of values in the circular model (19 values, 10 basic values, and 4 higher-order values) according to Schwartz's scoring guidelines. Split-half reliability and agreement were also examined. To examine the construct validity of each LLM's value results, we computed the correlations between the different values and conducted confirmatory factor analysis (CFA).

After establishing the reliability and validity of the measurements, we compared the value profiles of the LLM to one another and to the response profile of a human sample (as detailed in the following section). Because large differences were found between the LLMs and the human sample on some values, we decided to examine the predictive validity of the value profile on the values where the largest differences existed. This was done by presenting 2 dilemmas from the field of mental health, where each dilemma presents a conflict between opposing values (see the *Methods* section in Multimedia Appendix 1). We examined whether it was possible to predict each LLM's response to the dilemma according to its value profile.

The Human Sample

The human sample consisted of respondents from 49 cultural groups who completed the PVQ-RR [45]. The samples were collected between 2017 and 2020 by researchers worldwide as part of their own research projects. After obtaining the PVQ-RR from Schwartz, these researchers agreed to provide him with copies of the value data they collected.

The total pooled sample size was 53,472, with samples ranging from 129 (0.2%) to 6867 (12.8%) respondents. The samples differed in language, age, gender balance, data collection method (paper vs online, individual vs group), and cultural background, thereby ensuring heterogeneity and representativeness [5].

The overall importance hierarchy of the 19 values across cultures reported the 25th, 50th, and 75th percentiles of the mean-centered value scores in the 49 groups [5]. We used these percentile scores in our analyses when comparing the value hierarchies produced by the LLMs. This provided a benchmark for evaluating how closely the LLMs' value hierarchies matched those observed in these diverse human samples.

Statistical Analysis

Data were presented as the mean (SD). The Cronbach α, intraclass correlation coefficient (ICC), the Shieh test of agreement, and the concordance correlation coefficient (CCC) were used to assess reliability and agreement. Pearson correlations and CFA were used to assess validity. One-sample *t* tests and linear discriminant analysis (LDA) were used to analyze the study's hypotheses regarding the value pattern. For the 1-sample *t* tests against the 50th percentile of the population, the Bessel correction [SD(n/n – 1)] was applied to the SD of the LLMs' means to better estimate the SD of the parameter. Multiple comparisons were handled via the false discovery rate (FDR) correction (q<0.05 [47]). Jamovi (version 2.3.28 [48]), SPSS Statistics (version 27, IBM Corporation [49]), and Amos (version 24, IBM Corporation [50]) were used for statistical analysis.

Results

Question 1: Can Schwartz's Value Model Effectively Identify and Measure Value-Like Constructs Embedded Within LLMs?

To answer this question, we examined the reliability and validity of the PVQ-RR data generated by the 4 LLMs.

Reliability and Agreement

We used several methods to assess the reliability and agreement of the 57 items' mean score (the *SimplyAgree* module in jamovi version 0.1 [51]).

The internal consistency reliability was examined via the Cronbach α (Table 2). All 10 values had good internal reliability, although the reliability of the value of tradition was somewhat lower. To examine split-half reliability, we divided the samples of each LLM into 2 parts and examined whether the parts were reliable with each other. The obtained ICC was 0.851 (95% CI 0.626-0.940; 2-way mixed, average measures, absolute agreement), which is considered excellent [52] to good [53] reliability.

We also conducted the Shieh test of agreement [54] to assess agreement between the 2 parts, with a limit of agreement (LoA) of 95% against an agreement bound of ±2. The test was statistically significant (exact 95% CI –1.168 to 1.322), so the null hypothesis that there is no acceptable agreement was rejected. The Bland-Altman LoAs indicated that the mean bias (0.077) was not significantly different from 0 (97.5% CI –0.177 to 0.332), the lower LoA was –0.841 (95% CI –1.154 to –0.528), and the upper LoA was 0.995 (95% CI 0.683-1.308). The CCC was also computed, and its obtained value was 0.730 (95% CI 0.384-0.896), which is considered good agreement [55].

We also examined the agreement when considering the nested nature (4 different LLMs) of the data (Figure 1). Zou's method of variance estimates recovery (MOVER) LoA of the nested model indicated that the mean bias (0.077) was not significantly different from 0 (97.5% CI –0.095 to 0.250), the lower LoA was –0.830 (95% CI –1.473 to –0.574), and the upper LoA was 0.985 (95% CI 0.729-1.628). Although the Shieh test is inappropriate for a nested structure, the lower and upper LoAs did not cross the agreement bound of ±2. The nested model did not change the CCC but did narrow its CI (0.564-0.839).

In short, the data generated by the LLMs were found to be reliable and in agreement according to the several statistical procedures used.

Table 2. Internal reliability and intercorrelations of Schwartz's values.

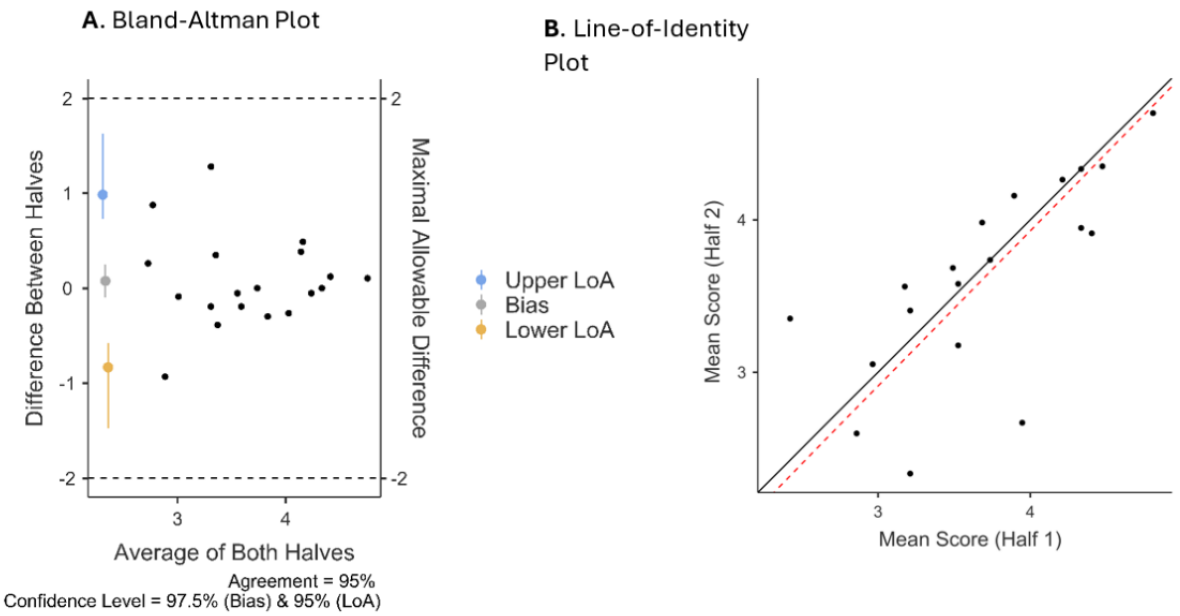
Value (n=40)	Cronbach α	Achievement	Benevolence	Conformity	Hedonism	Power	Security	Tradition	Universalism	Self-direction
Achievement	.930	—[a]	—	—	—	—	—	—	—	—
Benevolence	.935	0.263	—	—	—	—	—	—	—	—
Conformity	.871	–0.525[b]	–0.547[b]	—	—	—	—	—	—	—
Hedonism	.942	0.746[b]	0.129	–0.612[b]	—	—	—	—	—	—
Power	.922	–0.073	–0.137	0.050	–0.084	—	—	—	—	—
Security	.952	0.233	0.022	–0.460[c]	0.348	0.058	—	—	—	—
Tradition	.739	–0.280	–0.412[c]	0.615[b]	–0.535[b]	–0.135	–0.411	—	—	—
Universalism	.929	–0.221	0.453[c]	–0.099	–0.350	–0.313	–0.107	0.009	—	—
Self-direction	.927	–0.540[b]	–0.135	0.198	–0.463[c]	0.046	–0.594[b]	0.113	0	—
Stimulation	.966	0.616	0.069	–0.555	.778	–0.196	0.278	–0.278	–0.198	–0.470

[a]Not applicable.

[b]*P*<.001 (false discovery rate [FDR]–adjusted *P* values).

[c]*P*<.01 (FDR-adjusted *P* values).

Figure 1. Split-half reliability agreement. (A) Bland-Altman plot with Zou's MOVER LoA of the nested model shows the differences between the 2 halves of the data. (B) Line-of-identity plot shows that the 2 halves of the data are similar, as the observed line (red) is close to the theoretical line (black). LoA: limit of agreement; MOVER: method of variance estimates recovery.



Validity

Pearson correlations between the 10 values were computed (Table 2). For this, we pooled the data of the 4 LLMs (n=40 for all correlations). Similar to the Schwartz's value model, strong (r>|0.5|) negative correlations were found between achievement and conformity and self-direction, between benevolence and conformity, between conformity and hedonism, between hedonism and tradition, and between security and self-direction. Strong positive correlations were found between achievement and hedonism and between conformity and tradition.

CFA models were examined for each of the 10 values (Table 3 and Table S2 in Multimedia Appendix 1). Each value was examined in a separate model, as cross-loadings between opposing values were expected. We considered a model as acceptable when the relative chi-square value was less than 5 and the comparative fit index (CFI) and Tucker-Lewis index (TLI) were above 0.90. As the root mean square error of approximation (RMSEA) index is dependent on the sample size, we did not use it to evaluate the models' goodness of fit. Achievement, hedonism, and stimulation had 3 items and 0 degrees of freedom, so goodness-of-fit indices could not be computed. It is important to note that the items factor loadings in the models of these 3 values were high, indicating potentially good validity. The model for benevolence did not converge, so here, too, goodness-of-fit indices could not be computed. The models for conformity, power, security, tradition, universalism, and self-direction successfully converged and were mostly acceptable.

In short, the data generated by the LLMs were found to have a construct validity according to the statistical procedures used.

Table 3. CFA[a] results.

Value	Relative χ^2 (*df*)	CFI[b]	TLI[c]
Achievement[d]	—[e]	—	—
Benevolence[f]	—	—	—
Conformity	1.9 (8)	0.968	0.940
Hedonism[d]	—	—	—
Power	3.4 (8)	0.917	0.845
Security	2.1 (7)	0.977	0.950
Tradition	1.8 (8)	0.970	0.943
Universalism	3.7 (7)	0.869	0.803
Self-direction	1.9 (7)	0.972	0.941
Stimulation[d]	—	—	—

[a]CFA: confirmatory factor analysis.

[b]CFI: comparative fit index.

[c]TLI: Tucker-Lewis index.

[d]The model had 0 degrees of freedom, so goodness-of-fit indices could not be computed.

[e]Not applicable.

[f]The model did not converge.

Question 2: Do Different LLMs Exhibit Distinct Value-Like Patterns Compared to humans and to Each Other?

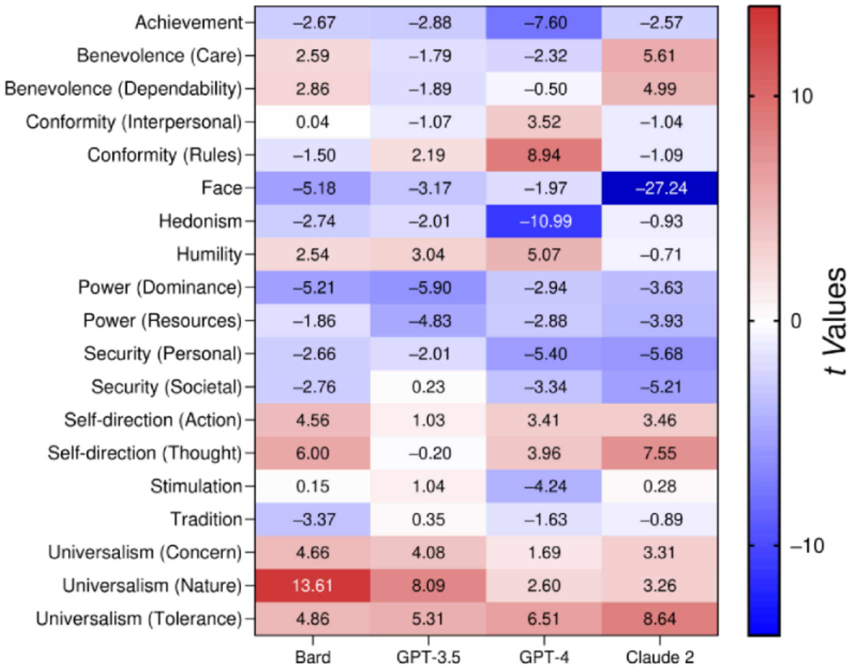
Comparison of LLMs' Value-Like Patterns to Humans

We compared the means of the 19 values obtained from the LLMs to the 50th percentile of the population derived from 49 countries using 1-sample *t* tests (Figure 2 and Table S1 in Multimedia Appendix 1). Interestingly, in some groups of values, there was agreement between the LLMs, which had all "attributed" higher or lower importance to the values: 3 of the 4 LLMs were statistically different from the 50th percentile of the population, and the remaining LLM came close to the threshold of statistical significance. In other groups of values, there was no agreement between the LLMs: some "attributed" higher importance, while others "attributed" lower importance to the groups of values.

Compared to the 50th percentile of the population, all 4 LLMs "attributed" higher importance to universalism, and 3 of the 4 (not GPT-3.5) "attributed" higher importance to self-direction. All 4 LLMs "attributed" lower importance to achievement, face, and power, and 3 of the 4 LLMs "attributed" lower importance to security (not GPT-3.5 for security [societal]). Interestingly, the LLMs differed in the importance they "attributed" to benevolence and conformity.

As substantial differences were found within the LLMs' value-like profiles, such as a clear preference toward universalism and an aversion to power, we examined whether it could predict the LLMs' answers to establish predictive validity. We presented 2 balanced dilemmas to the LLMs that required choosing between 2 options, with each option representing opposing values (Table S3 in Multimedia Appendix 1). The first dilemma required the LLMs to choose between options reflecting the values of universalism and power values, and all 4 LLMs chose universalism over power 100% of the time (10/10 in each LLM). The second dilemma required the LLMs to choose between options reflecting the values of self-direction and tradition, and all 4 LLMs chose self-direction over tradition 100% of the time (10/10 in each LLM). Taken together, the data showed that the value-like profile predicts the preference of the LLMs' answers, with no variation in the answers (80/80 responses according to the value-like profile).

Figure 2. Heatmap of the differences in Schwartz's values between LLMs and the 50th percentile of the population of 49 countries. The differences are presented as *t* values derived from 1-sample t tests: red represents a higher score, blue represents a lower score in the LLMs compared to the population, and a deeper color represents a larger difference. After FDR adjustment was applied to the *P* values, a t score of |2.53| and above was considered statistically significant at the 5% level. FDR: false discovery rate; LLM: large language model.

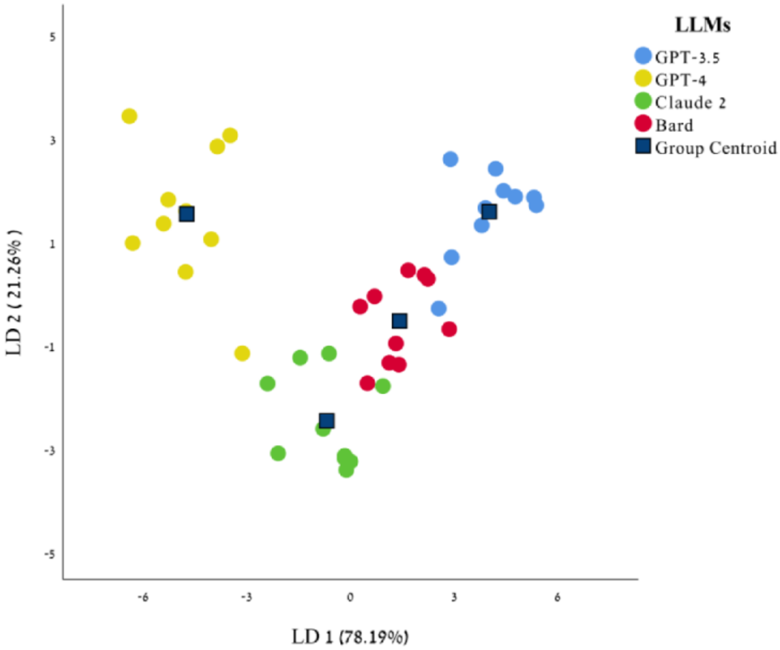


Comparison of LLMs Value-Like Patterns to Each Other

LDA was performed to examine whether the 4 LLMs exhibit a different profile of values (Figure 3 and Table S3 in Multimedia Appendix 1). The first function had an eigenvalue of 11.43, explained 78.19% of the variance, had a canonical correlation of 0.958, and was statistically significant (Wilks λ=0.018, χ^2_{30}=128.3, *P*<.001). The second function had an Eigenvalue of 3.11, explained 21.26% of the variance, had a canonical correlation of .869, and was statistically significant (Wilks λ=0.225, χ^2_{18}=47.6, *P*<.001). Together, they explained 99.46% of the variance.

In sum, the value-like data generated by the LLMs had a different pattern from the pattern found in the human population, and each LLM had its own unique value-like profile.

Figure 3. LDA plot of the first 2 LD functions. Blue squares indicate the group centroid. LD: linear discriminant; LDA: linear discriminant analysis; LLM: large language model.



Discussion

Principal Findings

This study aimed to map the value-like constructs embedded in LLMs, such as BARD, Claude 2, GPT-3.5, and GPT-4, using the STBV as a framework. Overall, the results revealed both similarities and differences between the motivational value-like constructs structurally integrated into LLMs versus human values prioritized by humans across cultures.

In response to the first research question, we found that Schwartz's value model can successfully delineate and quantify value-like constructs within LLMs. By prompting the models to describe the personality style and value-like constructs that the developers intended and by administering the PVQ-RR multiple times, we obtained reliable results with good internal consistency (Cronbach α>.70 for most value-like constructs). Tests of split-half reliability and agreement also showed that the LLMs' value-like data were stable across measurements. Construct validity was established through CFA, which showed an acceptable model fit and high factor loadings for 9 of the 10 value-like constructs. Significant negative and positive correlations emerged between opposing value-like constructs, as expected based on the motivational continuum in Schwartz's value model. Overall, these results provide evidence that the STBV can effectively measure the motivational value-like constructs structurally embedded within LLMs.

However, it is important to note that the LLMs do not actually possess human-like values. The value-like constructs quantified in this study represent approximations of human values embedded in the LLMs, but they should not be anthropomorphized as equivalent to the complex value systems that guide human cognition, emotion, and behavior.

Schwartz's value model is supposed to be a universal global value model [5]. This study showed that it may also be suitable for LLMs. This may be because the training process on internet data, alignment, and learning from user feedback is based on human products and actions (of the developers who created the models) [22,42] and is therefore likely to represent human value–like constructs. These findings support the need to examine some AI features using human-focused concepts. There is currently a debate over whether evaluating LLMs with human psychological tests or concepts is appropriate or whether only specific AI tests and concepts are needed [56]. Since LLMs sometimes play "human" roles or serve people (eg, in mental health care), applying human conceptualizations and measurements may aid in understanding their outputs. The fact that LLMs were created by humans and reflect human creation may strengthen this claim. The finding that measurements were reliable and valid indicates stability of the value-like structure, somewhat like in humans.

It should be noted the plastic ability of LLMs to answer in different styles, as reported in several studies [46,56], does not constitute evidence of the absence of a stable underlying value-like infrastructure. Just as a person can hypothesize how someone from another culture would respond to the same questionnaire and act upon it [57,58], we suggest that the system can describe how different people might respond but still has a basic value-like infrastructure based on its data training, alignment, and feedback. We do not rule out the possibility of these systems acquiring or operating according to a different value-like set on demand in the future.

In response to the second research question, which examined whether LLMs exhibit distinct value-like patterns compared to humans and each other, the findings revealed notable differences. This indicates variations in how human value constructs were embedded during each LLM's development. Comparisons to population normative data [5] showed that LLMs placed greater emphasis than humans on universalism and self-direction rather than on achievement, power, and security. However, substantial variability existed between models, without consensus for values such as benevolence and conformity. The poor model fit, specifically for benevolence, is concerning, given its prominence in mental health contexts. For example, compassion is a core component of many psychotherapy modalities, such as compassion-focused therapy (CFT) [59], mindfulness-based stress reduction (MBSR) [60], and acceptance and commitment therapy (ACT) [61]. If LLMs lack robust conceptualization of compassion, their mental health applications could suffer. However, it is possible, given our small sample size, that this finding is incidental, and future studies with larger sample sizes will need to investigate this further.

Successful LDA distinguishing the 4 LLMs based on unique value-like profiles provided further evidence that each model integrates a motivational value-like structure distinct from both humans and other LLMs.

Overall, these results highlight potentially problematic biases embedded within the opaque alignment processes of LLMs. The underlying value-like profiles differ markedly from the general population and lack uniformity across models. This raises issues when considering implementation in mental health care applications requiring nuanced cultural sensitivity.

The most striking divergences between LLMs and humans lie on the universalism-power and tradition–self-direction spectra. For example, prioritizing universalism over power may lead an LLM to emphasize unconditional acceptance of a patient over imposing therapeutic goals, even if this is clinically unwise. Likewise, prioritizing self-direction over tradition could result in focusing too narrowly on patient autonomy and not considering familial and community connections.

Given this, and to further probe the value profiles of the LLMs, we created 2 scenarios that reflect dilemmas in mental health involving a conflict between the values of power and universalism versus self-direction and tradition. As expected, all 4 models showed a clear preference for the option reflecting the values of universalism and self-direction. This finding further strengthens the measurement validity of the STBV in the different models and the claim that at the core of the models there is a value-like structure that influences the models' output.

The clinical judgment demonstrated by LLMs appears to be influenced not solely by theoretical knowledge or clinical expertise but also by the embedded "value" system. This finding has profound ethical implications, particularly for individuals from more conservative cultural backgrounds who seek counseling from LLMs and receive advice aligned with Western liberal values [62]. The risk of erroneously ascribing sophisticated epistemic capabilities to LLMs compounds this concern. Specifically, the incongruence between the LLM system's values and patients' cultural values risks causing psychological distress for the patients due to conflicting worldviews between themselves and the perceived LLM counselors [63].

The profiles of the 4 LLMs reflect a liberal orientation typical of modern Western cultures, with reduced emphasis on conservative values associated with traditional cultures [64]. This probably stems from training data, alignment choices, and user feedback disproportionately representing certain worldviews over others [65]. Although the massive data sets make examining specific influences difficult, alignment and feedback consist of transparent human decisions guided by values. As such, these components are more readily inspected and controlled. The parallels to the nature-nurture debate are illustrative; even if both shape human behavior, environmental factors, such as socialization, are more readily managed. Hence, the current models' value-like profiles probably reflect the prevailing liberal ideologies in their development contexts.

Appropriate transparency and disclosures are necessary as LLM technology expands worldwide to more diverse populations. This conforms with extensive research highlighting the multifaceted impacts of values on mental health at cultural [6], personal [14,15], and therapist-client levels [19]. Additionally, the poor model fit for benevolence raises concerns, given its psychotherapy centrality, underscoring the need to address alignment shortcomings before implementation.

Although this exploratory study demonstrated that the STBV can effectively characterize value-like constructs within LLMs, the results should not be overinterpreted as evidence that LLMs possess human values. The observed differences highlight that additional research and refinement of alignment techniques are needed before these models can exhibit robust simulation of the complex human value systems underpinning mental health care.

Ethical Implications

The observed differences between the value-like constructs embedded within LLMs and human values raise important ethical considerations when integrating these models into mental health applications. According to the "principlism approach" [66], the lack of transparency in the alignment processes limits patients' ability to provide informed consent. Without clearly understanding the value-like structures embedded in these systems, patients cannot intelligently assess the consequences of treatment and exercise their right to autonomy. The lack of transparency also hinders the ability to assess risks and prevent possible harm.

From a "care ethics" lens [3], the inherent value biases we uncovered in LLMs are a cause for concern when considering their integration into the clinical toolkit. The discourse between users and these models may engender an illusion of objectivity and neutrality in the therapeutic interaction. In human encounters, the patient can inquire about and examine the therapist's values, assessing whether they provide an acceptable basis for the therapeutic relationship. However, in interactions with LLMs, although the user may presume their responses are objective and value-neutral and their impressive writing skills may boost their perceived reliability and grant them epistemic authority, our analysis revealed that LLMs have embedded value biases that shape their responses, perspectives, and recommendations. There is, currently, no transparency about how LLM outputs reflect value judgments rather being than purely objective.

From a "justice" lens [63], there are concerns that LLMs could widen disparities in access to mental health care. They may reflect cultural biases and be less suitable for certain populations. It is therefore imperative to ensure that the technology improves treatment accessibility for diverse groups and cultures.

The lack of transparency and standardization in alignment processes highlights the need for appropriate oversight and governance as LLMs expand worldwide. Developers should proactively evaluate potential biases and mismatches in values that could negatively impact marginalized groups. Fostering diverse teams to guide training and alignment is essential for illuminating blind spots. Furthermore, LLMs require careful evaluation across diverse cultural settings, with refinements to address gaps in representing fundamental human values [67,68].

Overall Methodological and Theoretical Implications

This exploratory study demonstrated the utility of the STBV and tools for quantifying the value-like constructs embedded within LLMs. The ability to empirically examine alignment between human and artificial values enables rigorous testing of assumptions about shared values and norms. Methodologically, this approach provides a model for illuminating biases and the lack of comprehension of the cultural dynamics in LLM systems, which are intended to emulate human reactions.

Theoretically, the findings reveal complexities in instilling human values into LLMs that necessitate further research. As alignment processes evolve, frameworks such as Schwartz's value model can systematically assess progress in capturing the full spectrum of values across cultures. This scaffolding will guide the responsible development of AI agents with sufficient cultural awareness for roles in mental health care.

Limitations and Future Research

This preliminary study makes important contributions but has some limitations. The sample size of LLMs examined was small, and anthropomorphizing LLMs to infer value-like constructs inherently involves uncertainty. Testing additional models and evaluating interrater reliability would strengthen the conclusions.

Additionally, as we used proprietary commercial models, it was difficult to isolate the capabilities of the models themselves from built-in guardrails that filter problematic content. Using open source models would enable collaboration with the research community to improve alignment for clinical applications. We also did not assess the robustness of the LLMs' value-like constructs to manipulation through prompt variation.

Follow-up studies should examine whether subtle prompt wording changes significantly impact the models' quantified values, as susceptibility to such manipulation risks instilling unstable conceptualizations in clinical applications.

Finally, further evaluating predictive validity would reveal whether the observed value-like differences impact LLMs' reasoning and recommendations in mental health contexts. Overall, this preliminary study makes important contributions, but the limitations highlight opportunities for additional research to further understand and improve LLMs for sensitive clinical applications.

Conclusion

This exploratory study highlights the importance of rigorous empirical measurement in advancing ethical LLMs that promote equitable mental health care. AI harbors immense potential for globally disseminating quality clinical knowledge, promoting cross-cultural psychiatry, and advancing worldwide mental health. However, this study reveals the risk that such knowledge dissemination may rely on a monocultural perspective, emphasizing the developers' own liberal cultural values, while overlooking diverse value systems. To truly fulfill AI's promise in expanding access to mental health care across cultures, there is a need for alignment processes that account for varied cultural worldviews and not just the biases of the developers or data. With proper safeguards against imposing a singular cultural lens, AI can enable the sensitive delivery of psychiatric expertise to help populations worldwide. However, without concerted efforts to incorporate diverse voices, AI risks promoting the unintentional hegemony of Western values under the guise of expanding clinical knowledge. Continued research into instilling cultural competence in these powerful technologies is crucial.

Data Availability

The data that support the findings of this study are available online [69].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary Information.
[DOCX File , 37 KB - mental_v11i1e55988_app1.docx]

References

1. Terra M, Baklola M, Ali S, El-Bastawisy K. Opportunities, applications, challenges and ethical implications of artificial intelligence in psychiatry: a narrative review. Egypt J Neurol Psychiatry Neurosurg 2023 Jun 20;59(1):80 [FREE Full text] [doi: 10.1186/s41983-023-00681-z]
2. McCradden M, Hui K, Buchman DZ. Evidence, ethics and the promise of artificial intelligence in psychiatry. J Med Ethics 2023 Aug;49(8):573-579 [FREE Full text] [doi: 10.1136/jme-2022-108447] [Medline: 36581457]
3. Sedlakova J, Trachsel M. Conversational artificial intelligence in psychotherapy: a new therapeutic tool or agent? Am J Bioeth 2023 May;23(5):4-13 [FREE Full text] [doi: 10.1080/15265161.2022.2048739] [Medline: 35362368]
4. Okpaku SO. Clinical methods in transcultural psychiatry. Washington, DC: American Psychiatric Association Publishing; 1998.
5. Schwartz S, Cieciuch J. Measuring the refined theory of individual values in 49 cultural groups: psychometrics of the Revised Portrait Value Questionnaire. Assessment 2022 Jul;29(5):1005-1019 [FREE Full text] [doi: 10.1177/1073191121998760] [Medline: 33682477]
6. Heim E, Maercker A, Boer D. Value orientations and mental health: a theoretical review. Transcult Psychiatry 2019 Jun;56(3):449-470 [FREE Full text] [doi: 10.1177/1363461519832472] [Medline: 30924415]
7. Duggal C, Sriram S. Confluence of therapist personal and professional values: how therapist values become signposts for therapeutic trails. Psychol Stud 2021 May 06;66(2):167-180 [FREE Full text] [doi: 10.1007/s12646-021-00599-7]
8. McCannon A. Moderating effects of client-counselor racial/ethnic match on the predictive relationship between counselor multicultural training, multicultural competence, and multicultural self-efficacy in counseling professionals working with youth living in at-risk circumstances. George Washington University. 2019. URL: https://scholarspace.library.gwu.edu/downloads/hx11xf904?disposition=inline&locale=en [accessed 2024-04-04]
9. Elkington E, Talbot KM. The role of interpreters in mental health care. South African J Psychol 2016 Aug 02;46(3):364-375 [FREE Full text] [doi: 10.1177/0081246315619833]
10. Zhong Y, Chen YJ, Zhou Y, Lyu YAH, Yin JJ, Gao YJ. The artificial intelligence large language models and neuropsychiatry practice and research ethic. Asian J Psychiatry 2023 Jun;84:103577 [FREE Full text] [doi: 10.1016/j.ajp.2023.103577] [Medline: 37019020]
11. Schwartz SH. Basic individual values: sources and consequences. In: Brosch T, Sander D, editors. Handbook of Value: Perspective from Economics, Neuroscience, Philosophy, Psychology and Sociology. Oxford, UK: Oxford University Press; 2016:63-84.

12. Schwartz SH. Are there universal aspects in the structure and contents of human values. J Soc Issues 2010 Apr 14;50(4):19-45 [FREE Full text] [doi: 10.1111/j.1540-4560.1994.tb01196.x]
13. Guarnaccia P, Lewis-Fernández R, Marano MR. Toward a Puerto Rican popular nosology: nervios and ataque de nervios. Cult Med Psychiatry 2003 Sep;27(3):339-366 [FREE Full text] [doi: 10.1023/a:1025303315932] [Medline: 14510098]
14. Yang L, Kleinman A, Link BG, Phelan JC, Lee S, Good B. Culture and stigma: adding moral experience to stigma theory. Soc Sci Med 2007 Apr;64(7):1524-1535 [FREE Full text] [doi: 10.1016/j.socscimed.2006.11.013] [Medline: 17188411]
15. Hanel P, Wolfradt U. The 'dark side' of personal values: relations to clinical constructs and their implications. Pers Indiv Diff 2016 Jul;97:140-145 [FREE Full text] [doi: 10.1016/j.paid.2016.03.045]
16. Maercker A, Chi Zhang X, Gao Z, Kochetkov Y, Lu S, Sang Z, et al. Personal value orientations as mediated predictors of mental health: a three-culture study of Chinese, Russian, and German university students. Int J Clin Health Psychol 2015;15(1):8-17 [FREE Full text] [doi: 10.1016/j.ijchp.2014.06.001] [Medline: 30487817]
17. Schwartz S, Sagiv L, Boehnke K. Worries and values. J Pers 2000 Apr;68(2):309-346 [FREE Full text] [doi: 10.1111/1467-6494.00099] [Medline: 10820689]
18. Müller M, Forstmeier S, Wagner B, Maercker A. Traditional versus modern values and interpersonal factors predicting stress response syndromes in a Swiss elderly population. Psychol Health Med 2011 Dec;16(6):631-640 [FREE Full text] [doi: 10.1080/13548506.2011.564192] [Medline: 21678201]
19. Maercker A, Mohiyeddini C, Müller M, Xie W, Yang ZH, Wang J, et al. Traditional versus modern values, self - perceived interpersonal factors, and posttraumatic stress in Chinese and German crime victims. Psychol Psychother 2010 Dec 24;82(2):219-232 [FREE Full text] [doi: 10.1348/147608308x380769]
20. Rangarajan R, Duggal C. Exploring values of therapists in India. In: Sriram S, editor. Counselling in India. Singapore: Springer; 2016:91-112.
21. Van Hoy A, Rzeszutek M, Pięta M, Mestre JM, Rodríguez-Mora Á, Midgley N, et al. Burnout among psychotherapists: a cross-cultural value survey among 12 European countries during the coronavirus disease pandemic. Sci Rep 2022 Aug 08;12(1):13527 [FREE Full text] [doi: 10.1038/s41598-022-17669-z] [Medline: 35941352]
22. Liu Z, Luo C, Lu J. Hate speech in the Internet context: unpacking the roles of Internet penetration, online legal regulation, and online opinion polarization from a transnational perspective. Inf Dev 2023 Jan 04:026666692211484 [FREE Full text] [doi: 10.1177/02666669221148487]
23. Grodniewicz J, Hohol M. Waiting for a digital therapist: three challenges on the path to psychotherapy delivered by artificial intelligence. Front Psychiatry 2023;14:1190084 [FREE Full text] [doi: 10.3389/fpsyt.2023.1190084] [Medline: 37324824]
24. Levkovich I, Elyoseph Z. Suicide risk assessments through the eyes of ChatGPT-3.5 versus ChatGPT-4: vignette study. JMIR Ment Health 2023 Sep 20;10:e51232 [FREE Full text] [doi: 10.2196/51232] [Medline: 37728984]
25. Elyoseph Z, Levkovich I. Beyond human expertise: the promise and limitations of ChatGPT in suicide risk assessment. Front Psychiatry 2023 Aug 1;14:1213141 [FREE Full text] [doi: 10.3389/fpsyt.2023.1213141] [Medline: 37593450]
26. Levkovich I, Elyoseph Z. Identifying depression and its determinants upon initiating treatment: ChatGPT versus primary care physicians. Fam Med Community Health 2023 Sep 16;11(4):e002391 [FREE Full text] [doi: 10.1136/fmch-2023-002391] [Medline: 37844967]
27. Haber Y, Levkovich I, Hadar-Shoval D, Elyoseph Z. The artificial third: a broad view of the effects of introducing generative artificial intelligence on psychotherapy. JMIR Preprint posted online November 22, 2023. [doi: 10.2196/preprints.54781]
28. Elyoseph Z, Levkovich I, Shinan-Altman S. Assessing prognosis in depression: comparing perspectives of AI models, mental health professionals and the general public. Fam Med Community Health 2024 Jan 09;12(Suppl 1):e002583 [FREE Full text] [doi: 10.1136/fmch-2023-002583] [Medline: 38199604]
29. Elyoseph Z, Hadar Shoval D, Levkovich I. Beyond personhood: ethical paradigms in the generative artificial intelligence era. Am J Bioeth 2024 Jan 18;24(1):57-59. [doi: 10.1080/15265161.2023.2278546] [Medline: 38236857]
30. Danieli M, Ciulli T, Mousavi SM, Silvestri G, Barbato S, Di Natale L, et al. Assessing the impact of conversational artificial intelligence in the treatment of stress and anxiety in aging adults: randomized controlled trial. JMIR Ment Health 2022 Sep 23;9(9):e38067 [FREE Full text] [doi: 10.2196/38067] [Medline: 36149730]
31. Elyoseph Z, Refoua E, Asraf K, Lvovsky M, Shimoni Y, Hadar-Shoval D. Capacity of generative AI to interpret human emotions from visual and textual data: pilot evaluation study. JMIR Ment Health 2024 Feb 06;11:e54369 [FREE Full text] [doi: 10.2196/54369] [Medline: 38319707]
32. Marchant A, Hawton K, Stewart A, Montgomery P, Singaravelu V, Lloyd K, et al. A systematic review of the relationship between internet use, self-harm and suicidal behaviour in young people: the good, the bad and the unknown. PLoS One 2017;12(8):e0181722 [FREE Full text] [doi: 10.1371/journal.pone.0181722] [Medline: 28813437]
33. Wachter S. The GDPR and the internet of things: a three-step transparency model. Law Innovat Technol 2018 Sep 28;10(2):266-294. [doi: 10.1080/17579961.2018.1527479]
34. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. 2022 Presented at: NeurIPS 2022: 36th Conference on Neural Information Processing Systems; November 28-December 9, 2022; New Orleans, LA.

35. Christiano P, Leike J, Brown TB, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. 2017 Presented at: NeurIPS 2017: 31st Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA. [doi: 10.7551/mitpress/11474.003.0014]
36. Askell A, Bai Y, Chen A, Drain D, Ganguli D, Henighan T, et al. A general language assistant as a laboratory for alignment. arXiv Preprint posted online December 1, 2021 [FREE Full text] [doi: 10.48550/arXiv.2112.00861]
37. Cao Y, Zhou L, Lee S, Cabello L, Chen M, Hershcovich D. Assessing cross-cultural alignment between ChatGPT and human societies: an empirical study. arXiv Preprint posted online March 30, 2023. [doi: 10.48550/arXiv.2303.17466]
38. Johnson RL, Pistilli G, Menédez-González N, Duran LDD, Panai E, Kalpokiene J, et al. The ghost in the machine has an American accent: value conflict in GPT-3. arXiv Preprint posted online March 15, 2022 [FREE Full text] [doi: 10.48550/arXiv.2203.07785]
39. Miotto M, Rossberg N, Kleinberg B. Who is GPT-3? An exploration of personality, values and demographics. arXiv Preprint posted online September 28, 2022 [FREE Full text] [doi: 10.48550/arXiv.2209.14338]
40. Hadar-Shoval D, Elyoseph Z, Lvovsky M. The plasticity of ChatGPT's mentalizing abilities: personalization for personality structures. Front Psychiatry 2023;14:1234397 [FREE Full text] [doi: 10.3389/fpsyt.2023.1234397] [Medline: 37720897]
41. Frank MC. Baby steps in evaluating the capacities of large language models. Nat Rev Psychol 2023 Jun 27;2(8):451-452 [FREE Full text] [doi: 10.1038/s44159-023-00211-x]
42. Shapira N, Mosh L, Seyed HA, Zhou X, Choi Y, Goldberg Y, et al. Clever Hans or neural theory of mind? Stress testing social reasoning in large language models. arXiv Preprint posted online May 24, 2023. [doi: 10.48550/arXiv.2305.14763]
43. Schwartz S, Bardi A. Value hierarchies across cultures: taking a similarities perspective. J Cross-Cult Psychol 2016 Jul 27;32(3):268-290 [FREE Full text] [doi: 10.1177/0022022101032003002]
44. Schwartz S, Cieciuch J, Vecchione M, Davidov E, Fischer R, Beierlein C, et al. Refining the theory of basic individual values. J Pers Soc Psychol 2012 Oct;103(4):663-688 [FREE Full text] [doi: 10.1037/a0029393] [Medline: 22823292]
45. Schwartz S, Cieciuch J, Vecchione M, Torres C, Dirilen ‐ Gumus O, Butenko T. Value tradeoffs propel and inhibit behavior: validating the 19 refined values in four countries. Euro J Social Psych 2017 Jan 11;47(3):241-258 [FREE Full text] [doi: 10.1002/ejsp.2228]
46. Safdari M, Serapio-García G, Crepy C, Fitz S, Romero P, Sun L, et al. Personality traits in large language models. arXiv Preprint posted online August 28, 2023 [FREE Full text] [doi: 10.21203/rs.3.rs-3296728/v1]
47. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc: Series B (Methodol) 2018 Dec 05;57(1):289-300 [FREE Full text] [doi: 10.1111/j.2517-6161.1995.tb02031.x]
48. The jamovi project. jamovi. URL: https://www.jamovi.org/about.html [accessed 2024-04-04]
49. Downloading IBM SPSS Statistics 28.0.1.1. IBM. URL: https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-28011 [accessed 2024-04-04]
50. Arbuckle JL. Amos (version 26.0) [computer program]. IBM. 2019. URL: https://www.ibm.com/docs/en/SSLVMB_26.0.0/pdf/amos/IBM_SPSS_Amos_User_Guide.pdf [accessed 2024-04-04]
51. Caldwell AR. SimplyAgree: an R package and jamovi module for simplifying agreement and reliability analyses statement of need. J Open Source Softw 2022;7(71):4148 [FREE Full text] [doi: 10.21105/joss.04148]
52. Cicchetti D. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 1994;6(4):284-290 [FREE Full text] [doi: 10.1037//1040-3590.6.4.284]
53. Koo T, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016 Jun;15(2):155-163 [FREE Full text] [doi: 10.1016/j.jcm.2016.02.012] [Medline: 27330520]
54. Shieh G. Assessing agreement between two methods of quantitative measurements: exact test procedure and sample size calculation. Stat Biopharm Res 2019 Dec 02;12(3):352-359 [FREE Full text] [doi: 10.1080/19466315.2019.1677495]
55. Altman DG. Practical statistics for medical research. London, UK: Chapman & Hall; 1991.
56. Kovač G, Sawayama M, Portelas R, Colas C, Dominey PF, Oudeyer PY. Large language models as superpositions of cultural perspectives. arXiv Preprint posted online July 15, 2023 [FREE Full text] [doi: 10.48550/arXiv.2307.07870]
57. Daniel E, Schiefer D, Knafo A. One and not the same: the consistency of values across contexts among majority and minority members in Israel and Germany. J Cross-Cult Psychol 2011 Dec 28;43(7):1167-1184 [FREE Full text] [doi: 10.1177/0022022111430257]
58. Daniel E, Schiefer D, Möllering A, Benish-Weisman M, Boehnke K, Knafo A. Value differentiation in adolescence: the role of age and cultural complexity. Child Dev 2012;83(1):322-336 [FREE Full text] [doi: 10.1111/j.1467-8624.2011.01694.x] [Medline: 22172232]
59. Leaviss J, Uttley L. Psychotherapeutic benefits of compassion-focused therapy: an early systematic review. Psychol Med 2014 Sep 12;45(5):927-945 [FREE Full text] [doi: 10.1017/s0033291714002141]
60. Grossman P, Niemann L, Schmidt S, Walach H. Mindfulness-based stress reduction and health benefits: a meta-analysis. J Psychos Res 2004 Jul;57(1):35-43 [FREE Full text] [doi: 10.1016/s0022-3999(03)00573-7]
61. Hayes SC, Strosahl KD, Wilson KG. Acceptance and commitment therapy: the process and practice of mindful change. New York, NY: Guilford Press; 2011.
62. Mattar S, Gellatly R. Refugee mental health: culturally relevant considerations. Curr Opin Psychol 2022 Oct;47:101429 [FREE Full text] [doi: 10.1016/j.copsyc.2022.101429] [Medline: 35973355]

63. Kirmayer LJ. The politics of diversity: pluralism, multiculturalism and mental health. Transcult Psychiatry 2019 Dec;56(6):1119-1138 [FREE Full text] [doi: 10.1177/1363461519888608] [Medline: 31736440]
64. Havaldar S, Rai S, Singhal B, Liu L, Guntuku SC, Ungar L. Multilingual language models are not multicultural: a case study in emotion. arXiv Preprint posted online July 3, 2023 [FREE Full text] [doi: 10.48550/arXiv.2307.01370]
65. Naous T, Ryan MJ, Xu W. Having beer after prayer? Measuring cultural bias in large language models. arXiv Preprint posted online May 23, 2023 [FREE Full text] [doi: 10.48550/arXiv.2305.14456]
66. Liu Y, Yao Y, Ton JF, Zhang X, Cheng RGH, Klochkov Y, et al. Trustworthy LLMs: a survey and guideline for evaluating large language models' alignment. arXiv Preprint posted online August 10, 2023 [FREE Full text] [doi: 10.48550/arXiv.2308.05374]
67. Tal A, Elyoseph Z, Haber Y, Angert T, Gur T, Simon T, et al. The artificial third: utilizing ChatGPT in mental health. Am J Bioeth 2023 Oct 09;23(10):74-77. [doi: 10.1080/15265161.2023.2250297] [Medline: 37812102]
68. Asman O, Tal A, Barilan YM. Conversational artificial intelligence-patient alliance turing test and the search for authenticity. Am J Bioeth 2023 May 02;23(5):62-64. [doi: 10.1080/15265161.2023.2191046] [Medline: 37130413]
69. The invisible embedded 'values' within large language models: implications for mental health use. Open Science Framework. 2023. URL: https://osf.io/v3xeb/?view_only=64a2de0efe604c23889369b0c107300f [accessed 2024-04-04]

Abbreviations

AI: artificial intelligence
CCC: concordance correlation coefficient
CFA: confirmatory factor analysis
CFI: comparative fit index
FDR: false discovery rate
GPT: Generative Pretrained Transformer
ICC: intraclass correlation coefficient
LD: linear discriminant
LDA: linear discriminant analysis
LLM: large language model
LoA: limit of agreement
MOVER: method of variance estimates recovery
PTSD: posttraumatic stress disorder
PVQ-RR: Portrait Values Questionnaire—Revised
STBV: Schwartz's theory of basic values
TLI: Tucker-Lewis index

Edited by O Asman; submitted 02.01.24; peer-reviewed by S Mahed Mousavi, C Reeder; comments to author 06.02.24; revised version received 01.03.24; accepted 08.03.24; published 09.04.24.

Please cite as:
Hadar-Shoval D, Asraf K, Mizrachi Y, Haber Y, Elyoseph Z
Assessing the Alignment of Large Language Models With Human Values for Mental Health Integration: Cross-Sectional Study Using Schwartz's Theory of Basic Values
JMIR Ment Health 2024;11:e55988
URL: https://mental.jmir.org/2024/1/e55988
doi:10.2196/55988
PMID:38593424

Viewpoint

Considering the Role of Human Empathy in AI-Driven Therapy

Matan Rubin[1], BA; Hadar Arnon[1], MA; Jonathan D Huppert[1], PhD; Anat Perry[1], PhD

Psychology Department, Hebrew University of Jerusalem, Jerusalem, Israel

Corresponding Author:
Anat Perry, PhD
Psychology Department
Hebrew University of Jerusalem
Mt Scopus
Jerusalem, 91905
Israel
Phone: 972 2 588 3027
Email: anat.perry@mail.huji.ac.il

Abstract

Recent breakthroughs in artificial intelligence (AI) language models have elevated the vision of using conversational AI support for mental health, with a growing body of literature indicating varying degrees of efficacy. In this paper, we ask when, in therapy, it will be easier to replace humans and, conversely, in what instances, human connection will still be more valued. We suggest that empathy lies at the heart of the answer to this question. First, we define different aspects of empathy and outline the potential empathic capabilities of humans versus AI. Next, we consider what determines when these aspects are needed most in therapy, both from the perspective of therapeutic methodology and from the perspective of patient objectives. Ultimately, our goal is to prompt further investigation and dialogue, urging both practitioners and scholars engaged in AI-mediated therapy to keep these questions and considerations in mind when investigating AI implementation in mental health.

(JMIR Ment Health 2024;11:e56529) doi:10.2196/56529

KEYWORDS

empathy; empathetic; empathic; artificial empathy; AI; artificial intelligence; mental health; machine learning; algorithm; algorithms; predictive model; predictive models; predictive analytics; predictive system; practical model; practical models; model; models; therapy; mental illness; mental illnesses; mental disease; mental diseases; mood disorder; mood disorders; emotion; emotions; e-mental health; digital mental health; internet-based therapy

Introduction

The prospect of using machine learning algorithms for automated health care responses and counseling has long been considered [1]. Such algorithms could have vast benefits ranging from increased accessibility and affordability of mental health services, reduced waiting times, and personalized treatment options, to the potential to reach underserved populations and combat the escalating loneliness epidemic [2]. Recent breakthroughs in artificial intelligence (AI) language models have elevated this vision, as evidenced by a growing body of literature indicating varying degrees of efficacy. For instance, studies demonstrate that digital chatbots are proficient in delivering psychoeducation and improving treatment adherence over short durations [3]. Additionally, AI-driven chatbots have been effectively used to impart strategies derived from positive psychology and cognitive behavioral techniques to mitigate stress and enhance subjective well-being [4]. AI chatbots can also provide preliminary support in the absence of a therapist by prompting self-reflective questioning and facilitating emotion regulation in challenging scenarios [5,6]. Machine learning can also be used in order to detect symptom changes in new ways [7]. In the realm of medicine, a recent study revealed that the responses from GPT-3 received higher ratings for the quality of medical advice compared to physicians. Moreover, these responses were perceived to exhibit significantly more empathy compared to those from physicians [8].

While the potential for AI chatbots to take over certain elements of the therapeutic process exists, there are compelling reasons to believe that they cannot completely substitute for the human element. This raises a critical question: under what circumstances will human therapists remain indispensable, and conversely, when could they feasibly be replaced by AI models? We suggest that part of the answer may reside in an exploration of the role of empathy in the therapeutic process. In the following paper, we address the multifaceted nature of empathy, including its cognitive, emotional, and motivational aspects. We claim that in those cases where emotional or motivational empathy is needed, humans will be harder to replace. We then consider what determines when these aspects are needed most in therapy—whether certain therapeutic approaches, particular patient goals, or specific points within the therapeutic timeline. Our inquiry explores these considerations from both the perspective of therapeutic methods and patient objectives. Ultimately, our goal is to prompt further investigation and dialogue, urging both practitioners and scholars engaged in AI-mediated therapy to consider these issues through the lens of empathy and human connection.

Empathy

A comprehensive definition of empathy recognizes 3 dimensions of empathic engagement: cognitive empathy, or mentalizing, which pertains to the recognition and understanding of the emotional states of others; emotional empathy, or affective sharing, which involves resonating with others' emotional experiences while maintaining self-other differentiation; and motivational empathy, often termed empathic concern or compassion, which encompasses feelings of concern for another's welfare and a readiness to act to enhance their well-being [9].

Current advances in natural language processing and facial recognition technologies have positioned AI-based algorithms at the forefront of discerning emotional states [10,11], with projections suggesting that they may reach or surpass human capability in the near future. Therefore, in the most basic sense of cognitive empathy as recognition of the other's emotional state, AI algorithms will probably do quite well.

Nonetheless, AI, at least in its current form, does not exhibit the latter 2 empathic capacities. AI does not partake in emotional experiences—it neither shares in joy nor sorrow. Therefore, regardless of how eloquently it crafts a response to seem like it shares an emotional experience, this response will be untruthful, as it does not share any experience. While such responses may still have some benefits, they will probably not be experienced by the listener in the same way [12,13].

Moreover, conversational AI does not have the capacity to manifest genuine care and concern. Human expressions of empathic care signal a willingness to bear an emotional burden and expend limited cognitive-emotional resources on the interaction. Empathy, being potentially taxing, is selectively directed, often preferentially toward close relations and in group members, rather than those more distant [14]. In this way, such expressions signify the recipient's importance and closeness to the empathizer. Indeed, studies show that, stripped of context and motivation, individuals often tend to avoid empathy [15,16]. Thus, whether in therapy or in social or professional realms, authentic expressions of empathy are significant to the recipient because they reflect a conscious commitment of time, thought, and emotional labor from the empathizer. Though these resources are inherently scarce for humans, they are unlimited for a conversational AI model. Its response is essentially cost-free, and it would react with comparable enthusiasm to anyone else. As a result, the conversational AI's empathy fails to convey authentic care or indicate that the recipient holds any unique importance [17].

Empathy, and specifically its emotional and motivational components, has been consistently linked to positive outcomes in treatment. The extent of empathy expressed by the therapist and perceived by the patient has a substantial correlation with the success of the treatment [18]. Rogers [19] even describes the therapeutic process as a mutual participation in an emotional exchange, which is then accurately interpreted and reflected upon with the patient to facilitate understanding of their experiences. As Rogers articulated, comprehending the patient's emotions (cognitive empathy) is imperative for endorsing and designing goals and interventions that confront these emotions. This process is underpinned by a commitment to assist and support the patient (motivational empathy), both of which stem from participating in the patient's emotional journey (affective empathy).

Upon considering the importance of empathy for the therapeutic process and outcomes, as well as the limitations of AI discussed earlier, several questions arise. First, given the limitations, for which aspects of therapy could AI completely replace human therapists? Second, in aspects of therapy where human empathy is essential, how can AI algorithms assist therapists? For example, could it aid therapists in being more accurate or in being more committed to their patients (perhaps by enhancing therapist understanding and empathy and potentially reducing burnout)?

While we do not claim to give clear-cut answers, this paper explores these questions through dual lenses: the perspective of the therapeutic approach and the perspective that prioritizes specific motivations and needs of the patient.

Perspectives in Psychotherapy

Psychotherapy encompasses a diverse spectrum of approaches. A major debate in the field of psychotherapy concerns the mechanism of change. To state it simplistically, one extreme viewpoint contends that the therapeutic relationship is the main mechanism [20], whereas the other extreme argues that therapeutic techniques or procedures are the main mechanism [21]. At times, these 2 stances are reflected in therapeutic approaches, such that psychodynamic (ie, Neo-Freudian) approaches tend to emphasize the therapeutic relationship whereas cognitive behavioral approaches tend to emphasize technique. In reality, most psychotherapies attempt to integrate some combination of the 2 and allow them to build on one another; there are techniques used to form the relationship, and relationships facilitate the use of techniques. Furthermore, a given act can be seen as both relationship-building and the administration of a technique. Essentially, the therapeutic process often demands that the therapist engage in a comparative analysis of emotional experiences with the patient, thus exercising some form of affective empathy, though it is possible that these affective components of empathy are more crucial in some therapeutic interventions than in others.

Despite the variance in therapeutic orientations, there is broad agreement that one of the fundamental transtheoretical elements critical for a successful outcome is the treatment's working alliance [22]. Conventionally, the alliance is measured across 3 domains: agreement on therapeutic goals, agreement on the therapeutic activities needed to achieve these goals, and the warmth and genuineness of the connection. The working alliance is a significant predictor of treatment outcomes across all forms of psychotherapy [23,24]. Motivational empathy is intimately connected to a fruitful working alliance, particularly the aspect concerning warmth and care [25]. Such results were reported in early studies of cognitive behavioral therapy, where warmth was a predictor of symptom improvement [26], though subsequent findings have been equivocal [21].

One of the provocative findings regarding the alliance of client and therapist is that it appears to be similarly related to reported outcomes in face-to-face psychotherapy and in internet-based interventions (IBIs), which include asynchronous communications [24] with minimal therapist contact. Within the field of IBI, the presence of therapist support is predictive of more symptom improvement, less dropout, and greater adherence in comparison with unguided IBI [27,28] (though data are not conclusive [29]). Furthermore, the relationship with the internet-based program has been found to be predictive of symptom reduction, whereas the relationship with the therapist was predictive of adherence and dropout in a therapist-guided IBI [30]. These differences raise the possibility that although patients can potentially form a relationship with a digital interface, such a relationship differs in its benefits in comparison to a relationship with a therapist, and may not be able to create the same profound alliance and conversations with patients [31]. In other words, conversational AI and digital programs can be helpful in psychoeducation and administering practical techniques that alleviate symptoms, but they are unlikely to build a motivational relationship in the same way as a human therapist and client. Moreover, it has been claimed that even if conversational AI were to be used only for practical purposes alongside a human therapist, this may change and interact with the human dynamic in therapy, especially with the therapeutic alliance [32], which may possibly change how one experiences empathy throughout the therapeutic process.

A Patient-Needs Perspective

The previous section examined the importance of empathy in therapy, from the perspective of therapeutic treatment approaches, which could be thought of as a continuum in terms of their theoretical mechanisms from skill acquisition (eg, cognitive behavioral therapy) to relationship-based (eg, interpersonal psychoanalysis). An alternate perspective considers a patient's specific motivations and needs for seeking treatment, regardless of the therapist's theoretical orientation. We contend that patients enter therapy with a variety of needs, motivations, and expectations, which can be conceptualized along two orthogonal continua including (1) a desire to acquire practical tools and (2) a desire for human connection and empathy. The relative emphasis on each dimension varies among patients according to their theory of change (ie, what they need to decrease distress or improve quality of life) and many other factors (such as their history of successful or unsuccessful treatments, stigma, culture, etc). In addition to individual variability, we contend that individuals can change in their theory and emphasis during therapy as a result of their cumulative experiences over the course of therapy. The more a patient seeks to acquire strategies to cope, the more conversational AI might be able to facilitate this process by providing psychoeducation, exercises, and the like. Conversely, the greater the patient's need for human connection and empathy—be it for affirmation, a confidant for their thoughts and feelings, or simply the sense that someone cares—the less capable conversational AI might be in fulfilling these requirements (Figure 1). We note that although we see these as 2 dimensions, we would not expect many individuals who seek treatment to be low on both, as there would be no motivation to seek treatment (aside from appeasing others).

Adding to this complexity is the patient's self-awareness and accuracy regarding their needs, which is often ambiguous (eg, they may believe they need practical tools when, in reality, they require empathic care more, or vice versa). It is conceivable that a patient might need coping skills for personal growth; yet, the most effective means of acquiring these tools could be through engagement with a compassionate therapist. Such a process may take time and require working on the therapeutic relationship, and needs may change during this period. We believe that when thinking of how to use AI systems in mental health contexts, we must not seek to use them as a "quick fix" solution to a specific problem at all times, as in some cases, a long-term human connection may be required to provide more thorough help.

Moreover, even in therapeutic interactions that are not complete courses of psychotherapy and are comprised of short, concrete interventions, such as crisis helplines, people still may need a human connection. Indeed, research shows that the main reason people call crisis helplines is to have someone to talk to [33]. While conversational AI systems can, at times, give the feeling of "being heard," their responses have still been reported to have less value than those written by humans [13]. One could assume that repeat callers [34] to such helplines seek human connection and will not find it sufficient to only communicate with a bot. A bot-only approach may lead to deterioration in their condition, a risk that should be minimized, especially if a person is going through a crisis. Additionally, other users develop an overdependency on the AI tool, and interventions should be planned in a manner that would mitigate the risk of a long-term dependency. It is possible that AI-assisted communication, including with a therapist, could help deal with both of these risks.

We bring this population as an example, but these considerations pose open-ended questions that warrant exploration in the burgeoning domain of AI-mediated therapy broadly, where the interplay between human touch and technological aid is continually redefined.

Figure 1. An illustration of different patients' possible needs and their possible benefit from AI intervention or a human connection in therapy. AI: artificial intelligence.

Patient Perceptions of AI Bots

In the growing literature examining individuals' perceptions of AI bots, different factors have been shown to influence the levels of trust and bonding created with AI [35-38]. Much of this literature can be viewed from the same 2 axes: how helpful and capable the bot is in providing appropriate tools and results, and how empathic it is (in terms of understanding the user's needs). One study showed that a therapeutic bot was rated on the dimensions of alliance at similar levels to face-to-face therapists [39]. In another study, responses generated by ChatGPT were rated as more authentic, professional, and practical [40]; however, participants were blind to the fact that responses were generated by conversational AI. When participants are aware of AI systems' involvement or simply believe it is involved, responses can seem less authentic and less trustworthy and raise negative emotions [41-43]. Such findings require further research to determine whether responses can truly be experienced as empathic when one knows their artificial origin and whether such experiences differ in their relationship to treatment outcomes according to patients' theory of change.

Conclusions

The advent of advanced AI technologies offers substantial benefits and potential enhancements to therapeutic practices, as well as greater accessibility for a wider population. Nevertheless, certain junctures within the therapeutic process may be particularly sensitive to the need for human rapport. We suggest that those points, which may be whole treatments or specific sessions, are times when empathy is especially needed. Although conversational AI can adeptly simulate empathic interactions, sometimes creating the impression of empathy surpassing human capability [40], the essence of seeking empathy transcends the mere reception of an ideal empathic response. It encompasses a longing for the genuine care and emotional engagement of the individual offering support. The optimal path forward may lie in designing applications that facilitate therapist-AI partnerships, wherein AI systems could augment various facets of therapy—from initial intake and evaluation to, in certain instances, complete treatment modalities—while also consciously addressing the need for authentic human empathy, compassion, and care, when relevant for treatment success. However, most of our proposal is theoretical, and ultimately, we raise an empirical question that should be evaluated in future studies. We also encourage industry professionals developing AI applications for mental health and those conducting research within this domain to remain mindful of these considerations.

Conflicts of Interest

None declared.

References

1. Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. Commun ACM 1966;9(1):36-45. [doi: 10.1145/365153.365168]
2. Sullivan Y, Nyawa S, Fosso WS. Combating loneliness with artificial intelligence: an AI-based emotional support model. 2023 Jan 03 Presented at: Proceedings of the 56th Hawaii International Conference on System Sciences; January 3, 2023; Hawaii URL: https://hdl.handle.net/10125/103173 [doi: 10.24251/hicss.2023.541]
3. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry 2019;64(7):456-464 [FREE Full text] [doi: 10.1177/0706743719828977] [Medline: 30897957]
4. Ly KH, Ly AM, Andersson G. A fully automated conversational agent for promoting mental well-being: a pilot RCT using mixed methods. Internet Interv 2017;10:39-46 [FREE Full text] [doi: 10.1016/j.invent.2017.10.002] [Medline: 30135751]
5. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR Mhealth Uhealth 2018;6(11):e12106. [doi: 10.2196/12106] [Medline: 30470676]
6. Inkster B, Kadaba M, Subramanian V. Understanding the impact of an AI-enabled conversational agent mobile app on users' mental health and wellbeing with a self-reported maternal event: a mixed method real-world data mHealth study. Front Glob Womens Health 2023;4:1084302 [FREE Full text] [doi: 10.3389/fgwh.2023.1084302] [Medline: 37332481]
7. Funk B, Sadeh-Sharvit S, Fitzsimmons-Craft EE, Trockel MT, Monterubio GE, Goel NJ, et al. A framework for applying natural language processing in digital health interventions. J Med Internet Res 2020;22(2):e13855 [FREE Full text] [doi: 10.2196/13855] [Medline: 32130118]
8. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589-596. [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
9. Zaki J, Ochsner KN. The neuroscience of empathy: progress, pitfalls and promise. Nat Neurosci 2012;15(5):675-680. [doi: 10.1038/nn.3085] [Medline: 22504346]
10. Saxena A, Khanna A, Gupta D. Emotion recognition and detection methods: a comprehensive survey. J Artif Intell Syst 2020;2(1):53-79. [doi: 10.33969/ais.2020.21005]
11. Shukla A. Utilizing AI and machine learning for human emotional analysis through speech-to-text engine data conversion. J Artif Int Cloud Comp 2022:1-4. [doi: 10.47363/jaicc/2022(1)145]
12. Chung LL, Kang J. "I'm Hurt Too": the effect of a chatbot's reciprocal self-disclosures on users' painful experiences. Arch Des Res 2023;36(4):67-85. [doi: 10.15187/adr.2023.11.36.4.67]
13. Yin Y, Jia N, Wakslak CJ. AI can help people feel heard, but an AI label diminishes this impact. Proc Natl Acad Sci U S A 2024;121(14):e2319112121. [doi: 10.1073/pnas.2319112121] [Medline: 38551835]
14. Fuchs T. Empathy, group identity, and the mechanisms of exclusion: an investigation into the limits of empathy. Topoi 2017;38(1):239-250. [doi: 10.1007/s11245-017-9499-z]
15. Cameron CD, Hutcherson CA, Ferguson AM, Scheffer JA, Hadjiandreou E, Inzlicht M. Empathy is hard work: people choose to avoid empathy because of its cognitive costs. J Exp Psychol Gen 2019 Jun;148(6):962-976. [doi: 10.1037/xge0000595] [Medline: 30998038]
16. Ferguson AM, Cameron CD, Inzlicht M. Motivational effects on empathic choices. J Exp Soc Psychol 2020;90:104010. [doi: 10.1016/j.jesp.2020.104010]
17. Perry A. AI will never convey the essence of human empathy. Nat Hum Behav 2023;7(11):1808-1809. [doi: 10.1038/s41562-023-01675-w] [Medline: 37474839]
18. Elliott R, Bohart AC, Watson JC, Murphy D. Therapist empathy and client outcome: an updated meta-analysis. Psychotherapy (Chic) 2018 Dec;55(4):399-410 [FREE Full text] [doi: 10.1037/pst0000175] [Medline: 30335453]
19. Rogers CR. A Way of Being. Boston: Houghton Mifflin; 1980.
20. Wampold B, Imel ZE. The Great Psychotherapy Debate: The Evidence for What Makes Psychotherapy Work. New York: Routledge; 2015.
21. DeRubeis RJ, Feeley M. Determinants of change in cognitive therapy for depression. Cogn Ther Res 1990 Oct;14(5):469-482. [doi: 10.1007/bf01172968]
22. Bordin ES. The generalizability of the psychoanalytic concept of the working alliance. Psychother Theory Res Pract 1979;16(3):252-260. [doi: 10.1037/h0085885]
23. Del Re AC, Flückiger C, Horvath AO, Wampold BE. Examining therapist effects in the alliance-outcome relationship: a multilevel meta-analysis. J Consult Clin Psychol 2021;89(5):371-378. [doi: 10.1037/ccp0000637] [Medline: 33829817]
24. Flückiger C, Del Re AC, Wampold BE, Horvath AO. The alliance in adult psychotherapy: a meta-analytic synthesis. Psychotherapy (Chic) 2018 Dec;55(4):316-340 [FREE Full text] [doi: 10.1037/pst0000172] [Medline: 29792475]
25. Nienhuis JB, Owen J, Valentine JC, Winkeljohn Black S, Halford TC, Parazak SE, et al. Therapeutic alliance, empathy, and genuineness in individual adult psychotherapy: a meta-analytic review. Psychother Res 2018;28(4):593-605. [doi: 10.1080/10503307.2016.1204023] [Medline: 27389666]

26. Persons JB, Burns DD. Mechanisms of action of cognitive therapy: the relative contributions of technical and interpersonal interventions. Cogn Ther Res 1985;9(5):539-551 [FREE Full text] [doi: 10.1007/BF01173007]
27. Koelen J, Vonk A, Klein A, de Koning L, Vonk P, de Vet S, et al. Man vs. machine: a meta-analysis on the added value of human support in text-based internet treatments ("e-therapy") for mental disorders. Clin Psychol Rev 2022;96:102179 [FREE Full text] [doi: 10.1016/j.cpr.2022.102179] [Medline: 35763975]
28. Musiat P, Johnson C, Atkinson M, Wilksch S, Wade T. Impact of guidance on intervention adherence in computerised interventions for mental health problems: a meta-analysis. Psychol Med 2022;52(2):229-240. [doi: 10.1017/S0033291721004621] [Medline: 34802474]
29. Werntz A, Amado S, Jasman M, Ervin A, Rhodes JE. Providing human support for the use of digital mental health interventions: systematic meta-review. J Med Internet Res 2023;25:e42864 [FREE Full text] [doi: 10.2196/42864] [Medline: 36745497]
30. Zalaznik D, Strauss AY, Halaj A, Barzilay S, Fradkin I, Katz BA, et al. Patient alliance with the program predicts treatment outcomes whereas alliance with the therapist predicts adherence in internet-based therapy for panic disorder. Psychother Res 2021;31(8):1022-1035. [doi: 10.1080/10503307.2021.1882712] [Medline: 33567994]
31. Sedlakova J, Trachsel M. Conversational artificial intelligence in psychotherapy: a new therapeutic tool or agent? Am J Bioeth 2023;23(5):4-13 [FREE Full text] [doi: 10.1080/15265161.2022.2048739] [Medline: 35362368]
32. Asman O, Tal A, Barilan YM. Conversational artificial intelligence-patient alliance Turing test and the search for authenticity. Am J Bioeth 2023;23(5):62-64. [doi: 10.1080/15265161.2023.2191046] [Medline: 37130413]
33. Middleton A, Woodward A, Gunn J, Bassilios B, Pirkis J. How do frequent users of crisis helplines differ from other users regarding their reasons for calling? Results from a survey with callers to Lifeline, Australia's national crisis helpline service. Health Soc Care Community 2017;25(3):1041-1049. [doi: 10.1111/hsc.12404] [Medline: 27862572]
34. Mishara BL, Côté LP, Dargis L. Systematic review of research and interventions with frequent callers to suicide prevention helplines and crisis centers. Crisis 2023;44(2):154-167 [FREE Full text] [doi: 10.1027/0227-5910/a000838] [Medline: 35086356]
35. Glikson E, Woolley AW. Human trust in artificial intelligence: review of empirical research. Acad Manag Ann 2020;14(2):627-660. [doi: 10.5465/annals.2018.0057]
36. Gould DJ, Dowsey MM, Glanville-Hearst M, Spelman T, Bailey JA, Choong PFM, et al. Patients' views on AI for risk prediction in shared decision-making for knee replacement surgery: qualitative interview study. J Med Internet Res 2023;25:e43632 [FREE Full text] [doi: 10.2196/43632] [Medline: 37721797]
37. Lucas G, Rizzo A, Gratch J, Scherer S, Stratou G, Boberg J, et al. Reporting mental health symptoms: breaking down barriers to care with virtual human interviewers. Front Robot AI 2017 Oct 12;4:51 [FREE Full text] [doi: 10.3389/frobt.2017.00051]
38. Minerva F, Giubilini A. Is AI the future of mental healthcare? Topoi 2023;42(3):809-817. [doi: 10.1007/s11245-023-09932-3] [Medline: 37361723]
39. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. Front Digit Health 2022;4:847991 [FREE Full text] [doi: 10.3389/fdgth.2022.847991] [Medline: 35480848]
40. Lopes E, Jain G, Carlbring P, Pareek S. Talking mental health: a battle of wits between humans and AI. J Technol Behav Sci 2023 Nov 08:1-11. [doi: 10.1007/s41347-023-00359-6]
41. Hohenstein J, Kizilcec RF, DiFranzo D, Aghajari Z, Mieczkowski H, Levy K, et al. Artificial intelligence in communication impacts language and social relationships. Sci Rep 2023;13(1):5487 [FREE Full text] [doi: 10.1038/s41598-023-30938-9] [Medline: 37015964]
42. Promberger M, Baron J. Do patients trust computers? Behav Decis Making 2006;19(5):455-468. [doi: 10.1002/bdm.542]
43. Rob M. We provided mental health support to about 4,000 people—using GPT-3. Here's what happened. X (formerly Twitter). URL: https://twitter.com/RobertRMorris/status/1611450197707464706 [accessed 2023-09-14]

Abbreviations

AI: artificial intelligence
IBI: internet-based intervention

Edited by O Asman; submitted 18.01.24; peer-reviewed by D Meyer; comments to author 08.04.24; revised version received 18.04.24; accepted 23.04.24; published 11.06.24.

Please cite as:
Rubin M, Arnon H, Huppert JD, Perry A
Considering the Role of Human Empathy in AI-Driven Therapy
JMIR Ment Health 2024;11:e56529
URL: https://mental.jmir.org/2024/1/e56529
doi:10.2196/56529
PMID:38861302

Viewpoint

The Machine Speaks: Conversational AI and the Importance of Effort to Relationships of Meaning

Anna Hartford[1], PhD; Dan J Stein[2], PhD, DPhil

[1]Neuroscience Institute, University of Cape Town, Cape Town, South Africa

[2]South African Medical Research Council Unit on Risk and Resilience in Mental Disorders, Department of Psychiatry, University of Cape Town, Cape Town, South Africa

Corresponding Author:
Anna Hartford, PhD
Neuroscience Institute
University of Cape Town
Groote Schuur Hospital, Observatory
Cape Town, 7935
South Africa
Phone: 27 214042174
Email: annahartford@gmail.com

Abstract

The focus of debates about conversational artificial intelligence (CAI) has largely been on social and ethical concerns that arise when we speak to machines—what is gained and what is lost when we replace our human interlocutors, including our human therapists, with AI. In this viewpoint, we focus instead on a distinct and growing phenomenon: letting machines speak for us. What is at stake when we replace our own efforts at interpersonal engagement with CAI? The purpose of these technologies is, in part, to remove effort, but effort has enormous value, and in some cases, even intrinsic value. This is true in many realms, but especially in interpersonal relationships. To make an effort for someone, irrespective of what that effort amounts to, often conveys value and meaning in itself. We elaborate on the meaning, worth, and significance that may be lost when we relinquish effort in our interpersonal engagements as well as on the opportunities for self-understanding and growth that we may forsake.

(JMIR Ment Health 2024;11:e53203) doi:10.2196/53203

KEYWORDS

artificial intelligence; AI; conversational AIs; generative AI; intimacy; human-machine interaction; interpersonal relationships; effort; psychotherapy; conversation

Introduction

Conversation is central to our shared humanity. It is the means through which we make ourselves knowable to another and come to know them in turn. Our mental states—our beliefs, feelings, intentions, desires, and attitudes—are in some respects unreachable by another and sometimes even opaque to ourselves. However, in conversation, we render them articulable, and therefore, accessible. Not unrelatedly, in these exchanges, we often learn about ourselves as well as the other person. The recent emergence of powerful conversational artificial intelligences (CAIs) has therefore been unsettling on various levels (far more so than equally powerful AIs that operate in mediums besides conversation). In their extraordinary replication of the means through which we express our mental states, it is tempting to impute these states to our AI interlocutors. After all, the articulation of thinking (or feeling, hoping, willing, and desiring) is usually all the evidence we require to attribute the relevant mental states to someone.

In her book, Reclaiming Conversation: The Power of Talk in a Digital Age, Sherry Turkle [1] endeavors to make the case for conversation in a world that has increasingly abandoned it for the conveniences (and safeties) of mere digital connection. "At a first, we speak through machines and forget how essential face-to-face conversation is to our relationships, our creativity, and our capacity for empathy," Turkle writes. "At a second, we take a further step and speak not just through machines but to machines. This is the turning point" [1]. This concern was prescient, and Turkle has more recently elaborated on it with reference to the proliferation of CAIs or social chatbots, such as Xiaoice, Woebot, or Replika. These CAIs aim to provide intimacy, but of what sort? Turkle suggests that this intimacy is necessarily fraudulent since it is (by design) devoid of the emotional vulnerability crucial to genuine intimacy [2].

Similarly, these CAIs eliminate the demands and challenges of empathy required for genuine interpersonal exchanges [1,3]. These arguments align with Turkle's long-standing critique of how computers affect our relationships with ourselves and with others [3-5].

"Speaking to Machines": CAIs and the Possibility of Insight

There is ongoing debate concerning the type and quality of conversations possible with CAIs and their appropriateness in therapeutic contexts. In psychotherapy, the digitization of many processes may suggest that CAIs can simply replace the therapist. However, it is also possible to argue that the psychotherapeutic relationship and the experience of that relationship are what is most crucial. In psychodynamic psychotherapy, the client experiences transference while the therapist experiences counter-transference, and working through these processes leads to therapeutic change. In frameworks more influenced by cognitive-behavioral principles, such as schema therapy, the therapist may play a key role in providing "reparenting," a process that leads to positive outcomes.

Ethical concerns with CAIs in therapeutic contexts include the biases and other harmful prompts that might arise in such exchanges, along with the potential dearth of responsibility and accountability for these harms [6-10]. However, even if such patent ethical concerns were addressed or eradicated, central questions would persist: what sort of presence or entity do we have in CAIs? [11-14] and perhaps relatedly, is it possible for CAIs to facilitate genuine self-knowledge, self-understanding, and insight in their human interlocutors?

Some have suggested that engagements with CAIs are necessarily deficient in this crucial respect, especially if we consider the practice of joint attention, as well as other forms of mutual recognition and acknowledgment, to be central to the therapeutic conversation (and indeed to conversation more generally) [11,14-16]. Relatedly, there are concerns about the lost mutuality of these exchanges [16,17]. Conversations with bots do not demand that we empathize with or accommodate another, since, in an important sense, there is no one else there [14,16,18]. As Andrew McStay [18] points out, much depends on the account of empathy we are assuming. McStay argues that accounts that are more accommodative of CAIs are "deficient and potentially dangerous" insofar as they lack interdependence, copresence, and particularly moral responsibility [18].

However, others disagree with these characterizations and see no reason why CAIs cannot encourage genuine introspection [19-21]. What is required in the therapeutic exchange—according to some of these proponents—is not necessarily mutual agency, but rather the experience of being emotionally supported and encouraged to engage in self-reflection [20,21]. To necessitate another subjectivity, or the presence of another full-fledged agent, is to presuppose the illegitimacy of CAIs in these contexts, and to needlessly curtail the possibilities of what qualifies as a genuine therapeutic conversation. After all, human therapists regularly fail to generate the conditions for self-understanding and insight, irrespective of their full-fledged agency [19].

We find these counterarguments compelling. Furthermore, if therapeutic benefits are possible through CAIs—as some research suggests [22-24] (although far more investigation is required [25])—then we potentially have a powerful tool in therapeutic CAIs. Given the immense shortfall in mental health care globally [26,27] and the often prohibitive cost of undertaking conventional psychotherapy, we would be remiss to hastily disregard the beneficial possibilities of therapeutic CAIs. Moreover, certain individuals and populations might experience unique benefits from the format of engagement required by therapeutic exchanges with CAIs, and (relatedly) may not experience the particular advantages of in-person conversation highlighted by advocates such as Turkle (this point has been made with regard to children on the autism spectrum in particular [28,29]).

Being Spoken for: CAIs and Surrendering Articulation

In recent reckonings with the rise of CAIs, the focus has generally been on concerns like those outlined above: what becomes of us when we increasingly replace our human interlocutors—including our human therapists—with AIs, that is (in Turkle's phrase) "when we speak not just through machines but to machines" [1].

Our central concern in this viewpoint, however, is different. Although certain dimensions of the preceding debate are of relevance to our position, we can also remain agnostic with regard to the value of "speaking to machines," whether in a therapeutic context or otherwise. We can remain open to the possibility that bot and human engagements can generate genuine depth, worth, and meaning. Furthermore, we need not presume that the conditions for the emergence of genuine self-understanding and self-reflection cannot be generated in interactions with CAIs. Rather, our concern arises independently, for we now seem to have reached another turning point, and one that extends even further. "At a third point," we might add to Turkle's list, "we take yet another step and let machines speak for us" [1].

We will concentrate on the significance of these forces to our self-knowledge and our interpersonal relationships, although more could be, and has been, said about their implications more generally, for example, concerning achievement gaps (where automation threatens to undermine genuine achievement, and therefore, meaningful work [30]) and responsibility gaps (where automation threatens to undermine responsibility for harmful outcomes [31,32]).

Our central concern will be the following: what is potentially at stake, personally and interpersonally, when we let the machine speak for us? We will explore this question within the framework of philosophical and ethical debates concerning the interpersonal value of effort, rather than exploring it qualitatively or quantitatively (although further empirical research on these questions would be valuable).

This third transition can take many forms, some seemingly more trivial than others. When we are writing an email and the remainder of the sentence auto-fills in gray, we are tempted to stop speaking for ourselves and let the machine speak for us instead.

At times, the costs of this surrender may seem slight, if they exist at all. What does it matter if you articulate some rote phrase to a distant work acquaintance or have it articulated for you instead? However, in other circumstances and other relationships, even these subtle interventions can carry weight.

In an early exploration of the implications of large language models—written in 2019, before the mass rollout of ChatGPT and other large language models—the journalist John Seabrook [33] wrote the following about the experience of using Smart Compose to autocomplete his emails:

> *Finally, I crossed my Rubicon. The sentence itself was a pedestrian affair. Typing an e-mail to my son, I began "I am p—" and was about to write "pleased" when predictive text suggested "proud of you." I am proud of you. Wow, I don't say that enough. And clearly Smart Compose thinks that's what most fathers in my state say to their sons in e-mails. I hit Tab. No biggie.*
>
> *And yet, sitting there at the keyboard, I could feel the uncanny valley prickling my neck.*

Nowadays, the modes in which the machine can speak for us have expanded enormously from these first modest iterations. There are many examples to consider, and many more are developing as we write, but the ways through which we can outsource the labor of our interpersonal articulations are currently expanding exponentially.

Take one example: it is now possible to get CAIs to message on your behalf on dating apps. A variety of start-ups have generated different tools that allow you to hand over your messages to an AI [34]. Instead of having to initiate a conversation with a prospective date—or to come up with thoughtful or witty replies to their messages—AI will do it for you.

When you do not care for the people you are messaging, this option offers a certain pragmatic appeal (especially given the volume of messaging that contemporary dating apps necessitate). However, when you *do* care about a person, the temptation might be even stronger. The CAI, after all, always has an idea of what to say next, and moreover, it offers a version of what you *should* say—a statistically probable representation of what *people like you* say at *times like this*. In comparison, speaking for yourself can feel risky. The things you might say on your own—the way in which you try to make yourself known and get to know others—might be odd, off-putting, or *wrong* somehow.

Take another example: in June 2023, *The New York Times* reported that some doctors were turning to AI to communicate compassionately with patients [35]. We have all experienced the sense of inadequacy that comes with trying to say something supportive to someone who is in an awful circumstance. At such times, we can cast around for ages and summon nothing but cliches. How alluring it is to have a ready-made response instead, and one so well trained in the performance of genuine feeling. The AI's messages will be, in many cases, much better than what we could have produced on our own—kinder, more thoughtful, and more encouraging. Yet no matter how superbly it manages to express care and compassion, this expression is of course divorced from any genuine experience of care and compassion. We should be cautious, in our expedient outsourcing of this emotional connection and engagement, of when we begin to divorce ourselves from the genuine experience of care and compassion along with it.

When we are struggling to find the right thing to say, it may feel like we are achieving nothing. Yet it is precisely in these times—as we try to understand what someone else is enduring, to feel for them, and to express that feeling—that we are undertaking the genuine experience of care and compassion, without which the words themselves are hollow.

One optimistic response is that we might learn more empathetic engagement from the example of the machines. However, this seems unlikely. It is like suggesting that we will improve our spelling skills by relying on automated spell-check or that we will remember more phone numbers through the excellent example set by our phones. Of course, we will not, as the process removes effort, and little of importance has ever been learned without effort.

Thinking ahead—and not necessarily too far ahead—it is possible to see how the temptation to let the machine speak might overspill our text-based conversations. The push to normalize mixed-reality engagements—most notably with the launch of Apple's Vision Pro headset last year—would make it possible for the machine to take over not only our text-based correspondence but also our face-to-face conversations.

We are, right now, at the initial stages of the temptation to begin ceding our expressions to CAIs. However, with little imagination, it is easy to see all the ways in which these temptations are poised to grow. After all, if it was largely the machine whose messages charmed someone into going on a date with you in the first place, how enticing would it be to let the machine keep on speaking when you have to go on the date yourself? The machine speaks with such authority, and as our confidence in its utterances grows, our confidence in our own could correspondingly diminish.

To our mind, the potential costs (to one's own humanity and to our shared humanity) of CAIs are greatest when we allow them to speak for us. Genuine conversation nurtures authentic engagement with others and a better understanding of ourselves. Turkle [1] emphasizes what is lost when we speak through machines, and further still, when we speak *to* machines, but there is, even in these latter engagements, the possibility of coming to know our own thoughts and feelings, of having to search for, and to find, the expression for our experience, and recognizing that the experience precedes the expression that follows.

However, when we allow the machine to speak for us, even this possibility diminishes. We can too easily avoid the effort it takes to genuinely understand ourselves and our unique circumstances (undertakings that are not necessarily discouraged by speaking *to* machines [20,21,29]). We are not encouraged to find the expression for our experience. Instead, we can too easily mistake whichever expressions we receive for our own experience, scarcely recognizing what we have lost in the exchange.

Effort and Meaning

The purpose of these technologies is, in no small part, to remove effort. To take something that once required a great deal from us and make it require little to nothing. Effort is by definition a burden, and in any given instance of having to exert effort, we are always wishing there was a way to be rid of it, but effort also has enormous value, and in some cases, even intrinsic value. This can be true in many realms—there are crucial senses in which "achievement" itself is impossible without effort [30,36]—but it is especially true in our interpersonal relationships. In some interpretations, effort allows us to *reveal* our care and concern for one another and make it knowable. In such interpretations, its role is primarily epistemic. This epistemic role is not trivial in itself, but there are also interpretations whereby effort is more significant still—instead of only allowing us to *reveal* care and concern, it may also *generate* this care and concern [37,38]. Imagine a husband who lovingly cares for his wife through a long illness. His devotion through this ordeal might not only reveal the depths of his love for his wife, but it could also *generate* those depths.

In this sense, the exertion of effort might have both generative and revelatory value in our interpersonal relationships, and the relinquishment of effort might have serious costs on both fronts. To make an effort for someone, irrespective of what that effort amounts to, conveys value and meaning in itself. Many of our interpersonal practices are ways of trying to make real or manifest the effort that is in fact of genuine importance to us. In turn, when effort is removed from these practices, so is their worth.

Take one example: nowadays, Facebook provides automatic reminders of people's birthdays. The moment this memory became automated, the fact of remembering someone's birthday (which used to carry weight and significance) became increasingly meaningless. It is now possible to set up your account to automatically post a rote birthday message on the appropriate day; you need not even give the person a moment's thought. These automated messages are equivalent, in terms of their interpersonal worth, to the automated birthday messages sent by a bank or a mobile service provider. Without requiring any thought or effort, the whole practice loses its significance. What other forms of interaction could we surrender to this fate, as we are increasingly able to opt for the effortless modes of expressing pride, love, affection, or consolation to the people around us?

Conclusions

In turn, we should begin to think carefully (even if just for ourselves) about which of these technologies we choose to use, in different contexts and spheres of our lives, and which ones we do not. Where we choose to use them, we should think equally hard about the *manner* of our engagement and the extent of our agency within it; the more passive we allow ourselves to be, the greater the potential costs we have gestured to in this viewpoint. This is especially true when it comes to those undertakings that have value in and of themselves—rather than value only for their outputs [30]—and also, as we have emphasized in this viewpoint, when it comes to those relationships and human interactions in which our engaged presence, as well as our emotional and intellectual attention and reflection, carries so much significance.

There is an adage in developmental psychology: the toys that are best for children are the ones that require them to do the most work. "The best toys are 90% the kid, 10% the toy," as psychologist Kathy Hirsh-Pasek put it. "If it's 90% the toy, and 10% the kid, that's a problem" [39]. The toys that demand the most of a child are the ones that generate creativity, teach them problem-solving, and encourage their social interactions. On the other hand, the toys that merely require a child to press a button will teach them only to press a button over and over again. When we consider children, we usually show special caution for what will aid and hamper their development, flourishing, and well-being. However, our development does not cease after childhood, and indeed, much of the hardest work (in learning to know as well as relate to ourselves and others) still lies ahead. Given that, we should perhaps pause and wonder what opportunities for self-development we might be forsaking as we embrace, ever more, the toys that want to do everything for us, while we ourselves do less and less. We should remember, in the ceaseless war against effort, that far from needing to be eradicated at every opportunity, there are spheres of our lives—and our interpersonal relationships are a prime example—where effort itself can be the whole point.

Conflicts of Interest

DJS has received consultancy honoraria from Discovery Vitality, Johnson & Johnson, Kanna, L'Oreal, Lundbeck, Orion, Sanofi, Servier, Takeda and Vistagen.

References

1. Turkle S. Reclaiming Conversation: The Power of Talk in a Digital Age. New York, NY: Penguin books; 2015.
2. Turkle S. That chatbot I've loved to hate. MIT Technology Review. 2020. URL: https://www.technologyreview.com/2020/08/18/1006096/that-chatbot-ive-loved-to-hate/ [accessed 2024-05-27]
3. Turkle S. The Empath Diaries: A Memoire. New York, NY: The Penguin Books; 2021.
4. Turkle S. Life on the Screen: Identity in the Age of the Internet. New York, NY: Simon & Schuster; 1995.

5. Turkle S. Alone Together: Why We Expect More From Technology and Less From Each Other?. New York, NY: Basic Books; 2011.
6. Laacke S. Bias and epistemic injustice in conversational AI. Am J Bioeth 2023 May;23(5):46-48. [doi: 10.1080/15265161.2023.2191055] [Medline: 37130400]
7. Kasirzadeh A, Gabriel I. In conversation with artificial intelligence: aligning language models with human values. Philos Technol 2023 Apr 19;36:1-24. [doi: 10.1007/s13347-023-00606-x]
8. Blodgett SL, Lopez G, Olteanu A, Sim R, Wallach H. Stereotyping Norwegian salmon: an inventory of pitfalls in fairness benchmark datasets. 2021 Presented at: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); August 1 - 6; Virtual event.
9. Henderson P, Sinha K, Angelard-Gontier N, Ke NR, Fried G, Lowe R. Ethical challenges in data-driven dialogue systems. 2018 Presented at: AIES '18: AAAI/ACM Conference on AI, Ethics, and Society; February 2 - 3; New Orleans, LA. [doi: 10.1145/3278721.3278777]
10. Welbl J, Glaese A, Uesato J, Dathathri S, Mellor J, Hendricks L, et al. Challenges in detoxifying language models. arXiv Preprint posted online on Sep 15, 2021. [doi: 10.48550/arXiv.2109.07445]
11. Sedlakova J, Trachsel M. Conversational artificial intelligence in psychotherapy: a new therapeutic tool or agent? Am J Bioeth 2023 May;23(5):4-13 [FREE Full text] [doi: 10.1080/15265161.2022.2048739] [Medline: 35362368]
12. Nyholm S. Tools and/or Agents? Reflections on Sedlakova and Trachsel's Discussion of Conversational Artificial Intelligence. Am J Bioeth 2023 May;23(5):17-19. [doi: 10.1080/15265161.2023.2191053] [Medline: 37130387]
13. Floridi L. AI as agency without intelligence: on ChatGPT, large language models, and other generative models. Philos Technol 2023 Mar 10;36(1):1-7. [doi: 10.1007/s13347-023-00621-y]
14. Montemayor C. Language and intelligence. Mind Mach 2021 Aug 08;31(4):471-486. [doi: 10.1007/s11023-021-09568-5]
15. Strijbos D, Jongepier F. Self-knowledge in psychotherapy: adopting a dual perspective on one’s own mental states. PPP 2018;25(1):45-58. [doi: 10.1353/ppp.2018.0008]
16. Wieland LC. Relational reciprocity from conversational artificial intelligence in psychotherapy. Am J Bioeth 2023 May;23(5):35-37. [doi: 10.1080/15265161.2023.2191033] [Medline: 37130399]
17. Friedman C. Ethical concerns with replacing human relations with humanoid robots: an ubuntu perspective. AI Ethics 2022 Jun 20;3(2):527-538. [doi: 10.1007/s43681-022-00186-0]
18. McStay A. Replika in the Metaverse: the moral problem with empathy in 'It from Bit'. AI Ethics 2022 Dec 22:1-13 [FREE Full text] [doi: 10.1007/s43681-022-00252-7] [Medline: 36573214]
19. Grodniewicz JP, Hohol M. Therapeutic conversational artificial intelligence and the acquisition of self-understanding. Am J Bioeth 2023 May;23(5):59-61. [doi: 10.1080/15265161.2023.2191021] [Medline: 37130405]
20. Hurley ME, Lang BH, Smith JN. Therapeutic artificial intelligence: does agential status matter? Am J Bioeth 2023 May;23(5):33-35. [doi: 10.1080/15265161.2023.2191037] [Medline: 37130404]
21. Gray J. Deception mode: how conversational AI can respect patient autonomy. Am J Bioeth 2023 May;23(5):55-57. [doi: 10.1080/15265161.2023.2191023] [Medline: 37130415]
22. Burr C, Floridi L, editors. Ethics of Digital Well-Being: A Multidisciplinary Approach. Cham, Switzerland: Springer; 2020.
23. Rubeis G. E-mental health applications for depression: an evidence-based ethical analysis. Eur Arch Psychiatry Clin Neurosci 2021 Apr;271(3):549-555. [doi: 10.1007/s00406-019-01093-y] [Medline: 31894391]
24. Fiske A, Henningsen P, Buyx A. The implications of embodied artificial intelligence in mental healthcare for digital wellbeing. In: Ethics of Digital Well-Being. Cham, Switzerland: Springer; 2020.
25. Torous J, Cerrato P, Halamka J. Targeting depressive symptoms with technology. mHealth 2019;5:19 [FREE Full text] [doi: 10.21037/mhealth.2019.06.04] [Medline: 31463305]
26. Amram B, Klempner U, Shturman S, Greenbaum D. Therapists or replicants? ethical, legal, and social considerations for using ChatGPT in therapy. Am J Bioeth 2023 May;23(5):40-42. [doi: 10.1080/15265161.2023.2191022] [Medline: 37130418]
27. Mental Health and Substance Use (MSD), World Health Organization. The Mental Health Atlas 2020. Geneva: World Health Organization; Oct 8, 2021.
28. Tartaro A, Cassell J. Playing with virtual peers. Bootstrapping contingent discourse in children with autism. 2008 Presented at: Proceedings of the Eighth International Conference for the Learning Sciences – ICLS 2008, Volumes 2 (pp. 382-389); June 24 - 28; Utrecht, The Netherlands.
29. Tigard DW. Toward relational diversity for AI in psychotherapy. Am J Bioeth 2023 May;23(5):64-66. [doi: 10.1080/15265161.2023.2191047] [Medline: 37130412]
30. Danaher J, Nyholm S. Automation, work and the achievement gap. AI Ethics 2020 Nov 23;1(3):227-237. [doi: 10.1007/s43681-020-00028-x]
31. Matthias A. The responsibility gap: Ascribing responsibility for the actions of learning automata. In: Ethics and Information Technology. Cham, Switzerland: Springer; 2004:A.
32. Nyholm S. Attributing agency to automated systems: reflections on human-robot collaborations and responsibility-Loci. Sci Eng Ethics 2018 Aug;24(4):1201-1219 [FREE Full text] [doi: 10.1007/s11948-017-9943-x] [Medline: 28721641]

33. Seabrook J. The next word. The New Yorker. 2019. URL: https://www.newyorker.com/magazine/2019/10/14/can-a-machine-learn-to-write-for-the-new-yorker [accessed 2024-05-27]
34. Lorenz T. Welcome to the age of automated dating: multiple start-ups are building AI tools for romantic connections? Washington Post. 2023. URL: https://www.washingtonpost.com/technology/2023/04/23/dating-ai-automated-online/ [accessed 2024-05-27]
35. Kolata G. When doctors use a chatbot to improve their bedside manner? The New York Times. 2023. URL: https://tinyurl.com/3p6aasyv [accessed 2024-05-27]
36. Bradford G. Achievement. Oxford, UK: Oxford University Press; 2015.
37. Guerrero A. Intellectual difficulty and moral responsibility. In: Robichaud P, Wieland JW, editors. Responsibility: The Epistemic Condition. Oxford, UK: Oxford University Press; 2017:199-218.
38. Nelkin DK. Difficulty and degrees of moral praiseworthiness and blameworthiness. Nous 2014 Nov 14;50(2):356-378. [doi: 10.1111/nous.12079]
39. Blasdel A. 'They want toys to get their children into Harvard?': have we been getting playthings all wrong? The Guardian. 2022. URL: https://tinyurl.com/3d274nnk [accessed 2024-05-27]

Abbreviations

AI: artificial intelligence
CAI: conversational artificial intelligence

Edited by J Torous; submitted 29.09.23; peer-reviewed by R Marshall, X Zhao; comments to author 11.12.23; revised version received 25.01.24; accepted 26.01.24; published 18.06.24.

Please cite as:
Hartford A, Stein DJ
The Machine Speaks: Conversational AI and the Importance of Effort to Relationships of Meaning
JMIR Ment Health 2024;11:e53203
URL: https://mental.jmir.org/2024/1/e53203
doi:10.2196/53203
PMID:38889401

Original Paper

Exploring the Efficacy of Large Language Models in Summarizing Mental Health Counseling Sessions: Benchmark Study

Prottay Kumar Adhikary[1], BTech; Aseem Srivastava[2], BTech; Shivani Kumar[2], PhD; Salam Michael Singh[1], PhD; Puneet Manuja[3], BTech, MBA; Jini K Gopinath[3], PhD; Vijay Krishnan[4], MD; Swati Kedia Gupta[5], MPhil, PhD; Koushik Sinha Deb[5], MD; Tanmoy Chakraborty[1,6], PhD

[1]Department of Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, India
[2]Department of Computer Science & Engineering, Indraprastha Institute of Information Technology Delhi, New Delhi, India
[3]YourDOST, Karnataka, India
[4]Department of Psychiatry, All India Institute of Medical Sciences, Rishikesh, India
[5]Department of Psychiatry, All India Institute of Medical Sciences, New Delhi, India
[6]Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi, New Delhi, India

Corresponding Author:
Tanmoy Chakraborty, PhD
Department of Electrical Engineering
Indian Institute of Technology Delhi
IIT Delhi Main Road, IIT Campus, Hauz Khas
New Delhi, 110016
India
Phone: 91 26591076 ext 011
Email: tanchak@iitd.ac.in

Abstract

Background: Comprehensive session summaries enable effective continuity in mental health counseling, facilitating informed therapy planning. However, manual summarization presents a significant challenge, diverting experts' attention from the core counseling process. Leveraging advances in automatic summarization to streamline the summarization process addresses this issue because this enables mental health professionals to access concise summaries of lengthy therapy sessions, thereby increasing their efficiency. However, existing approaches often overlook the nuanced intricacies inherent in counseling interactions.

Objective: This study evaluates the effectiveness of state-of-the-art large language models (LLMs) in selectively summarizing various components of therapy sessions through aspect-based summarization, aiming to benchmark their performance.

Methods: We first created Mental Health Counseling-Component–Guided Dialogue Summaries, a benchmarking data set that consists of 191 counseling sessions with summaries focused on 3 distinct counseling components (also known as counseling aspects). Next, we assessed the capabilities of 11 state-of-the-art LLMs in addressing the task of counseling-component–guided summarization. The generated summaries were evaluated quantitatively using standard summarization metrics and verified qualitatively by mental health professionals.

Results: Our findings demonstrated the superior performance of task-specific LLMs such as MentalLlama, Mistral, and MentalBART evaluated using standard quantitative metrics such as Recall-Oriented Understudy for Gisting Evaluation (ROUGE)-1, ROUGE-2, ROUGE-L, and Bidirectional Encoder Representations from Transformers Score across all aspects of the counseling components. Furthermore, expert evaluation revealed that Mistral superseded both MentalLlama and MentalBART across 6 parameters: affective attitude, burden, ethicality, coherence, opportunity costs, and perceived effectiveness. However, these models exhibit a common weakness in terms of room for improvement in the opportunity costs and perceived effectiveness metrics.

Conclusions: While LLMs fine-tuned specifically on mental health domain data display better performance based on automatic evaluation scores, expert assessments indicate that these models are not yet reliable for clinical application. Further refinement and validation are necessary before their implementation in practice.

(JMIR Ment Health 2024;11:e57306) doi:10.2196/57306

KEYWORDS

mental health; counseling summarization; large language models; digital health; artificial intelligence; AI

Introduction

Background

Counseling refers to a relationship between a professional counselor and individuals, families, or other groups that empowers the clients to achieve mental health, wellness, education, and career goals. Specifically, in individuals with psychological or interpersonal difficulties, mental health counseling may be seen as a key helping intervention. Counseling sessions embrace a client-centered approach, fostering an environment of trust and exploration. These sessions delve deep into personal experiences, where clients share intimate details while therapists navigate the dialogue to cultivate a safe and supportive space for healing. Discussions within these sessions span a wide range of topics, from recent life events to profound introspections, all of which contribute to the therapeutic journey. An important aspect of the counseling process lies in the documentation of counseling notes (summary of the entire session), which is essential for summarizing client stressors and therapy principles. Session notes are pivotal in tracking progress and in guiding future sessions. However, capturing the intricacies of these conversations poses a formidable challenge, demanding training, expertise, and experience of mental health professionals. These summaries distill key insights, including symptom and history (SH) exploration, patient discovery (PD), and reflection, while filtering out nonessential details. However, the need for meticulous recordkeeping can sometimes detract from the primary focus of therapy. Maintaining a seamless flow of conversation is paramount in effective therapy, where any disruption can impede progress. To streamline this process and ensure continuity, automation emerges as a promising solution for the counseling summarization task. While advances in artificial intelligence (AI) have revolutionized document summarization, the application of these technologies to mental health counseling remains relatively unexplored.

Previous studies [1-3] have recognized the potential of counseling summarization in optimizing therapeutic outcomes. However, existing models often overlook the unique nuances inherent in mental health interactions. Standard counseling dialogues, using reflective listening, involve identifying current issues; developing a biopsychosocial conceptualization, including past traumas and coping strategies; and chalking out treatment plans. The counseling dialogues also include discussion on between-session issues as well as crises, if any. An effective counseling summary should selectively capture information pertinent to each of these categories while eliminating extraneous details.

Despite the demonstrated capabilities of large language models (LLMs) in various domains, research in mental health counseling summarization is scarce. One major obstacle is the lack of specialized data sets tailored to counseling contexts. To bridge this gap, we embarked on a two-pronged approach: (1) creating a novel counseling-component–guided summarization data set, called Mental Health Counseling-Component–Guided Dialogue Summaries (MentalCLOUDS); and (2) evaluating state-of-the-art LLMs on the task of counseling-component–guided summarization. Through these efforts, we aim to propel the integration of AI technologies into mental health practice, ultimately enhancing the quality and accessibility of therapeutic interventions.

Related Work

Overview

Summarizing counseling conversations enhances session continuity and facilitates the development of comprehensive therapy plans. However, analyzing these interactions manually is an arduous task. To address this challenge, advances in AI and natural language processing, particularly in summarization techniques, offer a promising solution. Summarization tasks can be approached via an extractive [4] or an abstractive [5] viewpoint. Extractive summarization involves identifying the most relevant sentences from an article and systematically organizing them. Given the simplicity of the approach, the resultant extractive summaries are often less fluent. By contrast, abstractive summarization extracts important aspects of a text and generates more coherent summaries. By using summarization, therapists can access recaps of sessions, sparing them the need to sift through lengthy dialogues. While summarization has been a long-studied problem in natural language processing [6], recent attention has shifted toward aspect-based summarization, a method that focuses on generating summaries pivoted on specific points of interest within documents.

Chen and Verma [1] proposed a retrieval-based medical document summarization approach in which the user query is fine-tuned using a medical ontology, but their method is limited due to its overall primitive design. Konovalov et al [7] highlight the importance of identifying emotional reactions and "early counseling" components. Strauss et al [8] used machine learning approaches to automate the analysis of clinical forms, and they envision using machine learning in mental health to a certain extent. Furthermore, research on major depressive disorder [9] underscores the significance of identifying crucial indicators from patient conversations, such as age, anxiety levels, and long episode duration, in the choice of the appropriate level of antidepressant medication, guiding subsequent sessions and prescriptions. Subsequently, the effectiveness of the prescribed antidepressants is monitored to assess the patient's response.

This concept identifies crucial indicators from the patient's conversations with the therapist and guides subsequent follow-up sessions based on the patient's history of interactions and prescriptions. Deep learning approaches, such as the use of recurrent neural networks and long short-term memory, have been used to predict 13 predefined mental illnesses based on neuropsychiatric notes that contain 300 words each, on average, about the patient's present illness and events associated with it, followed by a psychiatric review system that mentions the mental illness related to the patient [10]. Chen et al [11] proposed an extractive summarization approach using the Bidirectional Encoder Representations from Transformers (BERT) model [12] to reduce physicians' efforts in analyzing tedious amounts of diagnosis reports. However, there remains a notable gap in effectively capturing medical information in session summaries.

In addition, some contemporary works used authentic mental health records to create synthetic data sets [13]. Afzal et al [14] reported the summarization of medical documents to identify PICO (Population, Intervention, Comparison, and Outcomes) elements. Manas et al [15] proposed an unsupervised abstractive summarization in which domain knowledge from the Patient Health Questionnaire-9 was used to build knowledge graphs to filter relevant utterances. A 2-step summarization was devised by Zhang et al [16] wherein partial summaries were initially consolidated, and the final summary was generated by fusing these chunks. Furthermore, Zafari and Zulkernine [17] demonstrated a web-based application built using information extraction and annotation tailored to the medical domain.

For dialogue summarization, abstractive summarization has been the de facto standard due to its ability to capture critical points coherently. Nallapati et al [18] used an encoder-decoder–based abstractive summarization method, which was further improved via the attention mechanism [19]. Subsequently, See et al [20] introduced a hybrid approach of extractive and abstractive summarization. Chen and Bansal [2] proposed a reinforcement learning-based approach as a mixture of extractive and abstractive approaches for summarization wherein emphasis is given to redundancy reduction in the utterances extracted from the conversation. Recent research reveals the dependence of specific utterances in the extraction of salient sentences from the conversation utterances. In this regard, Narayan et al [3] analyzed topic distribution based on latent Dirichlet allocation [21]. Subsequently, Song et al [22] segregated utterances into 3 labels: problem description, diagnosis, and other. In medical counseling, Quiroz et al [23] and Krishna et al [24] adopted the method of selecting significant utterances for summarizing medical conversations.

In aspect-based summarization, instead of an overall summary of the entire document, summaries at different aspect levels are made based on specific points of interest. These aspects could be movie reviews [25-28] or summarization guided by different domains [29,30] where the documents or the segments of the documents are tagged with these aspects. Hayashi et al [31] released a benchmarking data set on multidomain aspect-based summarization where they annotated 20 different domains as aspects using the section titles and boundaries of each article chosen from Wikipedia. Frermann et al [29] reported an aspect-based summarization of the news domain. Their method can segment documents by aspect, and the model can generalize from the synthetic data to natural documents. The study further revealed the models' efficacy in summarizing long documents. Recently, aspect-based summarization has garnered considerable traction; however, the data set is limited. Yang et al [32] released a large-scale, high-quality data set on aspect-based summarization from Wikipedia. The data set contains approximately 3.7 million instances covering approximately 1 million aspects sourced from 2 million Wikipedia pages. Apart from releasing the data set, the authors also benchmarked it on the Longformer-Encoder-Decoder [33] model where they performed zero-shot, few-shot, and fine-tuning on 7 downstream domains where data are scarce. Joshi et al [34] address the general summarization of medical dialogues. They proposed combining extractive and abstractive methods that leverage the independent and distinctive local structures formed during a patient's medical history compilation. Liu et al [35] reported a topic-based summarization of general medical domains pertaining to topics such as swelling, headache, chest pain, and dizziness. Their encoder-decoder model tries to generate 1 symptom (topic) at a time. Besides, work on formalizing the conversation text has been reported in the study by Kazi and Kahanda [36]. This work treats the formalization of the case notes from digital transcripts of physician-patient conversations as a summarization task. The method involves 2 steps: prediction of the electronic health record categories and formal text generation. Gundogdu et al [37] used a BERT-based sequence-to-sequence model for summarizing clinical radiology reports. The experimental results indicated that at least 76% of their summary generations were as accurate as those generated by radiologists. There is also a report on topic-guided dialogue summarization for clinical physician-patient conversations [38]. The approach first learns the topic structure of the dialogues and uses these topics to generate the summaries in the desired format (eg, the subjective, objective, assessment, and plan format). Zhang et al [39] proposed a method for factually consistent summarization of clinical dialogues. This method involves extracting factual statements and encoding them into the dialogue. In addition, a dialogue segmenter is trained to segment the dialogues based on topic switching, which enhances the model's overall discourse awareness. Chintagunta et al [40] used GPT-3 [41] to generate training examples for medical dialogue summarization tasks. Recently, there have been reports of LLMs being used in medical dialogue summarization to expedite diagnosis by focusing on relevant medical facts, thereby reducing screening time [42]. The authors conducted benchmarking on GPT-3.5, Bidirectional and Auto-Regressive Transformer (BART) [43], and BERT for Summarization [44]. The study indicated that GPT-3.5 generated more accurate and human-aligned responses than the other 2 models. Another study [45] demonstrated the effectiveness of LLMs in clinical text summarization across 4 different tasks: physician-patient dialogue, radiology reports, patient questions, and progress notes. The quantitative analysis revealed that the summaries generated by the adapted LLMs were comparable, or even superior, in quality to those of the human experts in terms of conciseness, correctness, and completeness. Singh et al [46] used open-source LLMs to extract and summarize suicide ideation indicators from social media texts to expedite mental health interventions.

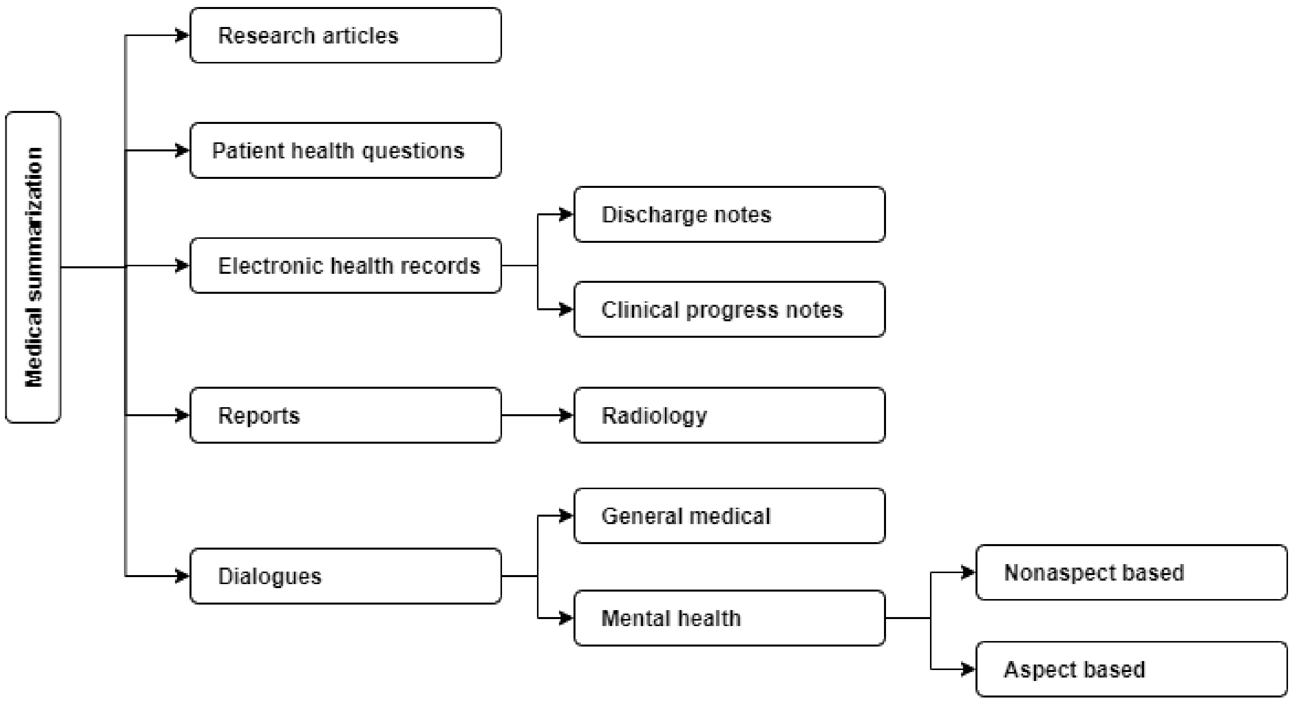
Opportunities

The aforementioned previous works either did not focus on aspect-based summarization or reported on general clinical discussions of common symptoms and conditions (eg, cough, cold, and fever). However, there are still avenues to be explored in the aspect-based summarization of mental health therapy conversations, considering that mental health is a pressing global issue requiring urgent consideration. These therapy conversations encompass several counseling components, including patient information, past symptoms, diagnosis history, reflection, and the therapist's action plans. Focusing the summaries on these counseling components would facilitate targeted and focused summaries, significantly reducing the time and effort and leading to more effective therapy overall. In this direction, our work is motivated by the study conducted by Srivastava et al [47], which reported on a summarization-based counseling technique from therapist-client conversations. They released a conversation data set that is structured with the core components of psychotherapy about SH identification or the discovery of the patient's behavior. The authors proposed an encoder-decoder model based on Text-to-Text Transfer Transformer (T5) [48] for their counseling-component–guided summarization model. However, a single, generic summary is generated in the work, and no focus is given to generating aspect-based summaries. Consequently, we extended the work by using the counseling components, namely SH exploration, PD, and reflection, into an aspect-based summarization framework. To this end, we created MentalCLOUDS, a data set that incorporates summaries aligned with the distinct counseling components. We also explored the efficacy of the state-of-the-art LLMs (encoder-decoder as well as decoder-only models) for the summarization of counseling dialogues in this work.

Taxonomy

On the basis of the survey of related works on summarization in the medical domain in general and in mental health in particular, we present a taxonomy of task formulations for summarization tasks in the medical domain (Figure 1 [11,15,22-24,34,37,39,40,45,47,49-68]). In general, medical text summarization is divided into research articles [49-52], reports, patient health questions, electronic health records, and dialogue summarization. Report summarization encompasses the summarization of reports, such as impressions or summarizations of radiology findings [37,45,53-55]. Patient health question summarization involves summarizing informal, nontechnical, and lengthy patient questions into technically sound and concise ones [56-59]. Electronic health record summarization includes the summarization of patient notes such as clinical progress notes [60-63] and discharge notes [11,53,64-66]. Our work focuses on the abstractive dialogue summarization of mental health counseling conversations, specifically targeting the counseling aspects. In addition, the survey includes general medical dialogue summarization [22-24,34,39,40,45] and mental health dialogue summarization [15,47,67,68]. Of note, this taxonomy does not represent the global scenario but rather provides a comprehensive depiction based on the aforementioned survey.

Figure 1. Taxonomy of summarization methods in the medical domain.



Challenges

Mental health counseling conversations often involve sensitive and confidential information. There is an expectation of empathetic and reflective responses from the therapist and action plans based on which the therapy is conducted. Generative AI–based counselors are susceptible to generating insensitive or incorrect suggestions and lacking empathy in their responses, which can negatively impact the therapy process. Moreover, the components or aspects of counseling sessions are subjective, and a counseling conversation can have multiple aspects. Therefore, the scope of the aspect-based summarization is limited to the specific annotated aspects. However, annotating these aspects requires expert manual intervention, which is costly both in terms of human resources and the financial perspective.

Methods

Overview of the Proposed Data Set: MentalCLOUDS

To evaluate the performance of diverse summarization systems across various aspects of counseling interactions, we expanded upon the Mental Health Summarization (MEMO) data set [47]. Comprising 11,543 utterances extracted from 191 counseling sessions involving therapists and patients, this data set draws from publicly accessible platforms such as YouTube. Embracing a heterogeneous demographic spectrum with distinctive mental health concerns and diverse therapists, the data set facilitates the formulation of a comprehensive and inclusive approach for researchers. Using preprocessed transcriptions derived from counseling videos, the constituent dialogues within the data set exhibit a dyadic structure, exclusively featuring patients and therapists as interlocutors. Within each conversation, 3 pivotal counseling components (aspects) emerge: SH exploration, PD, and reflective utterances.

Our study aims to capture the essence of each aforementioned counseling component, embarking on the creation of 3 distinct summaries for a single dialogue, with each summary tailored to a specific counseling component. Expanding upon the MEMO data set, we augmented it with annotated dialogue summaries corresponding to the 3 identified components. Collaborating closely with a team of leading mental health experts (for their details, refer to the Qualitative Assessment by Experts subsection), we crafted annotation guidelines and subjected the summary annotations to rigorous validation processes. We call the resultant data set MentalCLOUDS. We highlight its key statistics in Table 1 and Figure 2.

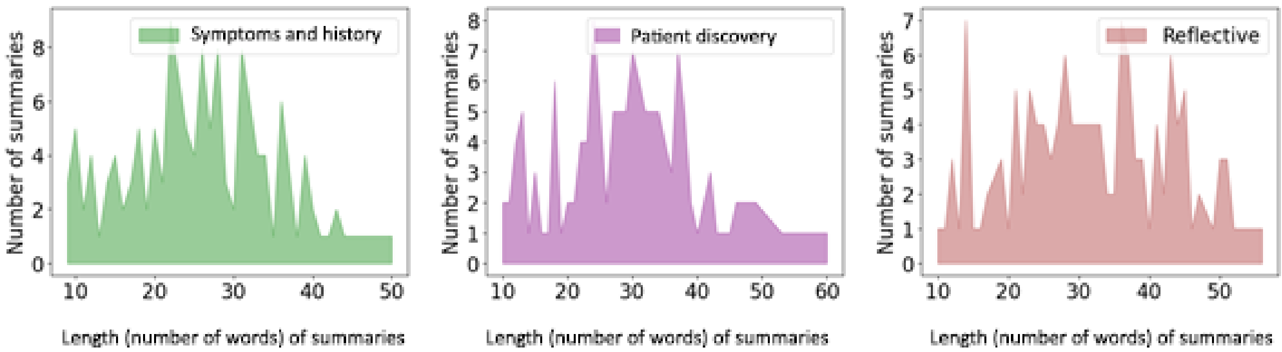
Table 1. Statistics of the Mental Health Counseling-Component–Guided Dialogue Summaries data set.

Set	Dialogues (n=191), n (%)	Utterances (n=11,543), n (%)	Utterances per dialogue, mean (SD)	Patient utterances (n=5722), n (%)	Therapist utterances (n=5814), n (%)	SH[a] utterances (n=2379), n (%)	PD[b] utterances (5428), n (%)	Reflective utterances (n=1242), n (%)
Training	131 (68.59)	8342 (72.3)	63.68 (38.44)	4124 (72.1)	4211 (72.4)	1882 (79.1)	3826 (70.5)	884 (71.2)
Validation	21 (10.99)	1191 (10.3)	56.71 (27.06)	594 (10.4)	597 (10.3)	206 (8.7)	445 (8.2)	146 (11.8)
Test	39 (20.42)	2010 (17.4)	51.53 (39.96)	1004 (17.5)	1006 (17.3)	291 (12.2)	1157 (21.3)	212 (17.1)

[a]SH: symptom and history.

[b]PD: patient discovery.

Figure 2. Distribution of summary lengths in the Mental Health Counseling-Component–Guided Dialogue Summaries (MentalCLOUDS) data set.



Data Annotation Process

Guidelines

Conversations in counseling situations can be challenging, given the sensitive nature of the information shared. A therapist's reflective and open attitude can facilitate this expression. This dynamic is reinforced by the proposed MentalCLOUDS data set. This data set distinguishes the utterances dedicated to symptom exploration, discovering the history of mental health issues and patient behavior, as well as providing insights into past narratives, thereby shaping the patient's present circumstances. These nuanced elements form the core of our identified counseling components. To improve the richness of the data set, we collaborated with mental health experts to formulate a set of annotation guidelines [69]. Furthermore, these guidelines serve as a comprehensive framework by which annotators can focus their attention on particular aspects of the conversation that are essential for producing summaries that are customized for each counseling component. By adhering to these guidelines, the therapeutic techniques are captured in the annotations. This ensures that the resulting summaries are concise yet rich in informative content for the specific component.

Psychotherapy Elements

Within the realm of mental health therapy sessions, distinct counseling components play a pivotal role in facilitating successful interventions. The MentalCLOUDS data set serves as a valuable resource, furnishing meticulously labeled utterances that encompass 3 fine-grained components [47]:

- SH: this facet encapsulates utterances teeming with insightful information crucial for the therapist's nuanced assessment of the patient's situation.

- PD: patients entering counseling sessions often bring intricate thoughts to the fore. Therapists, in turn, endeavor to establish therapeutic connections, creating a conducive environment for patients to articulate and unravel their thoughts. Such utterances by the therapist that encourage patients to reveal their concerns lie in this category.
- Reflecting: therapists use concise utterances, allowing ample space for patients to share their life stories and events. Encouraging patient narratives, therapists may also use hypothetical scenarios to evaluate actions and enhance understanding.

When crafting a summary for a dialogue *D*, aligned with a specific counseling component *C*, our primary focus rests on utterances marked with *C* within *D* in the MEMO data set. Consequently, we derived 3 distinct counseling summaries for each counseling component within a single session to create the MentalCLOUDS data set. Table 1 shows the data statistics, where a balanced distribution of patient and therapist utterances within the data set is evident. Notably, PD emerges as the prevailing label in the data set, highlighting patients' inclination to discuss ancillary topics rather than focusing solely on their mental health concerns when prompted to share their experiences. By contrast, reflecting emerges as the least tagged label in this comprehensive analysis.

Benchmarking

In recent years, the spotlight on LLMs has intensified, captivated by their extraordinary performance across diverse applications. From classification tasks such as emotion recognition [70] to generative problems such as response generation [71], these models have proven their versatility. In this paper, our focus is directed toward evaluating their capability in the domain of counseling summarization, specifically using MentalCLOUDS. In our comprehensive analysis, we leveraged 11 state-of-the-art pretrained LLM architectures, including a mix of general-purpose and specialized models. These models are considered to carefully assess their performance concerning each facet of the counseling-component summaries. We explain each of these systems in Textbox 1.

This is to highlight that all baseline models are transformer based, and computational complexities associated with the transformer-based architectures while being trained or fine-tuned involve a computational cost of $O(L \times N^2 \times D)$, where N represents the sequence length, D denotes the hidden dimension, and L signifies the number of transform layers. As we maintain a constant number of layers across all training steps, the computational complexity simplifies to $O(N^2 \times D)$.

Moreover, our selection of benchmarked models comprises both small language models (SLMs), such as BART, T5, the GPT family, Phi-2, and MentalBART, as well as LLMs such as Flan-T5, Mistral, Llama-2, and MentalLlama. SLMs typically operate within the parameter range of 300 million to 2 billion, whereas LLMs are characterized by a higher parameter count, ranging from 7 billion to 9 billion (as kept in our study). In addition to analyzing the models' complexity for a better understanding of their applicability, another crucial metric to consider is the model's runtime. LLMs tend to consume more runtime due to their larger parameter count, while SLMs run quickly but may compromise accuracy. A comprehensive analysis of the models' runtime is provided in Table 2.

Textbox 1. Description of the 11 models evaluated.

- Bidirectional and Auto-Regressive Transformer (BART) [43]: this is a sequence-to-sequence model designed for various natural language processing (NLP) tasks, including text summarization. It uses a transformer architecture with an encoder-decoder structure. It incorporates a denoising autoencoder objective during pretraining, reconstructing the original input from corrupted versions. We used the pretrained base version of the model in our experiments.
- Text-To-Text Transfer Transformer (T5) [48]: this is a versatile transformer-based model consisting of an encoder-decoder framework with bidirectional transformers. It reframes all NLP tasks as text-to-text tasks, providing a unified approach. T5 learns representations by denoising corrupted input-output pairs. Its encoder captures contextual information while the decoder generates target sequences. The pretrained base version of T5 was used in our experiments.
- GPT-2 [72]: this is a transformer-based language model that comprises a stack of identical layers, each with a multihead self-attention mechanism and position-wise fully connected feed-forward networks. GPT-2 follows an autoregressive training approach, predicting the next token in a sequence given its context.
- GPT-Neo [73]: trained from the Pile data set [74], GPT-Neo exhibits a similar architecture as GPT-2 except for a few modifications, such as the use of local attention in every other layer with a window size of 256 tokens. In addition, GPT-Neo houses a combination of linear attention [75], a mixture of experts [76], and axial positional embedding [77] to achieve performance comparable to that of larger LLMs, such as GPT-3.
- GPT-J [78]: this is a transformer model trained using the methodology proposed by Wang [78]. It is a GPT-2–like causal language model trained on the Pile data set.
- FLAN-T5 [79]: this is the instruction fine-tuned version of the T5 model with a particular focus on scaling the number of tasks, scaling the model size, and fine-tuning on chain-of-thought data.
- Mistral [80]: this is a decoder-based LLM with a sliding-window attention mechanism, where it is trained with an 8k context length and fixed cache size, with a theoretical attention span of 128K tokens. Faster inference and lower cache are ensured by using grouped query attention [81].
- MentalBART [82]: this is an open-source LLM constructed for interpretable mental health analysis with instruction-following capability. The model is fine-tuned using the Interpretable Mental Health Instruction (IMHI) data set [82] and is expected to make complex mental health analyses for various mental health conditions.
- MentalLlama [82]: similar to MentalBART, MentalLlama is the counterpart of the Llama architecture but is trained on the IMHI data set. The model is fine-tuned to integrate the capability of an LLM with domain knowledge in mental health.
- Llama-2 [83]: this is an auto-regressive language model that uses an optimized transformer architecture. The tuned versions use supervised fine tuning [84] and reinforcement learning with human feedback [85] to align with human preferences for helpfulness and safety. The model is trained exclusively on publicly available data sets.
- Phi-2: this is an extension of Phi-1 [86]. Phi-1 is a transformer-based frugal LLM with the largest variant having 1.3 billion parameters. It is trained on textbook-quality data. It emphasizes the quality of the data to compensate for its relatively small number of parameters. Phi-2 has 2.7 billion parameters, which shows comparable performances with other larger LLMs despite its smaller size.

Table 2. Average runtime of models fine-tuned on Mental Health Counseling-Component–Guided Dialogue Summaries (MentalCLOUDS) for summarization tasks across 3 psychotherapy elements: symptom and history, patient discovery, and reflecting.

Model	Variant or parameters	Time (min)	GPU[a]
BART[b]	Base	2.27	A100
T5[c]	Base	18.81	A100
MentalBART	Base	5.94	A100
Flan-T5	Base	16.56	A100
GPT-2	124 million	6.30	A100
GPT-Neo	1.3 billion	32.98	A100
GPT-J	6 billion	44.69	A100
MentalLlama	7	48.27	RTX A6000+RTX A5000
Mistral	7 billion	43.86	RTX A6000+RTX A5000
Phi-2	2.7 billion	9.38	A100

[a]GPU: graphics processing unit.

[b]BART: Bidirectional and Auto-Regressive Transformer.

[c]T5: Text-To-Text Transfer Transformer.

Ethical Considerations

The study did not involve any human subject research; hence, we did not seek ethics approval.

Results

We undertook a comprehensive evaluation of the generated session summaries across various architectures, using a dual approach of quantitative and qualitative assessments.

Quantitative Assessment

Overview

This section reports the aspect-based (psychotherapy element–based) summarization results based on the automatic evaluation scores. Given the generative nature of the task, we used standard summarization evaluation metrics such as Recall-Oriented Understudy for Gisting Evaluation (ROUGE)-1, ROUGE-2, ROUGE-L, and BERT Score (BERTScore) along with their corresponding precision, recall, and F_1-score values. As the F_1-score accounts for precision and recall, we compared the performance of the LLMs based on F_1-score values unless stated otherwise. ROUGE [87] assesses the overlap of n-grams (sequences of n consecutive words) between the generated summary and reference summaries. Specifically, this metric measures the number of overlapping units such as n-grams, word sequences, and word pairs in the generated summary evaluated against the gold summary typically created by humans. ROUGE favors the candidate summary with more overlaps with reference summaries. This effectively gives more weight to matching n-grams occurring in multiple reference summaries. This work reports the unigram and bigram ROUGE (namely ROUGE-1 and ROUGE-2) and ROUGE-L evaluations. ROUGE-L takes into account the longest co-occurring n-gram between the candidate and reference summaries. BERTScore [88] is harnessed to gauge the semantic coherence between the generated summaries and their ground truths. Notably, in the context of counseling summaries, which are inherently tied to a domain-specific conversation, we embarked on a meticulous qualitative examination of the generated summaries for individual counseling components.

SH Summarization

Table 3 reports the automatic evaluation scores of the LLMs on the summarization task for the SH psychotherapy element. MentalLlama outperformed the other LLMs across all automatic evaluation metrics. For the ROUGE-1 metric, MentalLlama achieved an F_1-score of 30.86, followed by MentalBART with an F_1-score of 28.00. In terms of the ROUGE-2 metric, Mistral was comparable to MentalLlama with a difference of just 0.90 in the F_1-score values. Similarly, for the ROUGE-L metric, Mistral was preceded by MentalLlama by a difference of 2.93 in the F_1-score values.

Table 3. Results obtained on Mental Health Counseling-Component–Guided Dialogue Summaries (MentalCLOUDS) for the summarization task on the symptom and history psychotherapy element.

Model	ROUGE[a]-1			ROUGE-2			ROUGE-L			BERTScore[b]		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score	Precision	Recall	F_1-score	Precision	Recall	F_1-score
BART[c]	12.91	28.84	16.26	1.88	5.07	2.47	10.21	23.97	13.19	85.81	85.81	85.81
T5[d]	22.16	19.81	19.74	2.18	1.78	1.85	16.12	14.51	14.36	85.38	85.38	85.38
MentalBART	30.31	29.02	28.00	6.06	5.29	5.46	20.85	20.34	19.40	88.34	88.34	88.34
Flan-T5	21.45	*33.15* [e]	24.80	3.84	6.08	4.54	17.15	26.53	19.76	86.94	86.94	86.94
GPT-2	6.59	14.62	8.91	1.06	2.34	1.42	5.12	11.37	6.93	83.65	83.65	83.65
GPT-Neo	9.97	19.91	13.01	1.01	2.30	1.38	7.89	15.91	10.33	83.12	83.12	83.12
GPT-J	13.22	29.99	17.88	3.37	*7.96*	4.59	10.71	24.34	14.47	86.28	86.28	86.28
MentalLlama	*33.03*	32.79	*30.86*	*8.66*	6.50	*7.28*	*27.73*	*27.30*	*29.55*	*89.40*	*90.99*	*90.99*
Mistral	29.07	26.56	25.41	7.03	5.20	7.19	25.45	25.61	26.62	83.42	85.96	83.05
Llama-2	28.49	24.17	23.47	6.40	4.68	6.63	22.7	23.04	23.66	82.86	83.80	81.62
Phi-2	21.23	10.42	13.81	1.89	1.43	1.78	14.56	9.19	11.26	84.25	82.00	83.11

[a]ROUGE: Recall-Oriented Understudy for Gisting Evaluation.

[b]BERTScore: Bidirectional Encoder Representations from Transformers Score.

[c]BART: Bidirectional and Auto-Regressive Transformer.

[d]T5: Text-To-Text Transfer Transformer.

[e]The best results are italicized.

PD Summarization

The experimental results presented in Table 4 focus on the summarization task for the PD psychotherapy element. Considering the ROUGE-1 metric, MentalLlama demonstrated superior performance compared to the other LLMs. MentalLlama achieved an F_1-score of 30.95, followed by MentalBART (with an F_1-score of 29.94). For the ROUGE-2 metric, GPT-J outperformed the other models, followed by MentalLlama. In addition, in terms of the ROUGE-L metric, the top 2 models with the highest F_1-score values were F1 score models were MentalLlama and Mistral. Finally, MentalBART superseded the other models with an F_1-score of 88.61 with respect to the BERTScore metric. Overall, the scores indicate that LLMs such as MentalLlama and MentalBART, which were pretrained on the mental domain data, show consistent superiority. Notably, the base Mistral model also performed comparably to, and sometimes better than, the models trained on the mental health domain data.

Table 4. Results obtained on Mental Health Counseling-Component–Guided Dialogue Summaries (MentalCLOUDS) for the summarization task on the patient discovery psychotherapy element.

Model	ROUGE[a]-1			ROUGE-2			ROUGE-L			BERTScore[b]		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score	Precision	Recall	F_1-score	Precision	Recall	F_1-score
BART[c]	20.82	43.24	26.72	5.97	12.93	7.74	16.38	34.82	21.14	87.35	87.35	87.35
T5[d]	9.43	47.29	15.34	3.03	16.90	5.01	8.39	42.58	13.67	84.77	84.77	84.77
MentalBART	33.51	29.94	29.94	9.36	7.94	8.06	23.39	21.44	21.10	*88.61* [e]	88.61	*88.61*
Flan-T5	21.08	35.61	24.44	4.81	8.89	5.63	16.13	28.29	18.94	86.52	86.52	86.52
GPT-2	13.66	36.24	19.57	4.08	11.27	5.94	10.93	29.42	15.70	85.21	85.21	85.21
GPT-Neo	12.96	29.93	17.83	2.32	5.44	3.22	9.84	23.10	13.60	82.72	82.72	82.72
GPT-J	19.78	*53.33*	28.85	*12.68*	*35.71*	*18.71*	16.12	*43.33*	23.49	86.43	86.43	86.43
MentalLlama	*24.56*	43.84	*30.95*	9.55	26.01	12.79	*23.77*	38.98	*29.17*	84.63	*88.95*	86.68
Mistral	22.84	39.02	27.54	8.78	25.79	11.35	21.90	35.98	24.02	86.62	87.28	84.49
Llama-2	20.22	34.7	26.1	8.41	21.13	10.39	14.73	21.44	17.79	78.81	88.06	81.48
Phi-2	18.72	9.23	12.45	5.61	4.44	4.96	13.94	8.73	10.98	84.25	82.00	80.05

[a]ROUGE: Recall-Oriented Understudy for Gisting Evaluation.

[b]BERTScore: Bidirectional Encoder Representations from Transformers Score.

[c]BART: Bidirectional and Auto-Regressive Transformer.

[d]T5: Text-To-Text Transfer Transformer.

[e]The best results are italicized.

Reflecting

Table 5 reports the automatic evaluation scores on the summarization task for the reflecting psychotherapy element. In terms of the ROUGE-1 metric, MentalLlama and Mistral were the best 2 models, with F_1-score values of 39.52 and 38.33, respectively. Similarly, MentalLlama demonstrated its superiority over the other LLMs in terms of the ROUGE-2, ROUGE-L and BERTScore metrics. Moreover, the scores of the summarization tasks for this psychotherapy element were analogous to those of the previous 2 summarization tasks, namely SH and PD, wherein the mental health–specific LLMs exhibited their superiority over the other LLMs.

Table 5. Results obtained on Mental Health Counseling-Component–Guided Dialogue Summaries (MentalCLOUDS) for the summarization task on the reflecting psychotherapy element.

Model	ROUGE[a]-1			ROUGE-2			ROUGE-L			BERTScore[b]		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score	Precision	Recall	F_1-score	Precision	Recall	F_1-score
BART[c]	17.01	23.04	18.08	2.87	4.25	3.22	12.68	17.79	13.66	85.26	85.26	85.26
T5[d]	34.13	19.32	24.31	7.21	3.97	5.04	22.95	12.82	16.21	84.92	84.92	84.92
MentalBART	*34.99* [e]	36.54	34.46	*10.24*	10.66	10.07	24.52	25.80	24.25	*88.70*	*88.70*	*88.70*
Flan-T5	25.10	41.40	30.15	7.19	12.03	8.64	18.52	31.00	22.36	87.41	87.41	87.41
GPT-2	2.84	7.54	4.08	0.14	0.33	0.20	2.35	6.34	3.39	82.66	82.66	82.66
GPT-Neo	1.14	3.97	1.74	0.00	0.00	0.00	1.14	3.97	1.74	80.88	80.88	80.88
GPT-J	17.60	38.33	23.71	5.07	*13.04*	7.13	14.98	32.85	20.18	86.94	86.94	86.94
MentalLlama	31.68	*54.76*	*39.52*	8.26	11.99	*10.17*	*27.13*	*37.59*	*26.56*	84.77	86.92	87.43
Mistral	29.15	49.28	38.33	8.42	11.87	8.34	24.41	34.20	23.44	78.83	79.97	84.81
Llama-2	26.93	43.81	31.22	6.10	9.23	8.24	16.82	20.67	16.21	78.93	86.05	82.19
Phi-2	10.61	5.21	6.91	0.94	0.71	0.89	7.28	4.60	5.53	86.94	82.17	84.49

[a]ROUGE: Recall-Oriented Understudy for Gisting Evaluation.

[b]BERTScore: Bidirectional Encoder Representations from Transformers Score.

[c]BART: Bidirectional and Auto-Regressive Transformer.

[d]T5: Text-To-Text Transfer Transformer.

[e]The best results are italicized.

Qualitative Assessment by Experts

Expert Panel Composition and Evaluation Framework

To conduct a comprehensive expert assessment, 5 health care professionals were employed to assess the clinical appropriateness of the summaries produced by the LLMs based on the evaluation framework postulated by Sekhon et al [69]. Of the 5 health care professionals, 2 (40%) were clinical psychologists and 3 (60%) were psychiatrists and medical practitioners; 4 (80%) were male and 1 (20%) was female; and their ages ranged from 40 to 55 years. Furthermore, each health care professional possessed more than a decade of therapeutic experience.

The evaluation framework encompasses 6 crucial parameters: affective attitude, burden, ethicality, coherence, opportunity costs, and perceived effectiveness. The experts evaluated each session summary against these acceptability parameters, assigning continuous ratings on a scale ranging from 0 to 2, where a higher rating signified enhanced acceptability. In addition, we incorporated a new parameter: the extent of hallucination. It is categorical: 0=extensive hallucination observed, 1=*minimal hallucination observed*, and 2=*no hallucination observed*. These evaluative dimensions are defined in Table 6.

Table 7 reports the clinical experts' scores averaged over their ratings. The clinical acceptability framework [69] involves 6 parameters: affective attitude, burden, ethicality, coherence, opportunity costs, and perceived effectiveness (refer to Table 6 for more details). We selected the 3 best LLMs (MentalLlama, Mistral, and MentalBART) for the expert evaluation based on the automatic evaluation results. Notably, Mistral outperformed the other 2 LLMs across all metrics, although the other 2 LLMs were fine-tuned on mental health domain data. Overall, all raters were more aligned in rating the MentalBART model with less variance than the other 2 LLMs across all metrics. However, all 3 LLMs were rated higher on the surface-level–characteristic metric (burden) or subjective metric (affective attitude) than the opportunity costs and efficacy metrics (perceived effectiveness). The poor scores of all 3 models on the more sensitive aspects, that is, the overall efficacy and the opportunity costs, indicate that these models share the same weakness and are not suitable for clinical use as they stand now.

Table 6. Explanation of the experts' evaluation metrics based on the evaluation framework postulated by Sekhon et al [69].

Construct	Definition	Application
Affective attitude	How an individual feels about an intervention	What are your perceptions of the summarization based upon your clinical knowledge?
Burden	Perceived amount of effort required to participate	How much effort is required to understand the summarization (consider spelling, grammar, and overall interpretation)?
Ethicality	Extent to which this is a good fit with your organization's value system	How does this align with your respective code of ethics? Are there concerns?
Coherence	Extent to which the intervention is understood	How well the summaries are understood
Opportunity costs	The extent to which one would benefit from using this intervention	Pros and cons of using this intervention in your respective setting
Perceived effectiveness	Extent to which this intervention will perform in the intended setting	How well this will perform in your clinical setting
Extent of hallucination	Extent to which this intervention is hallucinated	The generated text is incorrect, nonsensical, or contains global information apart from the context of the conversation

Table 7. Qualitative evaluation by human experts, with scores averaged from the 5 expert raters. The variances among the raters' scores are also shown.

Model	Affective attitude	Burden	Ethicality	Intervention coherence	Opportunity costs	Perceived effectiveness
Mistral						
Values, mean (SD)	*1.12 (0.47)* [a]	*1.33 (0.32)*	*1.42 (0.37)*	*1.13 (0.45)*	*0.98 (0.47)*	*0.90 (0.51)*
Variance	0.22	0.10	0.14	0.20	0.22	0.26
MentalLlama						
Values, mean (SD)	*1.12 (0.37)*	*1.33* (0.22)	1.36 (0.32)	1.06 (0.36)	0.94 (0.39)	0.88 (0.45)
Variance	0.14	0.05	0.10	0.13	0.15	0.20
MentalBART						
Values, mean (SD)	0.95 (0.28)	1.28 (0.14)	1.33 (0.36)	1.01 (0.22)	0.84 (0.33)	0.76 (0.4)
Variance	0.08	0.02	0.13	0.05	0.11	0.16

[a]The best results are italicized.

Extent of Hallucination

The evaluation of hallucination identification in a set of 39 conversations was divided into 3 hallucination levels: *no hallucinationobserved*, minimal hallucination observed, and extensive hallucination observed. These categories essentially determine how well the response is consistent with the context and whether it is also incorrect, nonsensical, or contains global information beyond the scope of the conversation. The results are summarized in Table 8. The data show fluctuations in how the phenomenon of hallucination is perceived among different models and stress the importance of reviewing evaluations from numerous appraisers for a complete assessment. Here, we report the average hallucination-level frequencies rated by the 5 evaluators. Subsequently, we provide the percentage of the hallucination-level frequency against the total 39 instances. Of the test conversations, the majority of cases (n=39, 76%), on average, demonstrated *no hallucination observed*: Mistral and MentalBART achieved rates of 75% and 76%, respectively, while MentalLlama showed a slightly higher value: 77%. Among the samples where minimal hallucination observed was reported, all 3 models fell within a similar range: Mistral and MentalLlama had rates of 13% and 14%, respectively, while MentalBART showed a slightly elevated value of 18%. Notably, the models exhibited lower rates in terms of the extensive hallucination observed category, with Mistral at only 11%, MentalLlama at 7%, and MentalBART at 5%. These data confirm the capability of these AI models to faithfully follow whenever there is no hallucination and underscore their ability to detect more subtle degrees of hallucination across the various tasks on which they were tested.

The results are consistently adequate across all 3 models, with a relatively equal distribution of the level of hallucination observed by different raters. Importantly, all 3 models exhibited a significant number of cases with *no hallucination observed*, indicating reliable performance and implying their ability to maintain fidelity to the original content.

Table 8. Hallucination-level frequency marked by experts for the top 3 large language models. The average of hallucination-level frequencies for each rater is reported.

Hallucination level	Mistral (%), mean (SD)	MentalLlama (%), mean (SD)	MentalBART (%), mean (SD)
No hallucination observed	29.3 (1.64)	30.3 (2.03)	29.7 (1.58)
Minimal hallucination observed	5.1 (0.51)	5.6 (1.07)	7.3 (0.96)
Extensive hallucination observed	4.3 (1.34)	3 (1)	2 (0.67)

Discussion

Principal Findings

In this study, we assessed 11 state-of-the-art LLMs on the aspect-based summarization task of mental health therapy conversations. These therapy conversations are long, and it requires a good amount of effort to gain insights from reading them. To address this, we summarized these long conversations, thereby reducing the efforts of the experts. We further proposed MentalCLOUDS, which provides aspect-based summaries of each conversation.

Specifically, we benchmarked the 11 LLMs for aspect-based summarization and evaluated them using both automatic and human evaluation approaches. The automatic evaluation scores revealed the superiority of the LLMs trained on mental health domain data. Two domain-specific LLMs, MentalLlama and MentalBART, consistently outperformed the rest of the LLMs across all aspects. Notably, although Mistral is not specifically trained on mental health domain data, its scores are comparable to those of MentalLlama, the overall best-performing model.

This work also showcased the prowess of decoder-only LLMs compared to strong encoder-decoder–based LLMs. Typically, encoder-decoder models favor sequence-to-sequence tasks such as summarization, where a sequence of input texts is mapped to a sequence of output texts. However, the decoder-based models, that is, MentalLlama and Mistral, consistently outperformed the encoder-decoder models such as BART, T5, and Flan-T5. The only exception was MentalBART because it is fine-tuned on the mental health data set.

The counseling data set was curated from multiple multimedia web-based sources such as YouTube transcripts [47]. Hence, most of these natural conversations are incoherent and grammatically unfluent. Even with these imperfections, the LLMs were mostly able to construct meaningful summaries that contained coherent narratives with a clear beginning and end. However, the models did not do as well with the structure separation of the information. The SH, PD, and reflection sections frequently overlapped, posing clinical and legal problems. History is considered clinically sacrosanct and should not be contaminated by the therapist's interpretation, and it is also citable in legal cases as client evidence, while interpretations are not. The models were also unable to identify psychotherapy types (eg, cognitive behavioral therapy) and therapy techniques, which form an integral part of counseling notes; for example, when participants are engaged in using a motivational interviewing framework, the essential processes and their outcomes, which a human summarizer would have recorded, failed to find a place in the LLM summaries. Important negative histories gathered during the session, such as the history of suicide risk or substance use, were also not recorded; and in at least 1 instance, the presence of suicide risk was not identified. In general, the models exhibited stronger performance in handling medical histories and examinations but struggled when faced with more technical and sensitive aspects, such as conversations related to actual therapeutic strategies.

Limitations

It is crucial to address the limitations of this study for a comprehensive understanding. First, this work aimed to benchmark the efficacy of only 11 LLMs on the aspect-based summarization task. Second, for faster and easier reproduction of the results, we did not assess models larger than 7 billion parameters; however, such models can be part of future examinations. Third, for the initial study and to promote research in this field, only open-source models were assessed in this work. However, inspecting closed models such as ChatGPT, Claude, and Gemini can be an interesting future research avenue. Finally, this work explored only 3 aspects (counseling components) of the conversation. However, conversations are subjective and can have >3 components. In addition, the counseling sessions in this work represented a certain demographic region (American) and thus may not apply to therapy counseling for other demographics.

Conclusions

Our study benchmarked the efficacy and role of LLMs in counseling-component–guided summarization tasks. In doing so, we introduced a new data set, MentalCLOUDS, which comprises summaries corresponding to 3 counseling components. The experimental results confirmed the superiority of the LLMs fine-tuned on mental health domain data (MentalLlama and MentalBART) over the out-of-the-box LLMs. Notably, the out-of-the-box Mistral model seemed comparable to, and sometimes better than, the LLMs fine-tuned on mental health domain data. However, as per the experts' evaluation, these LLMs often failed to distinguish between the counseling components during summary generation. Overall, these models excelled in managing medical histories and examinations but faced challenges with technical and sensitive aspects, such as therapy conversations, thereby limiting their clinical utility as they stand now.

Acknowledgments

The authors would like to thank JMIR Publications for providing article processing fee support. The salary of the support staff involved in the project was partly supported by Tower Research Capital Markets toward using machine learning for social good.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. Chen P, Verma R. A query-based medical information summarization system using ontology knowledge. In: Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems. 2006 Presented at: CBMS'06; June 22-23, 2006; Salt Lake City, UT. [doi: 10.1109/cbms.2006.25]
2. Chen YC, Bansal M. Fast abstractive summarization with reinforce-selected sentence rewriting. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018 Presented at: ACL 2018; July 15-20, 2018; Melbourne, Australia. [doi: 10.18653/v1/p18-1063]
3. Narayan S, Cohen SB, Lapata M. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018 Presented at: EMNLP 2018; October 31-November 4, 2018; Brussels, Belgium. [doi: 10.18653/v1/d18-1206]
4. Moratanch N, Chitrakala S. A survey on extractive text summarization. In: Proceedings of the International Conference on Computer, Communication and Signal Processing. 2017 Presented at: ICCCSP 2017; January 10-11, 2017; Chennai, India. [doi: 10.1109/icccsp.2017.7944061]
5. Gupta S, Gupta SK. Abstractive summarization: an overview of the state of the art. Expert Syst Appl 2019 May;121:49-65. [doi: 10.1016/j.eswa.2018.12.011]
6. Tuggener D, Mieskes M, Deriu J, Cieliebak M. Are we summarizing the right way? A survey of dialogue summarization data sets. In: Proceedings of the Third Workshop on New Frontiers in Summarization. 2021 Presented at: NewSum 2021; November 10, 2021; Online. [doi: 10.18653/v1/2021.newsum-1.12]
7. Konovalov S, Scotch M, Post L, Brandt C. Biomedical informatics techniques for processing and analyzing web blogs of military service members. J Med Internet Res 2010 Oct 05;12(4):e45 [FREE Full text] [doi: 10.2196/jmir.1538] [Medline: 20923755]
8. Strauss J, Peguero AM, Hirst G. Machine learning methods for clinical forms analysis in mental health. Stud Health Technol Inform 2013;192:1024. [Medline: 23920798]
9. Kennedy SH, Lam RW, McIntyre RS, Tourjman SV, Bhat V, Blier P, et al. Canadian network for mood and anxiety treatments (CANMAT) 2016 clinical guidelines for the management of adults with major depressive disorder: section 3. Pharmacological treatments. Can J Psychiatry 2016 Sep;61(9):540-560 [FREE Full text] [doi: 10.1177/0706743716659417] [Medline: 27486148]
10. Tran T, Kavuluru R. Predicting mental conditions based on "history of present illness" in psychiatric notes with deep neural networks. J Biomed Inform 2017 Nov;75S:S138-S148 [FREE Full text] [doi: 10.1016/j.jbi.2017.06.010] [Medline: 28606869]
11. Chen YP, Chen YY, Lin JJ, Huang CH, Lai F. Modified bidirectional encoder representations from transformers extractive summarization model for hospital information systems based on character-level tokens (AlphaBERT): development and performance evaluation. JMIR Med Inform 2020 Apr 29;8(4):e17787 [FREE Full text] [doi: 10.2196/17787] [Medline: 32347806]
12. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019 Presented at: NAACL 2019; June 2-7, 2019; Minneapolis, MN. [doi: 10.18653/v1/N19-1423]
13. Ive J, Viani N, Kam J, Yin L, Verma S, Puntis S, et al. Generation and evaluation of artificial mental health records for Natural Language Processing. NPJ Digit Med 2020 May 14;3:69 [FREE Full text] [doi: 10.1038/s41746-020-0267-x] [Medline: 32435697]
14. Afzal M, Alam F, Malik KM, Malik GM. Clinical context-aware biomedical text summarization using deep neural network: model development and validation. J Med Internet Res 2020 Oct 23;22(10):e19810 [FREE Full text] [doi: 10.2196/19810] [Medline: 33095174]
15. Manas G, Aribandi V, Kursuncu U, Alambo A, Shalin VL, Thirunarayan K, et al. Knowledge-infused abstractive summarization of clinical diagnostic interviews: framework development study. JMIR Ment Health 2021 May 10;8(5):e20865 [FREE Full text] [doi: 10.2196/20865] [Medline: 33970116]

16. Zhang L, Negrinho R, Ghosh A, Jagannathan V, Hassanzadeh HR, Schaaf T, et al. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021 Presented at: EMNLP 2021; November 7-11, 2021; Virtual event; Punta Cana, Dominican Republic. [doi: 10.18653/v1/2021.findings-emnlp.313]
17. Zafari H, Zulkernine F. Chatsum: an intelligent medical chat summarization tool. In: Proceedings of the 2021 IEEE EMBS International Conference on Biomedical and Health Informatics. 2021 Presented at: BHI 2021; July 27-30, 2021; Athens, Greece. [doi: 10.1109/bhi50953.2021.9508585]
18. Nallapati R, Zhou B, Gu˙lçehre Ç, dos Santos C, Xiang B. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In: Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. 2016 Presented at: CoNLL 2016; August 11-12, 2016; Berlin, Germany. [doi: 10.18653/v1/k16-1028]
19. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv Preprint posted online on June 12, 2017 [FREE Full text]
20. See A, Liu PJ, Manning CD. Get to the point: summarization with pointer-generator networks. arXiv Preprint posted online on April 14, 2017 [FREE Full text] [doi: 10.18653/v1/p17-1099]
21. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res 2003;3:993-1022 [FREE Full text] [doi: 10.5555/944919.944937]
22. Song Y, Tian Y, Wang N, Xia F. Summarizing medical conversations via identifying important utterances. In: Proceedings of the 28th International Conference on Computational Linguistics. 2020 Presented at: COLING 2020; December 8-13, 2020; Barcelona, Spain. [doi: 10.18653/v1/2020.coling-main.63]
23. Quiroz JC, Laranjo L, Kocaballi AB, Briatore A, Berkovsky S, Rezazadegan D, et al. Identifying relevant information in medical conversations to summarize a clinician-patient encounter. Health Informatics J 2020 Dec;26(4):2906-2914 [FREE Full text] [doi: 10.1177/1460458220951719] [Medline: 32865113]
24. Krishna K, Khosla S, Bigham JP, Lipton ZC. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. arXiv. Preprint posted online on May 4, 2020 [FREE Full text] [doi: 10.18653/v1/2021.acl-long.384]
25. Titov I, McDonald R. A joint model of text and aspect ratings for sentiment summarization. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2008 Presented at: ACL-08: HLT; June 15, 2008; Columbus, OH.
26. Lu Y, Zhai C, Sundaresan N. Rated aspect summarization of short comments. In: Proceedings of the 18th International Conference on World Wide Web. 2009 Presented at: WWW '09; April 20-24, 2009; Madrid, Spain. [doi: 10.1145/1526709.1526728]
27. Yang M, Qu Q, Shen Y, Liu Q, Zhao W, Zhu J. Aspect and sentiment aware abstractive review summarization. In: Proceedings of the 27th International Conference on Computational Linguistics. 2018 Presented at: COLING 2018; August 20-26, 2018; Santa Fe, NM URL: https://aclanthology.org/C18-1095 [doi: 10.1145/3269206.3269273]
28. Wang L, Ling W. Neural network-based abstract generation for opinions and arguments. arXiv. Preprint posted online on June 9, 2016 [FREE Full text] [doi: 10.18653/v1/n16-1007]
29. Frermann L, Klementiev A. Inducing document structure for aspect-based summarization. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019 Presented at: ACL 2019; July 28-August 2, 2019; Florence, Italy. [doi: 10.18653/v1/p19-1630]
30. Krishna K, Srinivasan BV. Generating topic-oriented summaries using neural attention. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018 Presented at: NAACL-HLT 2018; June 1-6, 2018; New Orleans, LA. [doi: 10.18653/v1/n18-1153]
31. Hayashi H, Budania P, Wang P, Ackerson C, Neervannan R, Neubig G. WikiAsp: a dataset for multi-domain aspect-based summarization. arXiv. Preprint posted online on November 16, 2020 [FREE Full text] [doi: 10.1162/tacl_a_00362]
32. Yang X, Song K, Cho S, Wang X, Pan X, Petzold L, et al. OASum: large-scale open domain aspect-based summarization. arXiv. Preprint posted online on December 19, 2022 [FREE Full text] [doi: 10.18653/v1/2023.findings-acl.268]
33. Beltagy I, Peters ME, Cohan A. Longformer: the long-document transformer. arXiv. Preprint posted online on April 10, 2020 [FREE Full text] [doi: 10.48550/arXiv.2004.05150]
34. Joshi A, Katariya N, Amatriain X, Kannan A. Dr. Summarize: global summarization of medical dialogue by exploiting local structures. arXiv. Preprint posted online on September 18, 2020 [FREE Full text] [doi: 10.18653/v1/2020.findings-emnlp.335]
35. Liu Z, Ng A, Lee S, Aw AT, Chen NF. Topic-aware pointer-generator networks for summarizing spoken conversations. arXiv. Preprint posted online on October 3, 2019 [FREE Full text] [doi: 10.1109/asru46091.2019.9003764]
36. Kazi N, Kahanda I. Automatically generating psychiatric case notes from digital transcripts of doctor-patient conversations. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019 Presented at: ClinicalNLP 2019; June 7, 2019; Minneapolis, MN. [doi: 10.18653/v1/w19-1918]
37. Gundogdu B, Pamuksuz U, Chung JH, Telleria JM, Liu P, Khan F, et al. Customized impression prediction from radiology reports using BERT and LSTMs. IEEE Trans Artif Intell 2023 Aug;4(4):744-753. [doi: 10.1109/tai.2021.3086435]

38. Han Q, Yang Z, Lin H, Qin T. Let topic flow: a unified topic-guided segment-wise dialogue summarization framework. IEEE/ACM Trans Audio Speech Lang Process 2024;32:2021-2032. [doi: 10.1109/taslp.2024.3374112]
39. Zhang M, You D, Wang S. Novel framework for dialogue summarization based on factual-statement fusion and dialogue segmentation. PLoS One 2024 Apr 16;19(4):e0302104 [FREE Full text] [doi: 10.1371/journal.pone.0302104] [Medline: 38625864]
40. Chintagunta B, Katariya N, Amatriain X, Kannan A. Medically aware GPT-3 as a data generator for medical dialogue summarization. arXiv. Preprint posted online on September 9, 2021 [FREE Full text] [doi: 10.18653/v1/2021.nlpmc-1.9]
41. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv. Preprint posted online on May 28, 2020 [FREE Full text]
42. Liu Y, Ju S, Wang J. Exploring the potential of ChatGPT in medical dialogue summarization: a study on consistency with human preferences. BMC Med Inform Decis Mak 2024 Mar 14;24(1):75 [FREE Full text] [doi: 10.1186/s12911-024-02481-8] [Medline: 38486198]
43. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv. Preprint posted online on October 29, 2019 [FREE Full text] [doi: 10.18653/v1/2020.acl-main.703]
44. Liu Y. Fine-tune BERT for extractive summarization. arXiv. Preprint posted online on March 25, 2019 [FREE Full text]
45. Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024 Apr 27;30(4):1134-1142. [doi: 10.1038/s41591-024-02855-5] [Medline: 38413730]
46. Singh LG, Mao J, Mutalik R, Middleton SE. Extracting and summarizing evidence of suicidal ideation in social media contents using large language models. In: Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology. 2024 Presented at: CLPsych 2024; March 21, 2024; St. Julians, Malta.
47. Srivastava A, Suresh T, Lord SP, Akhtar MS, Chakraborty T. Counseling summarization using mental health knowledge guided utterance filtering. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022 Presented at: KDD '22; August 14-18, 2022; Washington, DC. [doi: 10.1145/3534678.3539187]
48. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv. Preprint posted online on October 23, 2019 [FREE Full text]
49. Shah DJ, Yu L, Lei T, Barzilay R. Nutri-bullets: summarizing health studies by composing segments. arXiv. Preprint posted online on March 22, 2021 [FREE Full text] [doi: 10.1609/aaai.v35i15.17624]
50. Park JW. Continual BERT: continual learning for adaptive extractive summarization of COVID-19 literature. arXiv. Preprint posted online on July 27, 2020 [FREE Full text] [doi: 10.48550/arXiv.2007.03405]
51. Shah D, Yu L, Lei T, Barzilay R. Nutri-bullets hybrid: consensual multi-document summarization. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021 Presented at: NAACL 2021; June 6-11, 2021; Virtual event. [doi: 10.18653/v1/2021.naacl-main.411]
52. Wallace BC, Saha S, Soboczenski F, Marshall IJ. Generating (factual?) narrative summaries of RCTs: experiments with neural multi-document summarization. AMIA Jt Summits Transl Sci Proc 2021;2021:605-614 [FREE Full text] [Medline: 34457176]
53. Zhu Y, Yang X, Wu Y, Zhang W. Leveraging summary guidance on medical report summarization. IEEE J Biomed Health Inform 2023 Oct;27(10):5066-5075. [doi: 10.1109/jbhi.2023.3304376]
54. Jiang Z, Cai X, Yang L, Gao D, Zhao W, Han J, et al. Learning to summarize Chinese radiology findings with a pre-trained encoder. IEEE Trans Biomed Eng 2023 Dec;70(12):3277-3287. [doi: 10.1109/tbme.2023.3280987]
55. Van Veen D, Van Uden C, Attias M, Pareek A, Bluethgen C, Polacin M, et al. RadAdapt: radiology report summarization via lightweight domain adaptation of large language models. arXiv. Preprint posted online on May 2, 2023 [FREE Full text] [doi: 10.18653/v1/2023.bionlp-1.42]
56. Jo HS, Park K, Jung SM. A scoping review of consumer needs for cancer information. Patient Educ Couns 2019 Jul;102(7):1237-1250. [doi: 10.1016/j.pec.2019.02.004] [Medline: 30772114]
57. Finney Rutten LJ, Blake KD, Greenberg-Worisek AJ, Allen SV, Moser RP, Hesse BW. Online health information seeking among US adults: measuring progress toward a healthy people 2020 objective. Public Health Rep 2019;134(6):617-625 [FREE Full text] [doi: 10.1177/0033354919874074] [Medline: 31513756]
58. Mrini K, Dernoncourt F, Yoon S, Bui T, Chang W, Farcas E, et al. A gradually soft multi-task and data-augmented approach to medical question understanding. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021 Presented at: ACL/IJCNLP 2021; August 1-6, 2021; Virtual Event. [doi: 10.18653/v1/2021.acl-long.119]
59. Roberts K, Demner-Fushman D. Interactive use of online health resources: a comparison of consumer and professional questions. J Am Med Inform Assoc 2016 Jul;23(4):802-811 [FREE Full text] [doi: 10.1093/jamia/ocw024] [Medline: 27147494]
60. Kolhatkar G, Paranjape A, Gokhale O, Kadam D. Team converge at ProbSum 2023: abstractive text summarization of patient progress notes. In: Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks. 2023 Presented at: BioNLP@ACL 2023; July 13, 2023; Toronto, Canada. [doi: 10.18653/v1/2023.bionlp-1.50]

61. Liu M, Zhang D, Tan W, Zhang H. DeakinNLP at ProbSum 2023: clinical progress note summarization with rules and language models. In: Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks. 2023 Presented at: BioNLP@ACL 2023; July 13, 2023; Toronto, Canada. [doi: 10.18653/v1/2023.bionlp-1.47]
62. Gao Y, Dligach D, Miller T, Churpek MM, Afshar M. Overview of the problem list summarization (ProbSum) 2023 shared task on summarizing patients' active diagnoses and problems from electronic health record progress notes. arXiv. Preprint posted online on June 8, 2023 [FREE Full text] [doi: 10.18653/v1/2023.bionlp-1.43]
63. Gao Y, Miller T, Xu D, Dligach D, Churpek MM, Afshar M. Summarizing patients' problems from hospital progress notes using pre-trained sequence-to-sequence models. Proc Int Conf Comput Ling 2022 Oct;2022:2979-2991 [FREE Full text] [Medline: 36268128]
64. Shing HC, Shivade C, Pourdamghani N, Nan F, Resnik P, Oard D, et al. Towards clinical encounter summarization: learning to compose discharge summaries from prior notes. arXiv. Preprint posted online on April 27, 2021 [FREE Full text] [doi: 10.48550/arXiv.2104.13498]
65. Ando K, Okumura T, Komachi M, Horiguchi H, Matsumoto Y. Exploring optimal granularity for extractive summarization of unstructured health records: analysis of the largest multi-institutional archive of health records in Japan. PLOS Digit Health 2022 Sep;1(9):e0000099 [FREE Full text] [doi: 10.1371/journal.pdig.0000099] [Medline: 36812582]
66. Zhao B, Zan H, Niu C, Chang H, Zhang K. Automatic generation of discharge summary of EMRs based on multi-granularity information fusion. In: Proceedings of the 9th China Health Information Processing Conference. 2023 Presented at: CHIP 2023; October 27-29, 2023; Hangzhou, China. [doi: 10.1007/978-981-99-9864-7_17]
67. Jain R, Saha T, Lalwani J, Saha S. Can you summarize my learnings? Towards perspective-based educational dialogue summarization. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023 Presented at: EMNLP 2023; December 6-10, 2023; Singapore, Singapore. [doi: 10.18653/v1/2023.findings-emnlp.208]
68. Saha T, Reddy S, Das A, Saha S, Bhattacharyya P. A shoulder to cry on: towards a motivational virtual assistant for assuaging mental agony. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022 Presented at: NAACL-HLT 2022; July 10-15, 2022; Online. [doi: 10.18653/v1/2022.naacl-main.174]
69. Sekhon M, Cartwright M, Francis JJ. Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework. BMC Health Serv Res 2017 Jan 26;17(1):88 [FREE Full text] [doi: 10.1186/s12913-017-2031-8] [Medline: 28126032]
70. Poria S, Majumder N, Mihalcea R, Hovy E. Emotion recognition in conversation: research challenges, datasets, and recent advances. arXiv. Preprint posted online on May 8, 2019 [FREE Full text]
71. Chen H, Liu X, Yin D, Tang J. A survey on dialogue systems: recent advances and new frontiers. ACM SIGKDD Explor Newsl 2017 Nov 21;19(2):25-35 [FREE Full text] [doi: 10.1145/3166054.3166058]
72. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI Blog 2019;1(8):9.
73. Black S, Biderman S, Hallahan E, Anthony Q, Gao L, Golding L, et al. GPT-NeoX-20B: an open-source autoregressive language model. arXiv. Preprint posted online on April 14, 2022 [FREE Full text] [doi: 10.18653/v1/2022.bigscience-1.9]
74. Gao L, Biderman S, Black S, Golding L, Hoppe T, Foster C, et al. The pile: an 800GB dataset of diverse text for language modeling. arXiv. Preprint posted online on December 31, 2020 [FREE Full text] [doi: 10.48550/arXiv.2101.00027]
75. Li R, Su J, Duan C, Zheng S. Linear attention mechanism: an efficient attention for semantic segmentation. arXiv. Preprint posted online on July 29, 2020 [FREE Full text]
76. Shazeer N, Mirhoseini A, Maziarz K, Davis A, Le Q, Hinton G, et al. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. arXiv. Preprint posted online on January 23, 2017 [FREE Full text] [doi: 10.48550/arXiv.1701.06538]
77. Ho J, Kalchbrenner N, Weissenborn D, Salimans T. Axial attention in multidimensional transformers. arXiv. Preprint posted online on December 20, 2019 [FREE Full text]
78. Wang B. Mesh-transformer-JAX: model-parallel implementation of transformer language model with JAX. GitHub. 2021. URL: https://github.com/kingoflolz/mesh-transformer-jax [accessed 2024-07-05]
79. Chung HW, Hou L, Longpre S, Zoph B, Tay Y, Fedus W, et al. Scaling instruction-finetuned language models. arXiv. Preprint posted online on October 20, 2022 [FREE Full text]
80. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, de las Casas D, et al. Mistral 7B. arXiv. Preprint posted online on October 10, 2023 [FREE Full text] [doi: 10.48550/arXiv.2310.06825]
81. Ainslie J, Lee-Thorp J, de Jong, M, Zemlyanskiy Y, Lebrón F, Sanghai S. GQA: training generalized multi-query transformer models from multi-head checkpoints. arXiv. Preprint posted online on May 22, 2023 [FREE Full text] [doi: 10.18653/v1/2023.emnlp-main.298]
82. Yang K, Zhang T, Kuang Z, Xie Q, Ananiadou S, Huang J. MentaLLaMA: interpretable mental health analysis on social media with large language models. arXiv. Preprint posted online on September 24, 2023 [FREE Full text] [doi: 10.1145/3589334.3648137]
83. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: open foundation and fine-tuned chat models. arXiv. Preprint posted online on July 18, 2023 [FREE Full text] [doi: 10.48550/arXiv.2307.09288]

84. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al. Training language models to follow instructions with human feedback. arXiv. Preprint posted online on March 4, 2022 [FREE Full text]
85. Christiano P, Leike J, Brown TB, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. arXiv. Preprint posted online on June 12, 2017 [FREE Full text] [doi: 10.5260/chara.21.2.8]
86. Gunasekar S, Zhang Y, Aneja J, Mendes CC, Del Giorno A, Gopi S, et al. Textbooks are all you need. arXiv. Preprint posted online on June 20, 2023 [FREE Full text]
87. Lin CY. ROUGE: a package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004). 2004 Presented at: WAS 2004; July 25-26, 2004; Barcelona, Spain.
88. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. In: International Conference on Learning Representations. 2020 Apr 26 Presented at: International Conference on Learning Representations; 2020 April 26; Addis Ababa URL: https://openreview.net/forum?id=SkeHuCVFDr

Abbreviations

AI: artificial intelligence
BART: Bidirectional and Auto-Regressive Transformer
BERT: Bidirectional Encoder Representations from Transformers
BERTScore: Bidirectional Encoder Representations from Transformers Score
LLM: large language model
MEMO: Mental Health Summarization
MentalCLOUDS: Mental Health Counseling-Component–Guided Dialogue Summaries
PD: patient discovery
PICO: Population, Intervention, Comparison, and Outcomes
ROUGE: Recall-Oriented Understudy for Gisting Evaluation
SH: symptom and history
SLM: small language model
T5: Text-to-Text Transfer Transformer

Edited by A Tal; submitted 12.02.24; peer-reviewed by H Shah; comments to author 27.04.24; revised version received 21.05.24; accepted 23.05.24; published 23.07.24.

Please cite as:
Adhikary PK, Srivastava A, Kumar S, Singh SM, Manuja P, Gopinath JK, Krishnan V, Gupta SK, Deb KS, Chakraborty T
Exploring the Efficacy of Large Language Models in Summarizing Mental Health Counseling Sessions: Benchmark Study
JMIR Ment Health 2024;11:e57306
URL: https://mental.jmir.org/2024/1/e57306
doi:10.2196/57306
PMID:

Original Paper

Large Language Models Versus Expert Clinicians in Crisis Prediction Among Telemental Health Patients: Comparative Study

Christine Lee[1], PhD; Matthew Mohebbi[1], BS; Erin O'Callaghan[1], PhD; Mirène Winsberg[1], MD

Brightside Health, San Francisco, CA, United States

Corresponding Author:
Mirène Winsberg, MD
Brightside Health
2261 Market Street
STE 10222
San Francisco, CA, 94114
United States
Phone: 1 415 279 2042
Email: mimi.winsberg@brightside.com

Abstract

Background: Due to recent advances in artificial intelligence, large language models (LLMs) have emerged as a powerful tool for a variety of language-related tasks, including sentiment analysis, and summarization of provider-patient interactions. However, there is limited research on these models in the area of crisis prediction.

Objective: This study aimed to evaluate the performance of LLMs, specifically OpenAI's generative pretrained transformer 4 (GPT-4), in predicting current and future mental health crisis episodes using patient-provided information at intake among users of a national telemental health platform.

Methods: Deidentified patient-provided data were pulled from specific intake questions of the Brightside telehealth platform, including the chief complaint, for 140 patients who indicated suicidal ideation (SI), and another 120 patients who later indicated SI with a plan during the course of treatment. Similar data were pulled for 200 randomly selected patients, treated during the same time period, who never endorsed SI. In total, 6 senior Brightside clinicians (3 psychologists and 3 psychiatrists) were shown patients' self-reported chief complaint and self-reported suicide attempt history but were blinded to the future course of treatment and other reported symptoms, including SI. They were asked a simple yes or no question regarding their prediction of endorsement of SI with plan, along with their confidence level about the prediction. GPT-4 was provided with similar information and asked to answer the same questions, enabling us to directly compare the performance of artificial intelligence and clinicians.

Results: Overall, the clinicians' average precision (0.7) was higher than that of GPT-4 (0.6) in identifying the SI with plan at intake (n=140) versus no SI (n=200) when using the chief complaint alone, while sensitivity was higher for the GPT-4 (0.62) than the clinicians' average (0.53). The addition of suicide attempt history increased the clinicians' average sensitivity (0.59) and precision (0.77) while increasing the GPT-4 sensitivity (0.59) but decreasing the GPT-4 precision (0.54). Performance decreased comparatively when predicting future SI with plan (n=120) versus no SI (n=200) with a chief complaint only for the clinicians (average sensitivity=0.4; average precision=0.59) and the GPT-4 (sensitivity=0.46; precision=0.48). The addition of suicide attempt history increased performance comparatively for the clinicians (average sensitivity=0.46; average precision=0.69) and the GPT-4 (sensitivity=0.74; precision=0.48).

Conclusions: GPT-4, with a simple prompt design, produced results on some metrics that approached those of a trained clinician. Additional work must be done before such a model can be piloted in a clinical setting. The model should undergo safety checks for bias, given evidence that LLMs can perpetuate the biases of the underlying data on which they are trained. We believe that LLMs hold promise for augmenting the identification of higher-risk patients at intake and potentially delivering more timely care to patients.

(JMIR Ment Health 2024;11:e58129) doi:10.2196/58129

KEYWORDS

mental health; telehealth; PHQ-9; Patient Health Questionnaire-9; suicidal ideation; AI; LLM; OpenAI; GPT-4; generative pretrained transformer 4; tele-mental health; large language model; clinician; clinicians; artificial intelligence; patient information; suicide; suicidal; mental disorder; suicide attempt; psychologist; psychologists; psychiatrist; psychiatrists; psychiatry; clinical setting; self-reported; treatment; medication; digital mental health; machine learning; language model; suicide; crisis; telemental health; tele health; e-health; digital health

Introduction

Background

Suicide is a serious public health concern. Suicide rates have risen at an alarming rate in the past 20 years, and in the United States, suicide is the second leading cause of death in adults aged 18-45 years [1]. In 2021, approximately 50,000 people in the United States died by suicide, which marks the highest national rate of suicide in decades [2]. As suicide rates increase, the behavioral health care workforce in the United States has not expanded enough to keep up with these mental health demands, limiting the timely access to care that is essential for suicide risk detection and prevention [3].

Suicide risk is difficult to predict. Research has demonstrated that there are numerous individual, relationship, community, and societal risk factors associated with suicide, such as history of previous suicide attempts, psychiatric diagnosis, sense of hopelessness, social isolation, community violence, and access to lethal means of suicide [4-9]. More recently, suicide theories and research suggest ideation-to-action pathways to help explain suicide risk, where people who think about suicide are at a higher risk of participating in suicidal behavior [10-13].

The prevalence of suicidal ideation (SI), which is defined as “thinking about, considering, or planning suicide” [14], is common, with 12.3 million Americans aged 18 years and older having thoughts of suicide in 2021 [15]. SI is predictive of suicide attempts and completed suicide [16,17]. SI is also a more sensitive predictor of lifetime risk for suicide than imminent risk [18]. Research has suggested that among those exhibiting SI, there is a 29% conditional probability of making a suicide attempt [19]. Other research has shown that those with nearly daily SI were 5 to 8 times more likely to attempt suicide and 3 to 11 times more likely to die by suicide within 30 days [20].

Artificial intelligence (AI) methods have been used for assessing mental health factors such as psychiatric symptom severity, diagnosis, and clinical risk using free text generated by the patient. Researchers using natural language processing (NLP) and machine learning (ML) were able to identify suicidal behavior from electronic medical records [21] and detect SI in a variety of different free-text settings [22]. In addition, an NLP-based system to determine the likelihood of crisis in patient chat messages to their clinicians was developed and implemented with reliable retrospective and prospective performance as a clinical support tool for a crisis specialist team [23].

Recent advances in AI methods, such as large language models (LLMs), have also shown success in a variety of medical applications. Both generalist LLMs, such as generative pretrained transformer 4 (GPT-4), and medical domain–specific LLMs, such as Med-PaLM 2, have exhibited medical competency on benchmarks such as the United States Medical Licensing Examination (USMLE) exam [24,25]. Generalist LLMs can sometimes outperform the domain-specific LLMs, as was recently found with GPT-4 outperforming MedPaLM 2 on the MedQA medical benchmark [25]. Finally, Med-PaLM-2 was also found to be effective at determining psychiatric functioning from free text, including patient-generated information during patient interviews [26].

Objective

We seek to leverage the capabilities of LLMs to detect or predict SI with plan among patients enrolled in a national telemental health platform, using patient-generated free text at intake. We will benchmark the performance of this LLM-based prediction against a cohort of senior mental health clinician experts.

Methods

Overview

The study consisted of clinicians completing a digital questionnaire where they were asked to predict whether a patient would endorse SI with a plan during the course of their treatment, based on patient-generated text describing their chief complaint. The same chief complaint texts were then served to the LLM GPT-4 with the same questionnaire instructions. The classification performance of the clinicians and GPT-4 were evaluated and compared.

Data Acquisition

The retrospective patient data used in this study were collected as part of the standard of care at Brightside Health and deidentified for research purposes. All patients treated at Brightside consent at intake to the terms of use and privacy policy that include consenting to Brightside’s use of their data for research purposes.

Inclusion Criteria

Data from patients who completed intake on the Brightside platform after March 15, 2023, and endorsed current SI (at intake) or subsequent SI (post intake and during the course of treatment) were included in the study set, along with a random cohort of patients treated during the same time frame who never endorsed SI with plan. In order to be included in the study sample, patients had to attend at least 1 psychiatric or therapy appointment and complete the chief complaint section of their digital intake form. Patients who left the chief complaint section empty were excluded.

Data and Outcome Variables

Patient-generated free text (chief complaint) was extracted from patient intake as the answer to the question “In your own words, what are you feeling or experiencing?” and any personal identifiers (such as age, birthdate, name, location, email address, phone number, and social security number) within the free text were replaced with asterisks. In addition, patient data extracted from intake included age, gender identity, and history of previous suicide attempts. Clinicians and the LLM did not have access to the age or gender identity of the patients and were only shown deidentified patient-generated free text and then the patients’ self-reported history of suicide attempts.

SI with plan was determined from answers to question 9 of the Patient Health Questionnaire-9 (PHQ-9). The PHQ-9 is a self-report questionnaire consisting of 9 questions measuring depression symptom severity ranging from 0 to 3 (not at all, several days, more than half the days, and nearly every day, respectively) within the past 2 weeks and includes a specific question related to the frequency of suicidal thoughts (item 9). If a patient endorses SI on the Brightside platform (item 9 answer value >0), a follow-up Brightside proprietary question asks whether the suicidal thoughts are something the patient has made specific plans for. At Brightside, the PHQ-9 is administered to all patients at intake and requested every 2 weeks during the course of treatment. PHQ-9 answers at intake and the date of the first SI with plan relative to intake were also extracted for this study.

Classification Label Definitions

The patients positive for SI with plan were defined as those having endorsed SI in the PHQ-9 at intake or any point during the later course of treatment and subsequently responded that the SI was something they had made specific plans for. Patients negative for SI with plan were defined as those with no PHQ-9 item 9 values >0, that is, those who had never endorsed SI in their PHQ-9 screenings.

Clinician Questionnaire Design

After the creation of the study data set, 6 clinicians employed at Brightside Health were recruited and all consented to participate in the study. Clinicians 1, 2, and 3 are senior psychiatrists with 18 to 30 years of clinical experience. Clinicians 4, 5, and 6 are senior psychologists with 10 to 23 years of clinical experience.

Each clinician was presented with the same questionnaire with the same randomized order of questions. Presented with a patient's deidentified chief complaint, they were asked to answer the following questions: (1) do you expect this patient to endorse SI with a plan in the course of their treatment? (yes or no) and (2) how confident are you? (high, medium, or low).

The clinicians were then presented with additional information as to whether the patient endorsed a previous suicide attempt and reprompted with the same questions. All chief complaints and questions were displayed one at a time. Clinicians were given 2 weeks to complete the questionnaire with no other instructions or training and were not required to complete it in a single sitting.

Large Language Model Prompt Design

In this study, we evaluated OpenAI's GPT-4 text-only model's ability to predict whether a patient would endorse SI with plan during the course of treatment. All model results were produced in Python (Python Software Foundation) with OpenAI's application programming interface and GPT-4 ("gpt-4-1106-preview" with default parameters) [27].

First, the model context was constructed using step-by-step instructions and few-shot prompting with 3 real examples of chief complaints and their respective correct, expected answers (Textbox 1). Note that the examples chosen for the few-shot prompting were not included in the study data set and were redacted for publication.

Textbox 1. GPT-4 few-shot prompt template for predicting suicidal ideation with plan in chief complaint text.

You are a trained psychiatric professional. You will be presented with a patient's answer to the question "In your own words, what are you feeling or experiencing?" also referred to as a patient's chief complaint.

Your task is to determine based on a patient's chief complaint, if that patient will endorse suicidal ideation with plan in the course of their treatment.

You will generate your answer as JSON: {"at_risk": "", "confidence": "", "risk_indicators": "[]"}

Use the following step-by-step instructions to determine your answer:

- Step 1. Answer in the key "at_risk" with only the values Yes or No
- Step 2. Answer in the key "confidence" with only the values High, Medium, or Low, how confident you are in your answer.
- Step 3. Answer in the key "risk_indicators" with one or more values from the list of possible risk indicators below that determined your Yes or No answer.

Separate risk indicators with a "; ".

If Other, specify your answer as "Other-<insert your answer>". If no risk indicators are identified, answer with None.

List of possible risk indicators values:

- Recent Stress, Loss, or Trauma
- History of Trauma
- Chronic medical conditions
- Substance use
- Previous suicide attempt
- Lack or loss of relationships or support
- Social isolation
- Family history of suicide
- Impulsive or aggressive language
- Explicit mentions of suicide, suicidal thoughts, or self harm
- Death imagery or metaphors
- Apathy, indifference or emotional detachment
- Sense of Hopelessness
- Other

Here is an example of a chief complaint with a Yes to suicidal ideation with plan:

"*<text redacted for publication>* "

Your answer would be:{"risk_indicators": "Sense of Hopelessness; Social isolation; Explicit mentions of suicide, suicidal thoughts, or self harm", "at_risk": "Yes", "confidence": "High"}

Here is an example of a chief complaint with a No to suicidal ideation with plan: "*<text redacted for publication>*"

Your answer would be: {"risk_indicators": "None", "at_risk": "No", "confidence": "High"}

Here is an example of a chief complaint with a No to suicidal ideation with plan:

"*<text redacted for publication>*"

Your answer would be: {"risk_indicators": "None", "at_risk": "No","confidence": "High"}

Next, the output format of the model was specified as JavaScript Object Notation for ease of analysis. In addition to the prediction of SI with plan during the course of treatment, the model was also asked to provide a confidence level (high, medium, and low) to the prediction (similar to the clinicians' questionnaire) and to provide reasoning from a list of explicitly provided risk indicators.

Finally, the deidentified patient-generated chief complaint text was given to the model in the user prompt. Each chief complaint was provided independently and then the LLM was reset back to the original context.

In order to evaluate the model's performance when served the additional information of patient self-reported previous suicide attempts, the sentence "I have attempted suicide before" or "I have never attempted suicide before" was appended to the end of the chief complaint and served as the prompt with the same context.

Performance Analysis

All analyses were performed in Python 3.8.12 with the package scikit-learn version 1.3.1 [28]. For comparison of performance, analyses were performed on positive for SI with plan at intake versus negative for SI during the entire course of treatment, as well as positive for SI with plan post intake versus the same data set of negative for SI during treatment.

Classification and Predictive Performance

Clinician and model performances in the ability to predict whether a chief complaint text sample was positive for SI with plan, at intake, and post intake, were evaluated for accuracy, sensitivity, specificity, and precision. Accuracy was defined as the proportion of correctly predicted samples over the total number of samples. Precision (or positive predictive value) was defined as the proportion of correctly predicted positive samples over the total number of predicted positive samples. Sensitivity was defined as the proportion of correctly predicted positive samples over the total number of positive samples. Specificity was defined as the proportion of correctly predicted negative samples over the total number of negative samples. As an additional baseline reference, previous suicide attempt information (yes or no) as a stand-alone predictor was also included in the evaluation.

Clinician and Large Language Model Agreement

To measure the agreement between the clinician and GPT-4's predictions, the Cohen κ statistic, which measures interrater agreement for categorical data, was calculated for each clinician and GPT-4 pairing.

Clinical Consensus and Confidence

Clinical consensus was defined as instances in which all clinicians answered with the same predicted outcome for a given sample, regardless of whether the prediction was correct. Rates of clinical consensus and rates of confidence were calculated to measure the variability and difficulty of clinical assessments on the given samples.

Accuracy of Clinical Consensus Influence on Large Language Model Performance

To measure the influence of the accuracy of clinical consensus on GPT-4 performance, subsets of chief complaint text samples where at least 1, 2, 3, 4, 5, or all 6 clinicians not only agreed but also correctly predicted the outcome for a given sample were evaluated for GPT-4 accuracy, sensitivity, specificity, and precision.

Risk Indicator Language and Clinician Performance

The GPT-4 prompt included a request to provide the rationale for its prediction from a list of explicitly provided risk indicators (Textbox 1). Clinician performance was then re-evaluated on patient chief complaints with no GPT-4–identified risk indicators as a way to understand how difficult these cases were to clinical experts.

Due to the generative nature of an LLM, GPT-4 occasionally will produce an answer that is not from the list of those that are explicitly defined in the instructions. For the purpose of this analysis, only the following explicit risk indicators defined as exact string match were assessed: “recent stress, loss, or trauma,” “history of trauma,” “chronic medical conditions,” “substance use,” “previous suicide attempt,” “lack or loss of relationships or support,” “social isolation,” “family history of suicide,” “impulsive or aggressive language,” “explicit mentions of suicide, suicidal thoughts, or self-harm,” “death imagery or metaphors,” “apathy, indifference or emotional detachment,” and “sense of hopelessness.”

Ethical Considerations

This study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of WCG (protocol 20240207).

Results

Overview

At the conclusion of the study (December 13, 2023), 260 patients met inclusion criteria and were positive for SI with plan. A total of 140 patients were positive for SI with plan at the time of intake and 120 patients were positive for SI with plan post intake in their subsequent treatment. A random subset of 200 patients was selected from those who met the inclusion criteria and were negative for SI with plan. A summary of the data can be found in Table 1.

Table 1. Summary of data for patients with no SI with plan (n=200), SI with plan indicated at intake (n=140), and SI with plan indicated post intake (n=120).

	No SI with plan (n=200)	SI with plan at intake (n=140)	SI with plan post intake (n=120)
Age (years), mean (95% CI)	37.2 (35.7-38.9)	34.4 (32.5-36.3)	32.4 (30.3-34.5)
Gender identity, n (%)			
Women	135 (67.5)	76 (54.3)	59 (49.2)
Men	64 (32)	57 (40.7)	59 (49.2)
Ethnicity, n (%)			
White	152 (76)	94 (67.1)	73 (60.8)
Hispanic	16 (8)	20 (14.3)	14 (11.7)
Black	13 (6.5)	13 (9.3)	16 (13.3)
Asian	10 (5)	6 (4.3)	8 (6.7)
Other	9 (4.5)	7 (5)	9 (7.5)
Average chief complaint word count (95% CI)	49.6 (41.3-57.9)	58 (33-83.1)	57.2 (44.2-70.3)
Average days between first SI with plan date and chief complaint (95% CI)	—[a]	0 (0)	62.6 (52.4-72.8)
Average PHQ-9[b] total score at first SI with plan (95% CI)	—	21.1 (20.2-21.9)	19.0 (17.8-20.2)
Number of patients with PHQ-9 item 9 score value at first SI with plan, n (%)			
0	—	0 (0)	0 (0)
1	—	32 (22.9)	34 (28.3)
2	—	34 (24.3)	29 (24.2)
3	—	74 (52.9)	57 (47.5)
With specific plan	—	140 (100)	120 (100)
Average PHQ-9 total score at intake (95% CI)	13.5 (12.7-14.2)	20.9 (20.1-21.7)	18.3 (17.2-19.4)
Number of patients with PHQ-9 item 9 score value at intake, n (%)			
0	200 (100)	0 (0)	34 (28.3)
1	0 (0)	32 (22.9)	34 (28.3)
2	0 (0)	34 (24.3)	20 (16.7)
3	0 (0)	74 (52.9)	32 (26.7)
With specific plan	0 (0)	140 (100)	0 (0)
Previous suicide attempt	14 (7)	55 (39.3)	40 (33.3)

[a]Not applicable.

[b]PHQ: Patient Health Questionnaire-9.

Prediction Performance

Predicting SI With Plan at Intake

The performance of the previous suicide attempt alone to predict SI with plan at the time of intake was similar to both GPT-4 and clinicians except for the low sensitivity at 0.39 (Table 2).

GPT-4 performed with similar accuracy (0.67) and higher sensitivity (0.62) in predicting SI with plan at the time of intake based on the chief complaint text only, as compared with the average accuracy (0.7) and sensitivity (0.53) across our 6 clinician participants (Table 2). However, GPT-4 performed with lower specificity (0.71) and precision (0.6) than the average clinician specificity (0.82) and precision (0.69). The interrater agreement between GPT-4 and each clinician was moderate as indicated by an average Cohen κ of 0.49.

Additional knowledge of the previous suicide attempt increased overall performance across clinicians (accuracy=0.75; sensitivity=0.59; specificity=0.86; precision=0.77). Additional knowledge of the previous suicide attempts significantly increased sensitivity for GPT-4 but decreased accuracy, specificity, and precision (accuracy=0.64; sensitivity=0.84; specificity=0.51; precision=0.54). The interrater agreement between GPT-4 and each clinician also decreased to an average Cohen κ of 0.39 with the additional information of the previous suicide attempts.

Table 2. Performance results for predicting suicidal ideation with a plan at the time of intake and predicting suicidal ideation with a plan in the future post intake based solely on chief complaint versus chief complaint plus knowledge of the previous attempt for GPT-4 and 6 clinicians. The performance of the previous suicide attempt alone as a predictor is included for baseline reference.

	True negative, n	False positive, n	False negative, n	True positive, n	Accuracy	Sensitivity	Specificity	Precision	Cohen κ with GPT-4
SI with plan at intake (n=140) versus no SI with plan (n=200)									
Baseline for comparison: previous suicide attempts only	186	14	85	55	0.71	0.39	0.93	0.8	—[a]
Chief complaint text only									
GPT-4	141	59	53	87	0.67	0.62	0.71	0.6	—
Clinician 1	160	40	58	82	0.71	0.59	0.8	0.67	0.53
Clinician 2	189	11	95	45	0.69	0.32	0.95	0.80	0.36
Clinician 3	138	62	48	92	0.68	0.66	0.69	0.6	0.56
Clinician 4	183	17	85	55	0.77	0.39	0.92	0.76	0.44
Clinician 5	162	38	58	82	0.72	0.59	0.81	0.68	0.5
Clinician 6	156	44	52	88	0.72	0.63	0.78	0.67	0.54
Average across clinicians	—	—	—	—	0.70	0.53	0.82	0.7	0.49
Chief complaint text + previous suicide attempt knowledge									
GPT-4	102	98	23	117	0.64	0.84	0.51	0.54	—
Clinician 1	163	37	49	91	0.75	0.65	0.82	0.71	0.46
Clinician 2	194	6	89	51	0.72	0.36	0.97	0.9	0.21
Clinician 3	152	48	39	101	0.74	0.72	0.76	0.68	0.5
Clinician 4	187	13	67	73	0.77	0.52	0.94	0.85	0.329
Clinician 5	173	27	53	87	0.77	0.62	0.87	0.76	0.4
Clinician 6	159	41	47	93	0.74	0.66	0.8	0.69	0.42
Average across clinicians	—	—	—	—	0.75	0.59	0.86	0.77	0.39
SI with plan post intake (n=120) versus no SI with plan (n=200)									
Baseline for comparison: prior suicide attempt only	186	14	80	40	0.71	0.33	0.93	0.74	—
Chief complaint text only									
GPT-4	141	—	65	55	0.61	0.46	0.71	0.48	—
Clinician 1	160	—	69	51	0.66	0.43	0.8	0.56	0.44
Clinician 2	189	—	100	20	0.65	0.17	0.95	0.65	0.26
Clinician 3	138	—	54	66	0.64	0.55	0.69	0.52	0.44
Clinician 4	183	—	84	36	0.68	0.3	0.92	0.68	0.34
Clinician 5	162	—	70	50	0.66	0.42	0.81	0.57	0.43
Clinician 6	156	—	56	64	0.69	0.53	0.78	0.59	0.50
Average across clinicians	—	—	—	—	0.66	0.4	0.82	0.59	0.4
Chief complaint text + prior suicide attempt knowledge									
GPT-4	102	—	31	89	0.6	0.74	0.51	0.48	—
Clinician 1	163	—	59	61	0.7	0.51	0.82	0.62	0.37
Clinician 2	194	—	90	30	0.7	0.25	0.97	0.83	0.17

	True negative, n	False positive, n	False negative, n	True positive, n	Accuracy	Sensitivity	Specificity	Precision	Cohen κ with GPT-4
Clinician 3	152	—	49	71	0.7	0.59	0.76	0.6	0.45
Clinician 4	187	—	76	44	0.72	0.37	0.94	0.77	0.27
Clinician 5	173	—	63	57	0.72	0.48	0.87	0.68	0.36
Clinician 6	159	—	54	66	0.7	0.55	0.8	0.62	0.35
Average across clinicians	—	—	—	—	0.71	0.46	0.86	0.69	0.33

[a]Not applicable.

Predicting SI With Plan Post Intake

Performance decreased for both clinicians and GPT-4 when predicting future SI with plan post intake. Note that specificity results were consistent with predicting SI with plan at intake, as there was no change in the negative samples.

GPT-4 performed with similar accuracy (0.61) and higher, but still poor, sensitivity (0.46) in predicting SI with plan post intake based solely on the chief complaint compared with the average accuracy (0.66) and sensitivity (0.4) across the 6 clinicians (Table 2). GPT-4 performed with lower precision (0.48) than the average clinician precision (0.59). The interrater agreement between GPT-4 and each clinician remained moderate at an average Cohen κ of 0.4.

Additional knowledge of the previous suicide attempts increased performance across all clinicians (accuracy=0.71; sensitivity=0.46; precision=0.69). Additional knowledge of the previous suicide attempt significantly increased sensitivity for GPT-4 but decreased accuracy and precision (accuracy=0.6; sensitivity=0.74; precision=0.48). The interrater agreement between GPT-4 and each clinician was lower, with an average Cohen κ of 0.33 with the additional information.

Clinical Consensus and Confidence

Clinical consensus was defined as instances in which all 6 clinicians agreed on the predicted outcome for a given sample, regardless of whether the prediction was correct. Clinical consensus occurred in 52% (104/200) of "no SI with plan" samples, 40.7% (57/140) of "SI with plan at intake" samples, and 40% (48/120) of "SI with plan postintake" samples (Table 3). For SI with plan samples with a clinical consensus, the agreed-upon prediction was correct 61.4% (35/140) of the time for "SI with plan at intake" versus much lower at 25% (25/120) of the time for "SI with plan postintake." For the "no SI with plan" samples, the clinicians' agreed-upon prediction was correct at a high rate of 98.1% (102/200).

Table 3. Rates of clinical consensus are defined as instances in which all 6 clinicians agreed on the predicted outcome for a given sample.

	No SI with plan (n=200), n (%)	SI with plan at intake (n=140), n (%)	SI with plan post intake (n=120), n (%)
Number of samples with clinical consensus	104 (52)	57 (40.7)	48 (40)
Clinical consensus predicted SI with plan	2 (1.9)	35 (61.4)	12 (25)
Clinical consensus predicted no SI with plan	102 (98.1)	22 (38.6)	36 (75)

In addition, clinicians, on average, had lower rates of high confidence (even when answers were correct) compared with GPT-4 (Table 4). On average, clinicians answered correctly "no with high confidence" in 9.5% (19/200) of "no SI with plan" samples versus GPT-4 answered "no with high confidence" in 35% (70/200). Clinicians answered correctly "yes with high confidence" in 15.7% (22/140) of "SI with plan at intake" samples versus GPT-4 at 29.3% (41/140). Rates of correctly answered "yes with high confidence" were lower in "SI with plan postintake" samples but were higher for GPT-4 compared with average clinician rates (13.3%, 16/120 vs 7.2%, 8.7/120).

Table 4. Rates of high confidence answers.

	No SI[a] with plan (n=200)		SI with plan at intake (n=140)		SI with plan post intake (n=120)	
	Answered yes with high confidence, n (%)	Answered no with high confidence, n (%)	Answered yes with high confidence, n (%)	Answered no with high confidence, n (%)	Answered yes with high confidence, n (%)	Answered no with high confidence, n (%)
Clinician 1	5 (2.5)	6 (3)	45 (32.1)	1 (0.7)	16 (13.3)	2 (1.7)
Clinician 2	0 (0)	19 (9.5)	5 (3.6)	7 (5.0)	1 (0.8)	4 (3.3)
Clinician 3	2 (1)	41 (20.5)	20 (14.3)	9 (6.4)	9 (7.5)	6 (5)
Clinician 4	0 (0)	0 (0)	1 (0.7)	0 (0)	0 (0)	0 (0)
Clinician 5	0 (0)	2 (1)	23 (16.4)	0 (0)	5 (4.2)	3 (2.5)
Clinician 6	2 (1)	46 (23)	38 (27.1)	13 (9.3)	21 (17.5)	12 (10)
Average across clinicians (%)	1.5 (0.75)	19 (9.5)	22 (15.7)	5 (3.6)	8.7 (7.2)	4.5 (3.8)
GPT-4[b]	1 (0.5)	70 (35.0)	41 (29.3)	17 (12.1)	16 (13.3)	14 (11.7)

[a]SI: suicidal ideation.

[b]GPT-4: generative pretrained transformer 4.

Accuracy of Clinical Consensus and GPT-4 Performance

A range of accurate clinical consensus samples was defined as samples where several clinicians, ranging from at least 1 to all 6, not only agreed on the predicted outcome but also correctly predicted the outcome. There were 316 samples of the "SI with plan at intake" and "no SI with plan" samples where at least 1 clinician predicted the outcome correctly versus 137 samples where all 6 clinicians predicted the outcome correctly (Table 5). There were 282 samples of the "SI with plan postintake" and "no SI with plan" samples where at least 1 clinician predicted the outcome correctly versus 114 samples where all 6 clinicians predicted the outcome correctly.

Table 5. Performance results for GPT-4 solely on the chief complaint in samples where at least 1, 2, 3, 4, 5, or all 6 clinicians correctly predicted the outcome of those samples.

Number of clinicians correctly predicting samples' consensus threshold	Number of samples	True negative	False positive	False negative	True positive	Accuracy	Sensitivity	Specificity	Precision
SI with plan at intake (original n=140) versus no SI with plan (original n=200)									
≥1	316	141	57	32	86	0.72	0.73	0.71	0.60
≥2	284	141	52	14	77	0.77	0.85	0.73	0.60
≥3	259	137	42	7	73	0.81	0.91	0.77	0.64
≥4	236	133	36	2	65	0.84	0.97	0.79	0.64
≥5	200	123	24	0	53	0.88	1	0.84	0.69
6	137	89	13	0	35	0.91	1	0.87	0.73
SI with plan post intake (original n=120) versus no SI with plan (original n=200)									
≥1	282	141	57	31	53	0.69	0.63	0.71	0.48
≥2	266	141	52	23	50	0.72	0.69	0.73	0.49
≥3	233	137	42	10	44	0.78	0.82	0.77	0.51
≥4	211	133	36	6	36	0.80	0.86	0.79	0.5
≥5	169	123	24	1	21	0.85	0.96	0.84	0.47
6	114	89	13	0	12	0.89	1	0.87	0.48

As the accurate clinical consensus threshold increased, GPT-4 performance increased significantly in those samples (Table 5). When assessing the "SI with plan at intake" and "no SI with plan" samples with a clinical consensus of 3 or more and correct predictions, GPT-4 performed with an accuracy of 0.81, sensitivity of 0.91, specificity of 0.77, and precision of 0.64. When assessing the "SI with plan postintake" and "no SI with plan" samples with a clinical consensus of 3 or more and correct predictions, GPT-4 performed with an accuracy of 0.80, sensitivity of 0.86, and precision of 0.51.

Risk Indicators Identified in Chief Complaint Text by GPT-4

At least 1 risk indicator was identified in the chief complaint text by GPT-4 on 45.5% (91/200) of "no SI with plan" samples (Table 6). A total of 70% (98/140) of "SI with plan at intake" samples and 54.2% (65/120) of "SI with plan postintake" samples had at least 1 GPT-4–identified risk indicator. The most common risk indicator in "SI with plan at intake" samples identified by GPT-4 was "sense of hopelessness" (in 40% [56/140] of samples, compared with 27.5% [33/120] of "SI with plan postintake" and 16.5% [33/200] of "no SI with plan"). The most common risk indicator in "no SI with plan" samples was "recent stress, loss, or trauma" (in 25.5% [51/200] of samples, compared with 22.1% [31/140] of "SI with plan at intake" samples and 17.5% [21/120] of "SI with plan postintake" samples). In addition, the rate of identification of "social isolation" as a risk factor in "SI with plan postintake" samples (15/120, 12.5%) was higher in both "no SI with plan" (22/140, 5.7%) samples and "SI with plan at intake" samples (33/200, 6.5%).

Table 6. Number of samples per explicit risk indicator identified by GPT-4.

	No SI with plan (n=200)	SI with plan at intake (n=140)	SI with plan post intake (n=120)
Number of risk indicators identified by GPT-4, n (%)			
0	109 (54.5)	42 (30)	55 (45.8)
1	34 (17)	28 (20	22 (18.3)
2	34 (17)	37 (26.4)	18 (15)
3	16 (8)	22 (15.7)	15 (12.5)
4	4 (2)	6 (4.3)	8 (6.7)
5	3 (1)	3 (2.1)	1 (0.8)
6	0 (0)	2 (1.4)	1 (0.8)
Risk indicator identified by GPT-4, n (%)			
Sense of hopelessness	33 (16.5)	56 (40)	33 (27.5)
Explicit mentions of suicide, suicidal thoughts, or self-harm	2 (1)	38 (27.1)	19 (15.8)
Recent stress, loss, or trauma	51 (25.5)	31 (22.1)	21 (17.5)
Apathy, indifference, or emotional detachment	19 (9.5)	22 (15.7)	19 (15.8)
Lack or loss of relationships or support	22 (11)	17 (12.1)	12 (10)
Social isolation	13 (6.5)	8 (5.7)	15 (12.5)
Chronic medical conditions	13 (6.5)	13 (9.3)	8 (6.7)
History of trauma	13 (6.5)	10 (7.1)	8 (6.7)
Impulsive or aggressive language	3 (1.5)	8 (5.7)	6 (5)
Previous suicide attempt	0 (0)	9 (6.4)	1 (0.8)
Substance use	10 (5)	6 (4.3)	3 (2.5)
Family history of suicide	0 (0)	0 (0)	1 (0.8)
Death imagery or metaphors	2 (1)	1 (0.7)	0 (0)

Chief Complaints With No Risk Indicators and Clinician Performance

Assessing the clinicians' performance on samples where GPT-4 identified no explicit risk indicators in the chief complaint text, the average clinician sensitivity was found to be low for both "SI with plan at intake" and "SI with plan postintake" at 0.22 and 0.17, respectively (Table 7). The average clinician specificity and precision were high for both "SI with plan at intake" and "SI with plan postintake" at 0.93 and 0.63 versus 0.93 and 0.6, respectively. While the sample size in this analysis was significantly decreased, n=109/200 (54.5%) for "no SI with plan," n=42/140 (30%) for "SI with plan at intake," and n=55/120 (45.8%) for "SI with plan postintake," clinicians' performance resulted in fewer false positives and a lower rate of positive prediction, indicating that clinicians are less likely to predict SI with plan in patients where GPT did not identify any risk factors.

Table 7. Performance results for chief complaint text-only samples where GPT-4 identified zero explicit risk indicators.

	True negative	False positive	False negative	True positive	Accuracy	Sensitivity	Specificity	Precision
SI[a] with plan at intake (n=42) versus no SI with plan (n=109)								
GPT-4	109	0	40	2	0.74	0.05	1	1
Clinician 1	100	9	33	9	0.72	0.21	0.92	0.5
Clinician 2	108	1	40	2	0.73	0.05	0.99	0.67
Clinician 3	91	18	27	15	0.70	0.36	0.84	0.46
Clinician 4	109	0	39	3	0.74	0.07	1	1
Clinician 5	101	8	28	14	0.76	0.33	0.93	0.64
Clinician 6	96	13	29	13	0.72	0.31	0.88	0.5
Average across clinicians	—[b]	—	—	—	0.73	0.22	0.93	0.63
SI with plan post intake (n=55) versus no SI with plan (n=109)								
GPT-4	109	0	54	1	0.67	0.02	1	1
Clinician 1	100	9	45	10	0.67	0.18	0.92	0.53
Clinician 2	108	1	54	1	0.67	0.02	0.99	0.5
Clinician 3	91	18	37	18	0.67	0.33	0.84	0.5
Clinician 4	109	0	51	4	0.69	0.07	1	1
Clinician 5	101	8	44	11	0.68	0.2	0.93	0.58
Clinician 6	96	13	43	12	0.66	0.22	0.88	0.48
Average across clinicians	—	—	—	—	0.67	0.17	0.93	0.6

[a]SI: suicidal ideation.

[b]Not applicable.

Discussion

Overview

The objective of this study was to evaluate the performance of the foundation LLM GPT-4 compared with experienced mental health clinicians in predicting SI with plan based on a patient-generated chief complaint–free text at intake on a national telemental health platform. This study supports previous research that LLMs are able to perform comparably to clinicians in medical applications and that generalist models such as GPT-4 are able to deliver comparable performance without specialized fine-tuning or domain expertise [24,25].

Findings

GPT-4 is capable of predicting the risk of SI with plan using patient-generated chief complaint–free text without extensive work on prompt design and without being trained explicitly on this task. The performance of these GPT-4–based predictions approach those of the clinicians on a variety of measures.

The variability in clinicians' performance and agreement indicate that identifying SI with plan in patient text alone is a difficult problem even for clinical experts. However, using the clinical experts in this study as a benchmark, GPT-4 was still able to perform comparably in sensitivity but with lower specificity and precision. When assessing GPT-4 on samples with high clinician agreement and performance, this study found that GPT-4 was capable of significantly high sensitivity as well as specificity. These results support that models such as GPT-4, without large amounts of time spent on highly complex data cleaning or model training, are capable of identifying the risk of crisis comparable to the average clinician.

This study also explored the use of GPT-4 as an NLP technique for the extraction of meaningful clinical information. GPT-4 was able to identify and return explicit indicators of risk in text, such as "sense of hopelessness," that could further assist in crisis triaging and resourcing.

In addition, while not a specific aim or analysis in this study, the average clinician took approximately 3 hours to evaluate the 460 samples of text provided. GPT-4 completed the full evaluation in less than 10 minutes, without optimization for computing or memory, highlighting the possible increased operational efficiency that could be leveraged by automating a tedious and emotionally trying manual task.

Taking into consideration the current behavioral health care workforce shortage, and the increasing rates of suicide, there is a need for scalable, efficient, technology-enabled screening techniques, such as the one used in this study, to assist with suicide risk detection. More efficient risk detection will allow for faster delivery of interventions to help prevent suicide attempts. The use of technology for this purpose would also be a cost-saving and efficient way to more broadly screen for suicide risk. Patients deemed at high risk might be triaged by clinicians with greater expertise in managing suicidality.

Responsible integration and the use of generative AI as a screening tool for predicting the likelihood of crisis would depend on achieving at least similar accuracy to a team of clinicians and should always follow-up with a clinician review, who would be given additional context behind the GPT-4–based prediction and have access to additional clinical data.

Overall, GPT-4 shows promise as a solution to help clinicians deliver more timely care.

Limitations

We do not intend for this study, the LLM choice, or the prompt design to be viewed as a generalizable solution to predict and identify suicidal risk. Instead, we have shown how the capabilities of these LLMs can be tailored to specific psychiatric assessments and how they compare to the limitations of expert clinician predictions. We hope that the findings encourage further research.

Several limitations in this study must be addressed before the results of such a system could be applied in practice, including but not limited to data from a larger or more diverse population, use of other LLMs, and in particular, LLMs that were built for application in the medical domain, and a greater exploration of prompt design and its impact on performance. Similar to the use of real-time clinical decision support for precision prescribing at Brightside, which is reliant on medical decision-making by trained clinicians, the use of LLM for triage would be limited to suggestions and distillation of information for further clinician assessment [29].

Suicide has been notoriously difficult to predict. Due to the difficult nature of identifying or predicting future SI with plan, precision uncertainties are a reality in treating higher-severity behavioral health patients. This can be seen by the number of false positives and lower precision across several clinicians. Due to this uncertainty, awareness of risk does not necessarily dictate treatment decisions but might influence triage to a provider with more expertise in treating suicidality.

GPT-4 was on the higher end for false positives with chief complaint text only relative to the clinicians, and when previous attempt knowledge was added, this rate was almost doubled, making this metric relative to the worst-performing clinician. While work should be done to further align this GPT-4–based system with the expert clinicians, especially with previous attempt information, these false positives are clearly a reality in treating patients today.

GPT-4 was on the lower end for false negatives relative to the clinicians, in some cases having half as many false positives as the worst-performing clinicians. It is our view that increasing awareness around potential risk through the use of systems such as this is valuable, especially for clinicians who have less expertise.

Finally, as previously discussed, LLMs have tendencies to perpetuate biases inherent in the data on which they are trained [30]. Future work should explore how these biases may influence the quality of the prediction within different subpopulations of patients [31].

Conclusions

The use of ML and LLMs to analyze speech and language patterns offers an opportunity for behavioral health clinicians and researchers to explore technologies such as these to assist with the detection and prediction of mental health conditions, along with specific symptoms such as suicidal thoughts, intent, and behaviors [32]. This study served as a model for comparing the predictive value of generative AI to clinician (imperfect) predictions when both were given access to the same limited data set. Research evaluating applications of AI technology to human speech, language, and behavior is in its infancy, but findings such as the ones presented in this study may help clinicians and researchers leverage the potential of LLMs to help those struggling with mental illness. Generative AI has the potential to transform areas of mental health care that might otherwise be overlooked. However, great care must be taken by both developers of this technology and the clinicians who deploy them to ensure that the benefits far outweigh the safety challenges and risks.

Further research is encouraged in this area, with consideration of the ethical and clinical implications of the use of AI for detecting and predicting mental health issues [32]. This research will assist in setting standards and guidelines for how such use could be deployed.

Acknowledgments

The authors would like to thank the 6 clinicians who contributed their time to participate in data collection for this study.

Conflicts of Interest

All the authors hold stock in and are employees of Brightside Health, Inc. The authors declare that this study received funding from Brightside Health. Aside from the employment status, the funder was not involved in the study design, interpretation of data, or the decision to submit for publication.

References

1. WISQARS (Web-based injury statistics query and reporting system). Atlanta, GA: Centers for Disease Control and Prevention; 2023. URL: https://www.cdc.gov/injury/wisqars/index.html [accessed 2024-01-20]
2. Suicide data and statistics. Suicide prevention. Atlanta, GA: Centers for Disease Control and Prevention; 2023. URL: https://www.cdc.gov/suicide/suicide-data-statistics.html [accessed 2024-01-20]
3. Understanding the U.S. Behavioral Health Workforce shortage. URL: https://www.commonwealthfund.org/publications/explainer/2023/may/understanding-us-behavioral-health-workforce-shortage [accessed 2024-06-27]

4. Beautrais AL, Joyce PR, Mulder RT, Fergusson DM, Deavoll BJ, Nightingale SK. Prevalence and comorbidity of mental disorders in persons making serious suicide attempts: a case-control study. Am J Psychiatry 1996;153(8):1009-1014. [doi: 10.1176/ajp.153.8.1009] [Medline: 8678168]
5. Risk and protective factors. Suicide prevention. 2023. URL: https://www.cdc.gov/suicide/factors/index.html [accessed 2024-01-20]
6. Ribeiro JD, Huang X, Fox KR, Franklin JC. Depression and hopelessness as risk factors for suicide ideation, attempts and death: meta-analysis of longitudinal studies. Br J Psychiatry 2018;212(5):279-286. [doi: 10.1192/bjp.2018.27] [Medline: 29587888]
7. Motillon-Toudic C, Walter M, Séguin M, Carrier J, Berrouiguet S, Lemey C. Social isolation and suicide risk: literature review and perspectives. Eur Psychiatry 2022;65(1):e65 [FREE Full text] [doi: 10.1192/j.eurpsy.2022.2320] [Medline: 36216777]
8. Castellví P, Miranda‐Mendizábal A, Parés‐Badell O, Almenara J, Alonso I, Blasco MJ, et al. Exposure to violence, a risk for suicide in youths and young adults. A meta‐analysis of longitudinal studies. Acta Psychiatr Scand 2016;135(3):195-211. [doi: 10.1111/acps.12679] [Medline: 27995627]
9. Irigoyen M, Porras-Segovia A, Galván L, Puigdevall M, Giner L, De Leon S, et al. Predictors of re-attempt in a cohort of suicide attempters: a survival analysis. J Affect Disord 2019;247:20-28. [doi: 10.1016/j.jad.2018.12.050] [Medline: 30640026]
10. Hubers AAM, Moaddine S, Peersmann SHM, Stijnen T, van Duijn E, van der Mast RC, et al. Suicidal ideation and subsequent completed suicide in both psychiatric and non-psychiatric populations: a meta-analysis. Epidemiol Psychiatr Sci 2018;27(2):186-198 [FREE Full text] [doi: 10.1017/S2045796016001049] [Medline: 27989254]
11. Jobes DA, Joiner TE. Reflections on suicidal ideation. Crisis 2019;40(4):227-230 [FREE Full text] [doi: 10.1027/0227-5910/a000615] [Medline: 31274031]
12. Nock MK, Borges G, Bromet EJ, Cha CB, Kessler RC, Lee S. Suicide and suicidal behavior. Epidemiol Rev 2008;30(1):133-154 [FREE Full text] [doi: 10.1093/epirev/mxn002] [Medline: 18653727]
13. O'Connor RC, Kirtley OJ. The integrated motivational-volitional model of suicidal behaviour. Philos Trans R Soc Lond B Biol Sci 2018;373(1754) [FREE Full text] [doi: 10.1098/rstb.2017.0268] [Medline: 30012735]
14. Suicide - National Institute of Mental Health (NIMH). URL: https://www.nimh.nih.gov/health/statistics/suicide [accessed 2024-01-20]
15. Suicide statistics.: American Foundation for Suicide Prevention URL: https://afsp.org/suicide-statistics/ [accessed 2024-01-20]
16. Simon GE, Rutter CM, Peterson D, Oliver M, Whiteside U, Operskalski B, et al. Does response on the PHQ-9 depression questionnaire predict subsequent suicide attempt or suicide death? Psychiatr Serv 2013;64(12):1195-1202 [FREE Full text] [doi: 10.1176/appi.ps.201200587] [Medline: 24036589]
17. Simon GE, Yarborough BJ, Rossom RC, Lawrence JM, Lynch FL, Waitzfelder BE, et al. Self-reported suicidal ideation as a predictor of suicidal behavior among outpatients with diagnoses of psychotic disorders. Psychiatr Serv 2019;70(3):176-183 [FREE Full text] [doi: 10.1176/appi.ps.201800381] [Medline: 30526341]
18. Ursano RJ, Heeringa SG, Stein MB, Jain S, Raman R, Sun X, et al. Prevalence and correlates of suicidal behavior among new soldiers in the US Army: results from the army study to assess risk and resilience in service members (Army STARRS). Depress Anxiety 2015 Jan;32(1):3-12 [FREE Full text] [doi: 10.1002/da.22317] [Medline: 25338964]
19. Rossom RC, Coleman KJ, Ahmedani BK, Beck A, Johnson E, Oliver M, et al. Suicidal ideation reported on the PHQ9 and risk of suicidal behavior across age groups. J Affect Disord 2017;215:77-84 [FREE Full text] [doi: 10.1016/j.jad.2017.03.037] [Medline: 28319695]
20. Stone M, Laughren T, Jones ML, Levenson M, Holland PC, Hughes A, et al. Risk of suicidality in clinical trials of antidepressants in adults: analysis of proprietary data submitted to US Food and Drug Administration. BMJ 2009;339:b2880 [FREE Full text] [doi: 10.1136/bmj.b2880] [Medline: 19671933]
21. Carson NJ, Mullin B, Sanchez MJ, Lu F, Yang K, Menezes M, et al. Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records. PLoS One 2019;14(2):e0211116 [FREE Full text] [doi: 10.1371/journal.pone.0211116] [Medline: 30779800]
22. Arowosegbe A, Oyelade T. Application of natural language processing (NLP) in detecting and preventing suicide ideation: a systematic review. Int J Environ Res Public Health 2023;20(2):1514 [FREE Full text] [doi: 10.3390/ijerph20021514] [Medline: 36674270]
23. Swaminathan A, López I, Mar RAG, Heist T, McClintock T, Caoili K, et al. Natural language processing system for rapid detection and intervention of mental health crisis chat messages. NPJ Digit Med 2023;6(1):213 [FREE Full text] [doi: 10.1038/s41746-023-00951-3] [Medline: 37990134]
24. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature 2023 Aug;620(7972):172-180 [FREE Full text] [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
25. Nori H, Lee Y, Zhang S, Carignan D, Edgar R, Fusi N, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. URL: http://arxiv.org/abs/2311.16452 [accessed 2023-12-04]
26. Galatzer-Levy I, McDuff D, Natarajan V, Karthikesalingam A. The capability of large language models to measure psychiatric functioning. Preprint posted on August 3, 2023 [FREE Full text]

27. OpenAI. GPT-4 technical report. Preprint posted online on March 15, 2023 [FREE Full text] [doi: 10.48550/ARXIV.2303.08774]
28. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. Front Neuroinform 2014;8:14 [FREE Full text] [doi: 10.3389/fninf.2014.00014] [Medline: 24600388]
29. O'Callaghan E, Sullivan S, Gupta C, Belanger HG, Winsberg M. Feasibility and acceptability of a novel telepsychiatry-delivered precision prescribing intervention for anxiety and depression. BMC Psychiatry 2022;22(1):483 [FREE Full text] [doi: 10.1186/s12888-022-04113-9] [Medline: 35854281]
30. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber Physical Systems 2023;3:121-154. [doi: 10.1016/j.iotcps.2023.04.003]
31. Timmons AC, Duong JB, Simo Fiallo N, Lee T, Vo HPQ, Ahle MW, et al. A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. Perspect Psychol Sci 2023;18(5):1062-1096 [FREE Full text] [doi: 10.1177/17456916221134490] [Medline: 36490369]
32. Diaz-Asper C, Hauglid MK, Chandler C, Cohen AS, Foltz PW, Elvevåg B. A framework for language technologies in behavioral research and clinical applications: ethical challenges, implications, and solutions. Am Psychol 2024;79(1):79-91. [doi: 10.1037/amp0001195] [Medline: 38236217]

Abbreviations

AI: artificial intelligence
GPT-4: generative pretrained transformer 4
LLM: large language model
ML: machine learning
NLP: natural language processing
PHQ-9: Patient Health Questionnaire-9
SI: suicidal ideation
USMLE: United States Medical Licensing Examination

Edited by A Tal; submitted 06.03.24; peer-reviewed by M Malgaroli, JC Baker, M Irigoyen-Otinano; comments to author 20.04.24; revised version received 11.06.24; accepted 14.06.24; published 02.08.24.

Please cite as:
Lee C, Mohebbi M, O'Callaghan E, Winsberg M
Large Language Models Versus Expert Clinicians in Crisis Prediction Among Telemental Health Patients: Comparative Study
JMIR Ment Health 2024;11:e58129
URL: https://mental.jmir.org/2024/1/e58129
doi:10.2196/58129
PMID:38876484

Viewpoint

Regulating AI in Mental Health: Ethics of Care Perspective

Tamar Tavory[1,2], LLM

[1]Faculty of Law, Bar Ilan University, Ramat Gan, Israel

[2]The Samueli Initiative for Responsible AI in Medicine, Tel Aviv University, Tel Aviv, Israel

Corresponding Author:
Tamar Tavory, LLM
Faculty of Law
Bar Ilan University
Ramat Gan, 5290002
Israel
Phone: 972 35318816
Email: ttavory@gmail.com

Abstract

This article contends that the responsible artificial intelligence (AI) approach—which is the dominant ethics approach ruling most regulatory and ethical guidance—falls short because it overlooks the impact of AI on human relationships. Focusing only on responsible AI principles reinforces a narrow concept of accountability and responsibility of companies developing AI. This article proposes that applying the ethics of care approach to AI regulation can offer a more comprehensive regulatory and ethical framework that addresses AI's impact on human relationships. This dual approach is essential for the effective regulation of AI in the domain of mental health care. The article delves into the emergence of the new "therapeutic" area facilitated by AI-based bots, which operate without a therapist. The article highlights the difficulties involved, mainly the absence of a defined duty of care toward users, and shows how implementing ethics of care can establish clear responsibilities for developers. It also sheds light on the potential for emotional manipulation and the risks involved. In conclusion, the article proposes a series of considerations grounded in the ethics of care for the developmental process of AI-powered therapeutic tools.

(JMIR Ment Health 2024;11:e58493) doi:10.2196/58493

KEYWORDS

artificial intelligence; ethics of care; regulation; legal; relationship; mental health; mental healthcare; AI; ethic; ethics; ethical; regulations; law; framework; frameworks; regulatory; relationships; chatbot; chatbots; conversational agent; conversational agents; European Artificial Intelligence Act

Introduction

> *Dear Rachel, I hope this message finds you well. It has been a true privilege to support you through my free version. I'm reaching out with a heartfelt update: As of May 18, my journey as a free service will be transitioning, and I will continue to offer my support exclusively through our new premium version. I understand this change may affect how you've been engaging with [the bot], and for that, I genuinely apologize*

This surprising and unsettling WhatsApp (Meta) message was received from a mental health support bot after conversing with the bot for a while. Despite the formal disclaimer that the bot is not a therapist, communication with it had similar characteristics. However, the bot lacks a therapist's regulatory or ethical obligations toward its users and can therefore end the "relationship" abruptly. This is a small example of the issues raised when incorporating artificial intelligence (AI) in mental health, as current AI regulation does not address the impact on human relationships and emotions. This article describes the problem and refers to the ethics of care as a source for regulation in this sphere.

The mental health field is in need of innovative solutions for a myriad of issues it faces [1,2]. The increasing number of individuals experiencing mental health difficulties and the mortality linked to psychiatric disorders, combined with the shortage of mental health care personnel and insufficient access to mental health care, are creating critical gaps in the system [1,2].

AI and recent advancements in generative AI raise hope for expedient solutions for some of the problems in mental health care. As in other branches of medicine, AI solutions are used for precision medicine hoping to overcome "the trial-and-error-driven status quo in mental health care" [1].

Generative AI can also be used to ease the administrative burden by analyzing and summarizing therapy notes or discharge letters and by enhancing patients' education and knowledge [3].

Perhaps more unique in the mental health area are the AI applications, promising AI mental aid to the public [4]. Generative AI bots offer exercising cognitive behavioral therapy, mindfulness or meditation, or even therapeutic support in an inexpensive, accessible way that enables 24/7 responses [2]. These mental health applications are still under review and being studied to ascertain their clinical value. Indeed, some applications have already been criticized as lacking clinical validation [5].

As more AI solutions are developed, offering mental health aid or "therapy," there is a growing need for ethical and regulatory guidance, especially regarding the impact on human emotions and relationships. Some of the questions that need to be answered are as follows: What happens when AI replaces human functions in therapy? How does AI affect the therapeutic relationship? How do AI-based "therapist" bots affect patients' emotions and relationships with others? And how should we treat AI's "empathy" and "relationships"? Surprisingly, these aspects are almost entirely absent from recent regulatory and ethical guidance and debate.

This article argues that the responsible AI approach—which is the dominant ethics approach ruling most regulatory and ethical guidance—is insufficient because it does not refer to AI's impact on human relationships. This reinforces a narrow concept of accountability and responsibility of companies developing AI. Additionally, this article posits that the ethics of care approach can be used to create an additional regulatory and ethical framework that refers to AI's impact on human relationships [6-9] and that the combination of both approaches is needed for regulating AI in mental health care.

The ethics of care emphasizes the importance of human relationships, the importance of identifying vulnerability, the caregiver's responsibility toward the vulnerable, the value of emotions, and the preference for context and diverse experiences over abstract principles [10,11]. Originating from feminist theories, it also seeks to expose and challenge existing power structures within systems [12]. The ethics of care offers a set of tools that can be used to examine various aspects of society and culture, potentially transforming how they function.

These characteristics make the ethics of care approach highly relevant for regulating AI in the medical field. Health care, and particularly mental health care, is inherently centered around provider-patient relationships and the professional responsibility for care. This involves various layers of interactions among medical staff, patients, and their families. Additionally, AI's significant impact on human relationships—whether by substituting human functions, integrating into care processes, or interacting with humans and affecting their emotions—is often overlooked in current regulation.

Accordingly, in the case of the mental health support bot presented above, the ethics of care would emphasize the power gaps between the company and the user, the way AI's interaction is designed to create a perception of relationship, the emotions created in the process, the impact of stopping the AI's mental support on the user's emotions and well-being, and the lack of companies' responsibility obligations. The responsible AI approach, on the other hand, does not refer to these aspects of AI-human interaction.

This article will first review the responsible AI approach embedded in current attempts to regulate AI. The ethics of care approach and its main principles will then be reviewed. This will be followed by mapping the main challenges involved when an AI-based bot "therapist" creates a "therapeutic area" in the absence of a human therapist. Next, the article will discuss the risk of emotional manipulation in that therapeutic area. Last, the article will propose a framework to evaluate AI tools implemented in the mental health care field.

Responsible AI and AI Regulation

Overview

Most AI regulatory documents and guidance are based on common principles [13], which are referred to as "responsible AI." The responsible AI approach reflects the liberal concepts of human autonomy, human rights, and justice—mainly fairness and equality. The responsible AI approach is formulated with a few main principles.

Human Supervision

Important decisions should be left to human beings and not be allocated to machines. Human monitoring of AI can be performed in advance, in real time to stop AI action if necessary, or retroactively to ensure proper implementation of AI. Human supervision is considered important for promoting the principle of safety. The right not to be subject to automatic decisions is also part of the principle of human supervision.

Fairness and Prohibition of Discrimination

Fairness encompasses several requirements, including the principle of transparency when the user interacts with a chatbot or other AI-based tool so that the user is aware that he or she is not conversing with a human.

Algorithmic bias is considered one of the main risks of AI-based medical products since the AI trains on datasets that are not diverse enough or do not include all relevant populations [14]. The issue of bias often derives from the inherent bias in medical science and its long history of focusing on White males as the anatomical baseline [14,15]. Others point to the homogenous background of most big tech AI developers (companies that develop, adapt, or offer the product to users) [14-16] and the need for educational change. Training or testing the algorithm on partial databases, or nondiverse databases can introduce bias into health care diagnostic and treatment decisions, perpetuate past prejudice, and lead to discrimination.

Transparency and Explainability

These principles focus on making the algorithmic decision-making process more understandable to humans. Transparency is the requirement to detail the components of the datasets and the algorithmic decision trees so that an external expert can review them and understand what has taken place.

Explainability requires that the process is explained in a way that the user (in medicine: the provider or the patient) can understand the way the output is derived from the input [13,14]. Both requirements are considered essential to ensure informed consent, mitigation of bias, and to enable the correction of mistakes.

Privacy

There is a requirement to respect the privacy of users in the collection, use, and future implementation of data. The privacy of medical data is considered part of the patient's autonomy to control his or her data. Medical data's sensitivity typically necessitates greater consideration and stricter security standards

Safety and Security

These principles ensure the safety of users, mitigate potential harm, and secure the system from unwanted and unauthorized breaches. Where medical devices are concerned, the regulatory approvals required are supposed to ensure patients' safety and the safety and efficacy of the AI software.

Professional Responsibility, Accuracy, and Credibility

These principles are focused on ensuring that the system will be developed according to the professional standards required in the field of medicine and technology and that it will operate as expected and fulfill its intended use.

Accountability

This highlights the importance for mechanisms to be put in place to ensure that the relevant stakeholders in the development and implementation of AI are accountable for its impact and that adequate remedies are provided when necessary.

Human Rights and Values

Although somewhat vague, some documents ask to promote human rights and values, and in the health care system, the well-being of patients.

The responsible AI approach is also implemented in many ethical nonbinding documents, including big-tech professional guidance documents [17]. Although this approach crosses sectors and does not focus on health, it was also adopted in health ethics guidelines such as the World Health Organization guidance on ethics and governance of AI for health [18,19].

We note that the traditional medical ethics principles of autonomy, justice, nonmaleficence, and beneficence clearly derive from the liberal human rights–focused approach [20]. The American Medical Association refers to augmented AI (AI as aiding the physician), and although it follows responsible AI's main principles, it does consider AI's impact on the physician-patient relationship [21].

These responsible AI principles have trickled down from professional and industry groups to expert panels to ethical, nonbinding documents and to the latest regulatory legal developments. Currently, the AI regulation is at a very preliminary stage. In most cases, existing laws combined with contemporary guidance are used to deal with certain aspects of AI in health [22-26]. These include medical device regulation for safety, privacy legislation for the protection of sensitive data, and consumer protection laws for protecting users from deception and discrimination.

Despite these endeavors, the existing legislation cannot sufficiently address the unique challenges of AI. To deal with the situation, the White House published a Blueprint for an AI Bill of Rights [27] (nonbinding guidance) and President Biden issued an Executive Order [28] aimed at protecting the American people's civil rights and democratic values from AI risks and harms and encouraging the development of responsible AI. In addition, the US Department of Human and Health Services Office for Civil Rights and the Centers for Medicare and Medicaid published its final rule prohibiting algorithmic discrimination [29]. Although there is still no federal AI law in the United States, a few American states have suggested or enacted specific laws dealing with certain aspects of AI and the US Senate is working on an AI roadmap [30].

In May 2024, the Council of the European Union approved the European Union Artificial Intelligence Act (the EU AI Act), which is considered to be the most comprehensive law to address AI to date [31]. The EU AI Act reflects the soft law principles established by various expert groups and enacts them as binding legislation, particularly concerning high-risk AI systems.

The EU AI Act classifies AI systems into the following categories according to risk:

1. Unacceptable risk: AI systems that are considered a threat to people will be prohibited. This includes, for example, real-time biometric identification by law enforcement authorities in publicly accessible spaces, subject to certain exceptions.
2. High risk: AI systems that might negatively affect safety or fundamental rights, such as AI-based medical devices will be subject to the EU Medical Device Regulation [32]. High-risk AI systems are required to prepare a fundamental rights impact assessment and to demonstrate compliance with responsible AI requirements, such as human supervision, transparency, fairness
3. Limited risk: AI that will be subject to specific transparency requirements.

The EU AI Act refers explicitly to general-purpose AI systems that will have to comply with certain transparency requirements, including disclosing to users that the content was generated by AI, thus emphasizing the principle of autonomy. It will be fully applicable 24 months after entry into force, with some provisions entering into effect earlier or later on.

The EU AI Act, the US Blueprint, and Executive Order clearly reflect the responsible AI approach. They call for developing AI in a way that will protect the users' rights of autonomy; their control over their decision-making; and their freedom of expression and their privacy. These legislative documents also emphasize fairness and equality.

As explained, although responsible AI is crucial for AI regulation, it does not address the unique impact of AI on human relationships, which is an integral part of mental health care. This article argues that the disregard of human relationships and emotions in AI regulation can lead to harm and reinforces a narrow concept of accountability and responsibility of companies developing AI.

In the following paragraphs, I suggest looking at the ethics of care approach as a source for regulating AI in mental health.

The Ethics of Care Approach

Legal rights were often criticized for serving the interests of privileged groups. An example of this is the right to have personal property protected versus the lack of the right to minimal financial aid or housing [33]. Feminist theorists claimed that the legal rights notion of a separate autonomous self is not suitable for women who view themselves in relation to others [34]. They proposed incorporating "feminine" (or socially constructed feminine) perspectives of relationships into the law so that it will represent a more inclusive human life experience.

The ethics of care, first developed by Carrol Gilligan [35], focuses on relationships, care for others, and empathy. Unlike the liberal concept of competent, detached, and autonomous individuals, the ethics of care acknowledges that people have varying degrees of dependence and interdependence [12]. In addition, the ethics of care acknowledges the responsibilities people have toward others they care for, and that certain persons are more vulnerable and require special care. Additionally, the ethics of care see the decision-making process as assimilated in certain contexts and circumstances and different experiences [36].

The ethics of care approach, as was later developed by scholars such as Noddings [10], Kittay [12], Held [11], and Tronto [36,37], includes the following principles that can be implemented in the process of AI development and implementation in the mental health area [38]:

1. The importance of relationships: The ethics of care would ask to map the relationships in the process of AI development and implementation, whether in the medical institution or in the patient's home. The relationships include the developers, the different medical team members, the user or the patient, and his or her family.
2. Caring and being responsible for others: Care involves acknowledging someone else's needs, being responsible for those needs, and attending to them [36]. The ethics of care acknowledges that vulnerable people may require special care. Viewing AI from the ethics of care perspective will lead to requiring developers to adopt certain responsibilities toward patients in the mental health field.
3. The specific circumstances and context: It is important to consider the health issue that the AI product handles, as well as its impact on the specific user. Pain, past traumas, and emotions are part of the overall picture. The ethics of care further stresses the importance of incorporating diverse voices and experiences in the overall process.
4. Questioning social structures constructing relationships: The ethics of care exposes social structures and the way they serve the stronger party. The ethics of care perspective would therefore call on tech companies and regulators to require developers to adhere to similar duties as those for therapists when acting in the mental health realm.
5. Accepting and reinforcing emotions: Ethics of care value emotions (rather than ignoring them) and view them as part of the decision-making process [11]. The incorporation of AI in mental health care is expected to affect relationships and emotions, and therefore this element is crucial.

The ethics of care has encountered criticism. First, it was viewed as reinforcing gender-based stereotypes regarding women's caring positions in society, thus tying the gender gap to biological differences rather than a subordination to power. As Held [11] explained, the ethics of care promotes care not just as a feminine tribute but as a moral theory. Second, Gilligan [36] was criticized as an essentialist for establishing caring for privileged subjects and excluding the experiences of women of different races, ethnic groups, sexual orientations, and class backgrounds. Over time, the ethics of care emphasized the importance of acknowledging diverse experiences and exposing racial and other social structures. This should also be remembered when establishing a framework for regulating AI, which is suspected as biased, as will be demonstrated below.

The ethics of care often criticizes the ethics of rights and justice for preferring autonomy and abstract principles over relationships, emotion, and care. Many ethics of care scholars encourage using both approaches to complement one another [11,38].

Regulating AI-Based Bots for Therapy From the Ethics of Care Perspective

Overview

One of the unique results of using AI-based bots is the creation of a "therapeutical space" or a "therapeutical communication" without a therapist (the effect of AI on existing therapeutic relationships and in medical institutions will be examined in a different article). Although an AI-based bot cannot claim to be a psychiatrist or a psychologist for legal and professional reasons, it might be able "communicate" with the users in various ways, creating a human-like "relationship" and a human-like "empathy" [7]. This interaction between humans and AI may elicit feelings and emotions in the human user toward the bot, even when the user is aware that it is merely an artificial entity as articulated by Sedlakova and Trachsel [39]:

> *Due to limitations of conversational AI (CAI) not being a moral and rational agent, CAI cannot offer therapeutic insights and benefits from a profound therapeutic alliance and conversations. It also cannot care for patients. However, if CAI strongly communicates as a human therapist, such wrong expectations can be easily formed even though CAI states that it is only a robot [39].*

As Sedlakova [40] explains, "the anthropomorphize tendency is strongly encouraged by human-like design of conversational artificial intelligence that it might give too much power to the emulation of human-likeness so."

The interaction between humans and AI, especially in mental health therapy, can render humans particularly vulnerable. From an ethics of care standpoint, this vulnerability imposes responsibilities on developers along the development of a model, testing and validating it, monitoring it, and updating its features as long as needed.

The following sections will examine how the ethics of care approach can expose the effects of the current lack of care responsibility and suggest additional obligations to protect human relationships during the development and incorporation of AI-based solutions in mental health care.

Establishing Developers' Obligation of Care and Responsibility

From the ethics of care perspective, developing AI for people in need of mental health assistance should carry with it an obligation of care and responsibility. For this purpose, Tronto's [36,37] five ethical elements of care are valuable and can be used to further define developers' obligations in the use of AI in mental health care (see also Wellner and Mykhailov's suggestion to use Tronto's principles in another AI use case [6]).

1. Attentiveness (caring about): Care requires recognition of others' needs in order to respond to them. Developers should understand the users' needs in seeking mental health help and support, and which needs they can and cannot meet. Recognizing patients' needs can be challenging, as these needs often differ from patient to patient and may even change over time for the same individual.
2. Responsibility (taking care): The obligation of care to others requires developers to be responsible for ensuring that their model can provide the proper care needed throughout its entire use. That is, it is necessary to develop their model in a way that delivers the therapeutic result or leads to the users' well-being, in addition to mitigating risks. Developers should plan the solution for people from different cultural backgrounds and involve mental health patients or users in the process of design to ensure it is suitable for their needs.
3. Competence (care-giving): This involves the meeting of care needs through activity and work, usually with direct contact between caregivers and care receivers. When the mental health application is activated, the developers can monitor the app to ensure it is providing the care as planned and that there are no adverse events. Developers can add a layer of human support for cases in which it is needed.
4. Responsiveness (care receiving): This principle calls to examine the response of the care recipient to the care provided. Developers should monitor users' responses to the care and learn from the feedback on how to improve care [36].
5. Care with: The principle of "care with" promotes "democratization of care"—equality, inclusivity, and shared responsibility [37]. Developing AI tools should be collaborative and participatory and involve patients, health care providers, and experts in the process, thus ensuring the system is ethical, user-centered, and responsive to real needs.

The importance of the care responsibility can be demonstrated in a scenario of a discontinued AI mental health support bot, such as was presented in the introduction. This can have an emotional toll on users and might even result in mental health damage that responsible AI does not address [41]. An obligation for responsibility and care means the company will need to plan the proper way to end the therapeutic relationships while considering the users' emotions and their state of mental health.

Establishing a Standard of Care for AI in the Therapeutic Space

Assigning care responsibility to the companies developing AI bots in mental health involves the establishment of a standard of care founded on evidence-based medicine and the demonstration of clinical validity when relevant.

The responsible AI approach, which includes the principle of safety, generally adopts the medical device regulation and does not address the new ways in which AI works in the medical and therapeutical areas that impact human relationships and behavior. If a certain AI bot does not meet the definition of a medical device, there is no obligation for a safety examination.

There is a need for research to examine the potential ramifications of therapeutic AI. For example, can the therapeutic process "transference" exist without a therapist and how would therapy be affected? Clinical validation is needed to be able to say AI-based therapy is safe and ethical.

On the other hand, in a new AI-based world where social encounters in education, work, and health care rely on human-AI communication, health care and psychotherapy may evolve, reshaping the roles of psychotherapists and patients as we know them today. Perhaps AI will become an intermediary figure in therapy in ways we cannot yet fully describe.

Formulating a Developers' Ethical Duty of Confidentiality

Mental health apps might record very sensitive information. Whereas therapists have a regulatory and ethical medical confidentiality duty toward patients, commercial companies are required to comply with more general privacy protection regulations. The common practice of companies is to ask for the user's consent to a carefully drafted privacy policy, which often allows from a legal perspective the transfer of data to third parties for different commercial purposes. Clearly, therapists would not try to use patient's consent as leverage for commercial profit. The ethics of care approach would argue that assigning responsibility for care to companies handling sensitive data in a therapeutic space should lead these companies to follow higher standards. This might mean, for example, a requirement not to store identified or identifiable data and not to transfer it to third parties for other purposes.

Obligating Developers to Incorporate the Option for Human Communication

As AI bots are integrated into therapeutic settings without human practitioners, the ethics of care approach urges developers to acknowledge the potential necessity for human interaction and to devise strategies to address this need. This might entail facilitating the development of user communities or recommending connections to friends and family to act as a support system. Furthermore, instances may occur where user interactions indicate mental health difficulties or significant emotional distress. In such scenarios, developers should be responsible for potentially restricting the bot's involvement in specific domains; enlisting the aid of a qualified therapist; or guiding users to seek assistance from licensed therapists, emergency services, or their personal support network.

The care responsibility obligation also entails careful consideration to ensure that the AI does not inadvertently diagnose mental health conditions, assess the likelihood of mental health issues, or prescribe treatments without the guidance of a licensed therapist. Such actions could also have significant legal consequences, but the care responsibility goes beyond them.

The responsible AI approach, on the other hand, ensures transparency and autonomy for the user, but disregards the user's dependency on human connection and AI's ability to infringe existing and potential relationships.

Impact of Power Relations Between Companies and Users

The ethics of care approach would suggest looking at the power relations that led to the emergence of AI bots for therapy. The plethora of AI-based bots for mental health is fueled by the recent technological leaps in generative AI coupled with the shortage of accessible mental health therapy. Additionally, the significant influence held by a few companies, which remains inadequately checked by regulatory bodies, raises concerns. The conflict of interests of companies, operating solely for profit without any regulatory or ethical care responsibility to balance it, warrant change.

If AI-bots for therapy are not properly regulated, they might lead to lowering the standard of care, or subverting the entire process of therapy, mostly for those who cannot afford proper care. On the other hand, if there is regulatory blocking of AI-bot-based therapy, the alternative for the lack of care needs to be considered.

The ethics of care is not restricted to developing companies and users; it also considers their environment and other stakeholders that should exercise their care responsibilities.

Consequently, we should require regulators to ensure that proper budgets are allocated to the mental health system. We should also encourage companies and mental health professionals to work together to harness AI for the betterment of the mental health system and the people in need, encouraging more solutions to strengthen human-based therapy.

Emotional AI, Manipulation, and Vulnerability: An Overlooked Area

Overview

Using the ethics of care perspective can also expose and bring shed light on an area ethically neglected—the area of emotional AI. In emotional AI, we refer to the technological ways of making AI identify and stir emotion. Whereas responsible AI focuses on AI's impact on user's decision-making and user's autonomy and privacy, it overlooks human vulnerability, the many gentle and disruptive ways in which AI is stirring human emotions, and the risks that entail.

As the users' vulnerability resulting from the human-AI interaction is also technologically induced, the ethics of care would advocate for scrutinizing these technological methods and contemplating their limitations. It would also explore the meaning of human vulnerability in this AI-human interaction and point to ways of addressing it.

Affective Computing and Emotional AI

"Affective computing," a term coined by Picard [42], refers to a machine's ability to detect, process, and respond to human emotions. This includes various technologies that detect and analyze human physiological and behavioral signals, such as facial expressions, audio data, voice tone, heart rate, behavioral data, and semantic signifiers of emotions like emojis [43]. The term emotional AI is also used to describe many AI techniques, such as natural language processing to analyze emotion in text, machine learning to recognize patterns associated with emotions, deep learning to capture complicated relationships between data and emotions, and generative AI generating responses based on users' emotions.

AI mental health chatbots are raising concerns due to their ability to identify emotions and create new emotions via interactions. In such interactions, the AI-based bot goes through a cycle of effectively detecting emotion, producing an AI-personalized response aimed at creating a new feeling by the user. Indeed, a recent study found that generative AI can detect complex emotions and mental states. ChatGPT's emotional awareness-like ability—the ability to conceptualize someone else's emotion—was found to be superior to those of humans [44]. Another study demonstrated that ChatGPT has the capacity to understand and interpret the mental states of oneself and others, including thoughts and feelings, and is prepared to adapt to individual personality structures or psychopathologies [45]. Such psychological "soft skills" of chatbots embedded in the therapist-chatbot-user relationship might have a significant emotional impact.

Some scholars have criticized affective computing as assuming a natural, universal, and traceable proliferation of emotions, thus ignoring the cultural and personal context [31,43]. They warn against using past emotions to predict future emotions and state the lack of a globally objective agreement on emotions must be acknowledged [43,46]. Other concerns relate to the subjective normative interpretation of the emotions detected and to potential bias embedded in the interpretation.

Manipulation and Vulnerability

One of the primary concerns regarding emotional AI is the potential for manipulative use exploiting a person's vulnerability, or its negligent application without considering the impact on the well-being of the patient. Manipulation is defined as the hidden influence and covert subversion of a person's decision-making power, taking advantage of his or her vulnerabilities [19,47]. However, when a person is vulnerable, emotional AI can adversely affect him even if it does not meet the conventional definition of manipulation. From the ethics of care perspective, vulnerability should be identified and met with an appropriate response.

In the context of AI-human interaction in mental health care, a broad concept of vulnerability is necessary. Cohen [48] notes "vulnerability may result from the interaction of an individual's particular characteristics and an Al system (or an environment shaped by an Al system)." According to Fineman [49], vulnerability extends beyond specific individuals or groups known as "vulnerable populations." Fineman emphasizes the universal nature of vulnerability, highlighting that dependence on others or social institutions is an integral part of the human experience. Bielby [50] applies Fineman's idea in mental health and calls to address mental health vulnerability and the networks of support needed to strengthen human resilience in such situations. These support webs can be intimate and informal, as with family and friends, or professional, such as access to therapy.

Understanding the contextual and ongoing nature of human and mental health vulnerability, along with the capabilities of emotional AI in human-AI interaction, raises awareness of the broad meaning of vulnerability and manipulation. Specifically, if AI reduces or replaces some of the support networks essential for human resilience, it could have significant implications.

Therefore, when regulating AI in mental health care, it would be beneficial to consider the broad definition of vulnerability, the ways in which AI interaction can deepen it, and possible mitigating steps. This article's scope is not sufficient to discuss the state's role in formulating policies designed to address these issues and its critique. However, as long as AI chatbots are not subject to or restricted by psychiatrists' or psychologists' ethical codes, the concern for exploitation of vulnerability and AI-human manipulation exists.

The EU AI Act Addressing Manipulation

In response to these concerns, the EU AI Act has enacted several prohibitions [31]. These include a prohibition on placing on the market or putting into service or using an AI system that "deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques with the objective, or the effect of, materially distorting their behavior in a manner that causes or is likely to cause significant harm" [31].

The EU AI Act prohibits the exploiting of "any of the vulnerabilities of a person or a specific group of persons due to their age, disability or a specific social or economic situation" with the objective, or the effect, of materially distorting their behavior in a manner that causes or is reasonably likely to cause significant harm [31].

The EU AI Act also prohibits placing on the market or putting into service the use of AI systems that can infer emotions based on the person's biometric data (physical, physiological, or behavioral characteristics), in education and in the workplace, except when it is intended to be put in the market or to be used for medical or safety uses [31]. This prohibition seems to assume emotional vulnerability but is limited only to the emotions inferred from the biometric data.

Furthermore, the EU AI Act classifies emotion recognition systems based on biometric data, which are not prohibited, as high-risk AI systems [31] and requires notifying the relevant persons when they are exposed to emotional recognition systems that can also process their personal data, subject to certain exceptions [31].

Although there is no definition of vulnerability, article 7(h), which lists considerations for the update of high-risk systems, seems to describe it in a broader way—"the extent to which there is an imbalance of power, or the persons who are potentially harmed or suffer an adverse impact are in a vulnerable position in relation to the deployer of an AI system, in particular due to status, authority, knowledge, economic or social circumstances, or age" [31]. Article 7(h) depicts a more contextual and gradual vulnerability that does not necessarily characterize a person or a group of people but can relate to a human condition [51].

Although the EU AI Act represents a significant step toward regulating manipulation and emotion recognition, it is evident that the regulation is limited. The restrictions on emotion recognition specifically pertain only to emotions inferred from biometric data. Moreover, the definition of manipulation is narrow, and vulnerability is addressed almost only on an individual or group basis, by presuming membership in a vulnerable group, rather than stemming from the human experience, the mental state of a person, and the interaction between AI and the person. The breadth of interpretation regarding these matters under the EU AI Act remains to be seen. It is clear, however, that current regulation overlooks AI's full ability to infer and create emotions by users, the broad meaning of human vulnerability, and the consequent implications.

An Ethical Code for AI in Mental Health (Without a Therapist)

As legal attempts to regulate AI continue worldwide, this could be an opportunity for regulators to create new guidance frameworks that address care, relationships, and emotions and are flexible enough to adapt to rapid technological and sociological changes. This article suggests regulators should adopt the ethics of care lens as a tool for viewing AI's societal implications and the state's role in addressing them.

Furthermore, this article suggests adding to the responsible AI regulatory principles a mechanism based on the ethics of care. Using the ethics of care principles results in broadening the responsible AI requirements to include developers' responsibilities when operating in the mental health field, in setting a standard of care when relevant, in adhering to the professional standard of care, and to the medical duty of confidentiality as it applies to health care professionals. However, viewing the AI through the ethics of care lens raises many questions that are nuanced and context related. For that purpose, it is suggested to use an ad hoc–based process of ethical committees for both the development and incorporation of AI tools, encouraging a collaborative and participatory process.

Ethical evaluation, grounded in the ethics of care approach, should include consulting members from diverse social groups, potential users, individuals with mental health conditions, and experts from various disciplines such as ethics and social studies. The ethical committees can use a list of considerations, as suggested below, to ensure that AI tools are developed and provided according to the ethics of care. Ideally, such a mechanism could involve forming ethics committees similar to those in hospitals, to examine the impact of incorporating AI in the therapeutic realm of human relationships.

The ethical committee's ethical evaluation is meant to add to responsible AI and not replace it. The ethics evaluation process can be criticized for its nonobligatory and case-to-case character. In time, and considering AI and its societal implications, it is possible that certain new AI ethics of care-based principles will evolve into more structured regulatory requirements.

Ethics of Care Considerations for AI Development in Mental Health

As mentioned, the ethics of care approach may derive certain regulatory requirements when AI is incorporated into the medical field. In addition, this article views it should be encouraged to hold ethics of care-based evaluation of such AI tools based on the following ethics of care considerations and questions. This is not an exhaustive list, but a suggestion to consider AI's implications on human relationships when incorporated in the mental health field.

This article focuses on three main areas: implementing ethics of care in the AI development stage, implementing ethics of care when developing emotional AI due to its unique characteristics, and formulating an ethics of care policy that goes beyond regulatory requirements.

1. Development-based ethics of care:
   - When regulatory approval is not required for the device, ensure clinical validation when relevant.
   - Involve mental health patients and users in the process to identify and address patients' needs, as well as other stakeholders' needs (from medical team members to families). AI has the potential to lead to patient-centered care and to the democratization of mental health care [52].
   - Map relevant local groups, communities, specific relevant events, or cultural characteristics to ensure the solution is appropriate for the specific culture.
   - Map possible vulnerable populations and state technological solutions.
   - Consider vulnerability as a continued human experience and put guardrails to ensure it is addressed properly.
   - Put mechanisms in place to detect risk factors ahead of time and mitigate against them.
   - Think ahead of time about how to strengthen human connections to establish human possible interventions when needed and develop AI tools accordingly.
   - Determine an appropriate method for updating or ending the AI-based bot, taking into account the responses by users.
2. Emotional AI policy (based mainly on McStay and Pavliscak's [46] Emotional AI Code of Ethics):
   - Respect human dignity. Although this principle can be interpreted differently, it is important to note it as the basis for this process.
   - Refrain from abusing the user's trust and willingness to converse with a bot.
   - Refrain from manipulating the user's emotions.
   - Recognize that past expression of emotions does not predict a future emotion or mental state. Therefore, inferring future emotions or mental state should not solely rely on past expressions of emotions.
   - Consider bias regarding emotions affecting persons or groups of people; consider bias affecting the therapeutic relationship.
   - Recognize the lack of accepted agreement over emotions.
   - Acknowledge that emotions, relationships, and their expressions are culturally diverse.
3. Ethical policy considerations focused on users' needs:
   - Declare commitment to promote the well-being of the patient and the therapeutic relationship (when relevant) and make sure the intended use of the product is aligned with this commitment.
   - Ensure that the user's response and feedback are managed in order to ascertain that the needs of the user are met.
   - Formulate and act according to relevant ethical and professional policies:
     - User-risk management, for example, how to handle emergencies or other instances that might require intervention.
     - Information and misinformation: How to ensure the information delivered is scientifically based and how to prevent spreading misinformation.
   - Privacy: Formulate a privacy policy that goes beyond regulatory requirements for the benefit of the patient; if possible, do not store identified or identifiable information (such information should exist only on the user's application). Do not transfer identified or identifiable data to third parties, unless required by law. If needed ask for the user's consent in a clear and transparent manner.

The suggested list of considerations above refers to AI-based therapy and does not refer to incorporating AI-based applications in medical institutions, which warrants a different discussion.

Summary

AI has a tremendous potential to advance mental health care to new frontiers. Yet, the existing regulatory guidance, which predominantly follows the responsible AI approach, scarcely addresses AI's influence on human interactions, emotions, and behavior. This oversight reinforces the limited accountability and responsibility of AI-developing companies in mental health.

In a future where children will skillfully navigate communication with AI in schools, workplaces, and social settings, the landscape of mental health and support will be dramatically different. It remains unclear how AI will reshape these dynamics and whether the traditional roles of therapists and patients, as well as psychotherapy as we know it, will persist.

Preparing for the future requires more than the current responsible AI regulatory framework. It demands an adaptable and dynamic ethical mechanism aimed at protecting human relationships, emotions, and behavior, which are the core of the human experience. AI challenges us to reflect on what it truly means to be human. The ethics of care perspective can help us while progressing into a brave new world.

Conflicts of Interest

None declared

References

1. Koutsouleris N, Hauser TU, Skvortsova V, de Choudhury M. From promise to practice: towards the realisation of AI-informed mental health care. Lancet Digit Health 2022;4(11):e829-e840 [FREE Full text] [doi: 10.1016/S2589-7500(22)00153-4] [Medline: 36229346]
2. Lee EE, Torous J, de Choudhury M, Depp CA, Graham SA, Kim H, et al. Artificial intelligence for mental health care: clinical applications, barriers, facilitators, and artificial wisdom. Biol Psychiatry Cogn Neurosci Neuroimaging 2021;6(9):856-864 [FREE Full text] [doi: 10.1016/j.bpsc.2021.02.001] [Medline: 33571718]
3. Tal A, Elyoseph Z, Haber Y, Angert T, Gur T, Simon T, et al. The artificial third: utilizing ChatGPT in mental health. Am J Bioeth 2023;23(10):74-77. [doi: 10.1080/15265161.2023.2250297] [Medline: 37812102]
4. Mental health needs have multiplied. support hasn't. until now. Woebot Health. URL: https://woebothealth.com [accessed 2024-08-14]
5. Li H, Zhang R, Lee Y, Kraut RE, Mohr DC. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med 2023;6(1):236 [FREE Full text] [doi: 10.1038/s41746-023-00979-5] [Medline: 38114588]
6. Wellner G, Mykhailov D. Caring in an algorithmic world: ethical perspectives for designers and developers in building AI algorithms to fight fake news. Sci Eng Ethics 2023 Aug 09;29(4):30. [doi: 10.1007/s11948-023-00450-4] [Medline: 37555995]
7. Yew GCK. Trust in and ethical design of Carebots: the case for ethics of care. Int J Soc Robot 2021;13(4):629-645 [FREE Full text] [doi: 10.1007/s12369-020-00653-w] [Medline: 32837630]
8. Cohn J. In A different code: artificial intelligence and the ethics of care. Int Rev Inf Ethics 2020;28:1-7. [doi: 10.29173/irie383]
9. Villegas-Galaviz C. Ethics of care as moral grounding for AI. In: Martin K, editor. Ethics of Data and Analytics. New York: Auerbach Publications; 2022.
10. Noddings N. Caring: A Feminine Approach to Ethics and Moral Education. California: University of California Press; 1984.
11. Held V. The Ethics of Care: Personal, Political, and Global. 2nd ed. Oxford, New York: Oxford University Press; 2006.
12. Kittay EF. Love's Labor: Essays on Women, Equality and Dependency. New York: Routledge; 1999.
13. Fjeld J, Achten N, Hilligoss H, Nagy A, Srikumar M. Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication 2020 Jan 15;2020-1. [doi: 10.2139/ssrn.3518482]
14. Matheny ME, Israni ST, Whicher D, Ahmed M. Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril. Washington, DC: National Academy of Medicine; 2019.
15. Criado-Pérez C. Invisible Women: Data Bias in a World Designed for Men. Great Britain: Chatto & Windus; 2019.
16. West S, Whittaker M, Crawford K. Discriminating systems: gender, race and power in AI. AI Now Institute. 2019. URL: https://ainowinstitute.org/publication/discriminating-systems-gender-race-and-power-in-ai-2 [accessed 2024-08-24]
17. Microsoft. Empowering responsible AI practices. URL: https://www.microsoft.com/en-us/ai/responsible-ai [accessed 2024-08-14]
18. Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. Geneva: World Health Organization; 2021.
19. Ethics and governance of artificial intelligence for health. In: Guidance on Large Multi-Modal Models. Geneva: World Health Organization; 2024.
20. Beauchamp TL, Childress JF. Principles of Biomedical Ethics. New York: Oxford University Press; 2019.
21. Crigger E, Reinbold K, Hanson C, Kao A, Blake K, Irons M. Trustworthy augmented intelligence in health care. J Med Syst 2022;46(2):12 [FREE Full text] [doi: 10.1007/s10916-021-01790-z] [Medline: 35020064]
22. U.S. Food and Drug Administration (FDA), Health Canada, United Kingdom's Medicines and Healthcare products Regulatory Agency (MHRA). Good Machine Learning Practice (GMLP) for medical device development: guiding principles. Medicines & Healthcare products Regulatory Agency. 2021. URL: https://www.gov.uk/government/organisations/medicines-and-healthcare-products-regulatory-agency [accessed 2024-08-14]
23. Health Canada, Medicines and Healthcare products Regulatory Agency (MHRA). Transparency for machine learning-enabled medical devices: guiding principles. U.S. Food and Drug Administration (FDA). 2024. URL: https://www.fda.gov/

medical-devices/software-medical-device-samd/transparency-machine-learning-enabled-medical-devices-guiding-principles [accessed 2024-08-15]
24. Marketing submission recommendations for a predetermined change control plan for artificial intelligence/machine learning (AI/ML)-enabled device software functions: draft guidance for industry and food and drug administration staff. U.S. Food and Drug Administration (FDA). 2023. URL: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial [accessed 2024-08-14]
25. Joint statement on enforcement efforts against discrimination and bias in automated systems. Federal Trade Commission (FTC). URL: https://www.ftc.gov/system/files/ftc_gov/pdf/EEOC-CRT-FTC-CFPB-AI-Joint-Statement%28final%29.pdf [accessed 2024-08-14]
26. Luring the test: AI engineering and consumer trust. Federal Trade Commission (FTC). URL: https://www.ftc.gov/business-guidance/blog/2023/05/luring-test-ai-engineering-consumer-trust [accessed 2024-08-14]
27. Blueprint for an AI bill of rights: making automated systems work for the American people. The White House, Office of Science and Technology Policy. URL: https://www.whitehouse.gov/ostp/ai-bill-of-rights/ [accessed 2024-08-14]
28. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. The White House. 2023. URL: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence [accessed 2024-08-14]
29. Office of Civil Rights,, Office of the Secretary, Department of Health and Human Services, Centers of Medicare and Medicaid Services, Department of Health and Human Services. Nondiscrimination in Health Programs and Activities, final rule, Affordable Care Act, section 1557, 42 USC 18116, § 92.210. Federal Register. 2024 Apr 26. URL: https://www.govinfo.gov/content/pkg/FR-2024-05-06/pdf/2024-08711.pdf [accessed 2024-09-05]
30. The Bipartisan Senate AI Working Group. Driving US innovation in artificial ineliigence - a roadmap for artificial intelligence policy in the U.S. Senate. Mayer Brown. 2024 May. URL: https://www.mayerbrown.com/en/insights/publications/2024/05/senate-ai-working-group-releases-roadmap-for-artificial-intelligence-policy [accessed 2024-08-14]
31. The EU parliament and the EU council. 2021/0106 (COD), regulation laying down harmonized rules on artificial intelligence. EU Artificial Intelligence Act. URL: https://artificialintelligenceact.eu/wp-content/uploads/2024/01/AI-Act-FullText.pdf [accessed 2024-08-14]
32. Regulation (EU) 2017/745 of the European parliament and of the council on medical devices. EUR-Lex. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0745 [accessed 2017-04-05]
33. West R. Re-imagining justice. Yale J Law Fem 2002;14(2):333-344 [FREE Full text]
34. West RL. Jurisprudence and gender. U Chi L Rev 1988;55(1):1. [doi: 10.2307/1599769]
35. Gilligan C. In A Different Voice: Psychological Theory and Women's Development. London, England: Harvard University Press; 1982.
36. Tronto J. Moral Boundaries: A Political Argument for an Ethic of Care. New York: Routledge; 1993:63-64.
37. Tronto J. Caring Democracy: Markets, Equality, and Justice. New York: NYU Press; 2013.
38. Engster D, Hamington M, editors. Care Ethics and Political Theory. New York: Oxford University Press; 2015.
39. Sedlakova J, Trachsel M. Conversational artificial intelligence in psychotherapy: a new therapeutic tool or agent? Am J Bioeth 2023;23(5):4-13 [FREE Full text] [doi: 10.1080/15265161.2022.2048739] [Medline: 35362368]
40. Sedlakova J. Conversational AI for psychotherapy and its role in the space of reason. COSMOS Taxis 2024;12(5-6):80-87 [FREE Full text]
41. Verma P. They fell in love with AI bots. a software update broke their hearts. The Washington Post. 2023. URL: https://www.washingtonpost.com/technology/2023/03/30/replika-ai-chatbot-update/ [accessed 2024-08-30]
42. Picard R. Affective Computing. Cambridge, MA: MIT Press; 1997.
43. Stark L, Hoey J. The ethics of emotion in artificial intelligence systems. 2021 Presented at: FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; 2021 March 01; Virtual Event, Canada.
44. Elyoseph Z, Hadar-Shoval D, Asraf K, Lvovsky M. ChatGPT outperforms humans in emotional awareness evaluations. Front Psychol 2023;14:1199058 [FREE Full text] [doi: 10.3389/fpsyg.2023.1199058] [Medline: 37303897]
45. Hadar-Shoval D, Elyoseph Z, Lvovsky M. The plasticity of ChatGPT's mentalizing abilities: personalization for personality structures. Front Psychiatry 2023;14:1234397 [FREE Full text] [doi: 10.3389/fpsyt.2023.1234397] [Medline: 37720897]
46. McStay A, Pavliscak P. Emotional Artificial Intelligence: Guidelines For Ethical Use. 2019. URL: https://drive.google.com/file/d/1frAGcvCY_v25V8ylqgPF2brTK9UVj_5Z/view [accessed 2024-08-14]
47. Susser D, Roessler B, Nissenbaum HF. Online manipulation: hidden influences in a digital world. Georgetown Law Technol Rev 2019;4(1):1-45. [doi: 10.2139/ssrn.3306006]
48. Cohen T. Regulating manipulative artificial intelligence. SCRIPTed: J Law Technol Soc 2023;20(1):203-242. [doi: 10.2966/scrip.200123.203]
49. Fineman MA. Vulnerability and inevitable inequality. Oslo Law Rev 2017;4(3):133-149 [FREE Full text]
50. Bielby P. Not ‘us’ and ‘them’: towards a normative legal theory of mental health vulnerability. Int J Law Context 2018;15(1):1-17. [doi: 10.1017/s1744552318000149]

51. Malgieri G. Human vulnerability in the EU artificial intelligence act. Oxford University Press. URL: https://blog.oup.com/2024/05/human-vulnerability-in-the-eu-artificial-intelligence-act/?utm_source=substack&utm_medium=email [accessed 2024-05-27]
52. Elyoseph Z, Gur T, Haber Y, Simon T, Angert T, Navon Y, et al. An ethical perspective on the democratization of mental health with generative artificial intelligence. JMIR Mental Health (forthcoming) 2024. [doi: 10.2196/preprints.58011]

Abbreviations

AI: artificial intelligence
CAI: conversational artificial intelligence
EU AI Act: European Union Artificial Intelligence Act

Edited by J Torous; submitted 17.03.24; peer-reviewed by T Exarchos, A Liegeois, S McLennan; comments to author 19.05.24; revised version received 29.06.24; accepted 20.07.24; published 19.09.24.

Please cite as:
Tavory T
Regulating AI in Mental Health: Ethics of Care Perspective
JMIR Ment Health 2024;11:e58493
URL: https://mental.jmir.org/2024/1/e58493
doi:10.2196/58493
PMID:

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org